Systematic identification of relevant features for the statistical modeling of materials properties of crystalline solids

Dissertation
zur Erlangung des Doktorgrades der
Naturwissenschaften (Dr. rer. nat.)

am Fachbereich Physik der Freien Universität Berlin.

vorgelegt von
Benjamin Regler

Berlin, 2021
Freie Universität Berlin
Fritz-Haber-Institut der Max-Planck-Gesellschaft

Vom Fachbereich Physik
der Freien Universität Berlin als Dissertation am 1. Oktober 2021 angenommen.

Betreuer:	Dr. Luca M. Ghiringhelli (Gruppenleiter, FHI Berlin)
Erstgutachter:	Prof. Dr. Matthias Scheffler (Ehemaliger Direktor der Theorie-Abteilung, FHI Berlin)
Zweitgutachter:	Prof. Dr. Jens Eisert (Professor im Fachbereich Physik, FU Berlin)
Prüfungskommission:	Prof. Dr. Robert Bittl (Vorsitzender des Promotionsausschusses)
	Prof. Dr. Holger Dau (Vorsitzender der Prüfung)
	Prof. Dr. Matthias Scheffler (Erstgutachter)
	Prof. Dr. Jens Eisert (Zweitgutachter)
	Dr. David Rosenberger (Promovierter Wissenschaftlicher Mitarbeiter)
	Doguscan Ahiboz (Student im Aufbaustudium)

Tag der mündlichen Prüfung am 6. Juli 2022.

Declaration of authorship

Name: Regler
First name: Benjamin

I declare to the Freie Universität Berlin that I have completed the submitted dissertation independently and without the use of sources and aids other than those indicated. The present thesis is free of plagiarism. I have marked as such all statements that are taken literally or in content from other writings. This dissertation has not been submitted in the same or similar form in any previous doctoral procedure.

I agree to have my thesis examined by a plagiarism examination software.

Date: ________________ Signature: ___________________________

Abstract

Designing materials with desired properties is essential to developing new materials for today's challenges. Historically, new materials have been discovered through trial and error. Nowadays, materials can be simulated and designed on the computer before they are synthesized in the laboratory. However, despite increasingly powerful computational resources and automatized experiments, this process is still comparatively demanding.

Given the anticipated potential diversity of materials, a brute-force search for candidate materials with desired properties is impractical. In recent years, algorithms for building statistical models, especially machine learning, have been used to estimate properties from available materials data. These models relate a set of materials properties – the so-called features of the data set – to a property of interest. Because there is no standardized procedure for selecting a set of features related to a property of interest, materials data sets can have hundreds to thousands of features. As a result, models are often complex, placing high demands on computational resources.

This thesis proposes a systematic approach to reduce the number of features prior to statistical modeling and a framework for automatically constructing and estimating the prediction uncertainty of statistical models. The information-theoretic approach presented first allows a ranking of the identified features by quantifying the relevance of features in terms of their mutual dependence to the property of interest. Whereas traditional methods work well for discrete data, a method for continuous data is developed for the application to materials data. A framework for feature identification is designed that can be applied to information-theoretic methods as well as to machine-learning algorithms. The framework is based on the branch-and-bound algorithm and iteratively combines sets of features with the goal of identifying the features related to a property of interest with either the highest mutual dependence or the best prediction performance.

Examples with known as well as empirically identified feature-property relationships are used to compare the information-theoretic method and the developed framework with established methods. The framework is then applied to actual materials data sets. The information-theoretic method is robust in the presence of inter-correlated features and is stable with increasing numbers of data samples, but requires more data to identify the same set of features than machine-learning algorithms for feature identification. Generated machine-learning models therefore resulted in higher prediction errors. The same framework, but using machine-learning algorithms, required fewer features to achieve a comparable prediction performance to the models reported in the literature.

The framework identifies different sets of features that leads to an ensemble of statistical models with similar prediction performance. A number of additional tools are developed to further identify feature inter-correlations and to estimate the prediction error within a probabilistic tolerance. These tools can be used to assess the limitations of the generated models in predicting the desired property of new materials, to determine which materials cannot be predicted, and to find the features related to the property of interest in a model-independent framework for feature identification and model construction.

Zusammenfassung

Das Design von Materialien mit gewünschten Eigenschaften ist für die Entwicklung neuer Materialien für heutige Herausforderungen von entscheidender Bedeutung. Historisch gesehen wurden neue Materialien hauptsächlich durch Versuch und Irrtum entdeckt. Heutzutage können Materialien am Computer simuliert und entworfen werden, bevor sie im Labor synthetisiert werden. Doch trotz immer leistungsfähigerer Rechenressourcen und automatisierter Experimente ist dieser Prozess vergleichsweise anspruchsvoll.

Angesichts der Vielfalt an Materialien ist eine Suche durch simples Ausprobieren von Kandidatenmaterialien mit gewünschten Eigenschaften ungeeignet. In den letzten Jahren wurden Algorithmen zur Erstellung statistischer Modelle verwendet, darunter Maschinelles Lernen, um Materialeigenschaften aus verfügbaren Daten zu schätzen. Diese Modelle setzen einen Menge von Materialeigenschaften – die sogenannten Features des Datensatzes – in Beziehung zu der gesuchten Eigenschaft. Da es kein standardisiertes Verfahren zur Auswahl der Features in Bezug auf die interessierende Eigenschaft gibt, können Materialdatensätze Hunderte bis Tausende von Features aufweisen. Infolgedessen sind erstellte statistische Modelle oft komplex und stellen hohe Anforderungen an die Rechenressourcen dar.

In dieser Doktorarbeit wird ein systematischer Ansatz entwickelt, um die Anzahl der Features vor der statistischen Modellierung zu reduzieren, sowie ein Framework, um statistische Modelle automatisch zu erstellen und deren Vorhersageunsicherheit abzuschätzen. Der vorgestellte informationstheoretische Ansatz quantifiziert zunächst die Relevanz von Features in Bezug auf ihre gegenseitige Abhängigkeit zur gesuchten Eigenschaft, was eine Rangfolge der identifizierten Features ermöglicht. Weil herkömmliche Methoden nur für diskrete Daten geeignet sind, wird eine Methode für kontinuierliche Daten entworfen und auf Materialdaten angewendet. Darauf aufbauend wird ein Verfahren zur Identifizierung von Features entwickelt, das sowohl auf informationstheoretische Methoden als auch auf maschinelle Lernalgorithmen angewendet werden kann. Das Verfahren basiert auf dem Branch-and-Bound-Algorithmus und kombiniert iterativ Teilmengen von Features mit dem Ziel, die Features mit der höchsten gegenseitigen Abhängigkeit oder der besten Vorhersageleistung zu identifizieren.

Anhand von Beispielen mit bekannten sowie empirisch ermittelten Feature-Eigenschafts-Beziehungen wird die informationstheoretische Methode und das entwickelte Framework mit etablierten Methoden verglichen. Das Framework wird dann auf konkrete Materialdaten angewendet. Die informationstheoretische Methode ist robust bei untereinander korrelierten Features und ist stabil mit zunehmender Anzahl von Datenproben, benötigt aber mehr Datenproben zur Identifizierung derselben Anzahl an Features als maschinelle Lernalgorithmen zur Feature-Identifizierung. Die daraus generierten Machine-Learning-Modelle führen daher zu höheren Vorhersagefehlern. Das gleiche Framework, aber unter Verwendung von maschinelle Lernalgorithmen, benötigte weniger Features, um eine vergleichbare Vorhersageleistung wie bei den in der Literatur beschriebenen Modellen zu erzielen.

Das Framework identifiziert verschiedene Teilmengen von Features, die zu einem Ensemble von statistischen Modellen mit ähnlicher Vorhersageleistung führen. Eine Reihe von zusätzlichen Werkzeugen wurde entwickelt, um Feature-Interkorrelationen zu identifizieren und den Vorhersagefehler innerhalb einer probabilistischen Toleranz abzuschätzen. Diese Werkzeuge werden verwendet, um Einschränkungen der generierten Modelle bei der Vorhersage der gewünschten Eigenschaft neuer Materialien zu beurteilen, um festzustellen, welche Materialien nicht vorhergesagt werden können, und um Features zu finden, die mit der gewünschten Eigenschaft in einem modellunabhängigen Framework für die Feature-Identifikation und Modellkonstruktion zusammenhängen.

Nomenclature

Symbols

Y	Property of interest
X, X'	A single feature of the data set $\mathscr{D}$
$\vec{X} = \{X_1, \ldots, X_n\}$	A set of n features of the data set $\mathscr{D}$
$\vec{X}$	Feature vector, shorthand notation for $\vec{X} = \{X_1, \ldots, X_n\}$
$f : \vec{X} \longrightarrow Y$	A mapping (i.e., a statistical model) between the features $\vec{X}$ of the data set $\mathscr{D}$ and the property of interest Y
$\lvert\vec{X}\rvert$	Cardinality of a feature subset $\vec{X}$, i.e., the number of features $(i = 1, \ldots, n)$ in $\vec{X}$
$\vec{X} \setminus X_i$	Feature subset without the ith-feature
$J(\vec{X}, \vec{X}')$	Jaccard similarity (coefficient) between two feature subsets $\vec{X}$ and $\vec{X}'$
p, P	Probability and cumulative (probability) distribution
P'	Residual cumulative (probability) distribution $(P' = 1 - P)$
$\hat{\mathscr{E}}$	Empirical estimator of a quantity $\mathscr{E}$
α	Confidence level $\in [0, 1]$
ϵ	Significance level $= 1 - \alpha$

Abbreviations

i.i.d.	Independent and identically distributed

Acronyms

ML	Machine learning
CV	Cross validation
FS	Feature selection
GBDT	Gradient-boosting decision trees
RFE	Recursive feature elimination
TCMI	Total cumulative mutual information
SISSO	Sure-independence screening and sparsifying operator
TB3	Tolerance-based branch-and-bound (algorithm)

To my wife and my daughter, my mother, and my family.
In love, in memoriam, in deep gratitude.

Contents

Chapter 1

Computational materials discovery and design

Materials discovery and design is essential to developing new materials for today's challenges ranging from catalysts and superconductors to batteries and renewable energy sources [1–4]. Historically, new materials have been discovered through phenomenological observations in metallurgy and mineralogy and were often found by pure serendipity, trial-and-error approaches, or by analogies to existing systems. Nowadays, materials can be simulated and designed on the computer before they are synthesized in the laboratory. Such computationally-guided design and synthesis of materials have become characteristic for the present and past decade joining experiment, theory, and computation to comprise a fundamental framework for materials science and engineering referred to as the "fourth paradigm of science" [5, 6].

1.1 Motivation

The nascent field of materials informatics combines materials science and engineering with informatics to develop new techniques for computationally-guided materials discovery and design [6–9]. Materials informatics aims at facilitating data acquisition, storage, management, and dissemination of materials data [10, 11] and at systematically searching the space of possible materials to discover and design new compounds with optimized properties [12]. In recent years, automated workflows have been adopted to materials science [13–24], using conventional tools (such as density-functional theory [25, 26]) to calculate materials properties or structures based on quantum mechanics [27–29]. To date, millions of properties and structures have been computed which can be accessed through extensive and centralized materials databases [30–37]. However, even though materials databases are constantly growing at unprecedented rates, taken together, they only cover a tiny fraction of the anticipated potential diversity to be as large as a googol[1] of theoretical materials [38, 39]. As a result, neither the storage nor the screening of all possible materials for new applications are viable. Therefore, alternative approaches must find promising candidate materials for the targeted applications from limited theoretical, simulated, and experimental data.

[1] A googol is the large number represented as one followed by 100 zeros (10^{100}).

1.2 Scope and contents of the thesis

Over the past two decades, statistical tools have become indispensable in the materials-science pipeline of materials discovery and design [4, 40–43], in reducing the inherently time-consuming and costly process of developing and manufacturing new materials [9, 44–46][2]. In particular, artificial intelligence [49, 50] and its sub-field machine learning [51, 52] are increasingly being used to accelerate the search for new materials and to estimate materials properties, driven by a growing infrastructure of data-science tools for generating, testing, and refining scientific models. Unique to machine learning is the identification of candidate materials based on statistical correlations of a portion of characterized materials, whose models improve as more data are obtained from further calculations or experiments. This process of generating mathematical models, whose prediction performance improves with the availability of data [53, 54], makes machine learning ideally suited for tasks where conventional tools fail due to the combinatorial explosion of the materials-space exploration or to the explosion in the computational or experimental cost and time.

The accelerated search for materials is a collaborative effort of experimental validation and integration of theory, simulations, and experiment. Among the major challenges are the quantitative characterization of the materials with respect to the property of interest [55], the development of general-purpose machine-learning algorithms [56], and the robust estimation of the properties of interest from the limited amount of materials data [57, 58]. So far, machine learning is applied on a case-by-case basis, requiring expertise and a careful analysis of the materials data set. In addition, there are potential risks with relatively small and heterogeneous data sets that are confined to a particular domain of the materials space: Due to inherent assumptions of the machine-learning algorithms, this can lead to limited predictive capabilities of generated machine-learning models [59].

This thesis addresses some of the challenges of materials discovery and design by proposing a systematic data-driven and model-independent framework to identify the set of the features that are multivariately and non-linearly related to a property of interest (e.g., materials properties[3]) prior to statistical modeling. The framework includes a novel method (total cumulative mutual information, Section 3.3.3, and the tolerance-based branch-and-bound algorithm, Section 4.1) for identifying these sets of features and uses these sets to automatically construct machine-learning models. In addition, the framework uses conformal prediction [60–64] to estimate the model's prediction uncertainties (Section 4.3) and includes a novel method for estimating the prediction capabilities of machine-learning models (referred to as credibility, Section 4.4), along with tools for identifiying multivariate non-linear feature relationships within a probabilistic tolerance (Section 4.2).

The focus of the thesis is the extensive and systematic search of a set of features related to the materials properties of interest, the construction of reliable machine-learning models, and the relation of the identified feature sets to the prediction performance of the generated models. Throughout this thesis, available materials data are used to investigate different feature-identification methods and to compare their prediction performance with established machine-learning approaches. The main theme is to uniquely characterize each material by a set of features through the use of tabulated

[2] The typical time span to just discover a new material is in the range of 6–12 months [47], whereas the time span between the discovery of a material to an actual application typically takes more than 20 years [48].

[3] These are input properties used to numerically identify the material and estimate the property of interest.

and easily accessible materials properties defined at varying levels of complexity [55, 65, 66]: from the microscopic level based on molecular or atomic properties to higher-level structural properties, emergent properties at the macroscopic level, and environmental conditions such as temperature and pressure. The key is to establish a mapping between the features of the data set and the properties of interest to find the relevant features that are related to the property of interest from a larger set of candidate features, to get a comprehensive understanding of the feature-property relationships, and to determine where the model fails or succeeds even if the model may operate on unphysical principles [67, 68].

1.3 Thesis structure

The structure of this thesis is as follows (Fig. 1.1):

Chapter 2 introduces the concepts used in the thesis (Section 2.1) and highlights some of the recent developments in materials science (Sections 2.2 and 2.4): the application of high-throughput methods to first-principles calculations, the emergence of materials databases, and the application of machine learning in the context of materials discovery and design. First-principles calculations are based on solving the Schrödinger equation of a quantum-mechanical system within a set of approximations[4]. For electronic-structure calculations, density-functional theory is one of the most accurate theories to compute (ground-state) materials properties from quantum mechanics. Its mathematical foundation and a brief history are summarized in Section 2.2. Despite computationally intensive first-principles calculations and costly experiments, high-throughput approaches have generated large amounts of theoretical as well as experimental data that are increasingly being stored in materials databases [30–37] (Section 2.3). To harness and use these data in scientific applications other than those for which the first-principles calculations or experiments were performed, alternative methodologies from statistics and computer science are used to identify patterns based on existing data to construct mathematical models for estimating the targeted properties of interest. The application of statistics and computer-science methodologies to materials science is outlined in Section 2.4 and the existing challenges and limitations are discussed in Section 2.6.

Despite the substantial impact statistics and methodologies from computer science have had on materials science and other fields, there is not yet a consensus on best practices for computational materials discovery and design. Furthermore, in a field where predictions about the behavior of materials have historically been based purely on the fundamental laws of physics (as given by the rigorous theoretical foundation of quantum mechanics and the Schrödinger equation [69]), there is a great interest in gaining insight into the machine-learning models to understand the models' predictions and to find empirical relationships for designing new materials.

Typically, the set of features used to characterize each material in a data set depends on the property of interest and therefore requires expertise and knowledge of the materials classes as well as of the targeted application. Chapter 3 discusses the representation of materials and the challenges therein, i.e., the quantitative characterization of the materials via features of the data set and the identification of features related to a property of interest. Because the materials representation may comprise

[4] Born-Oppenheimer approximation, Hartree or Hartree-Fock approximation, choice of exchange-correlation functional, etc.

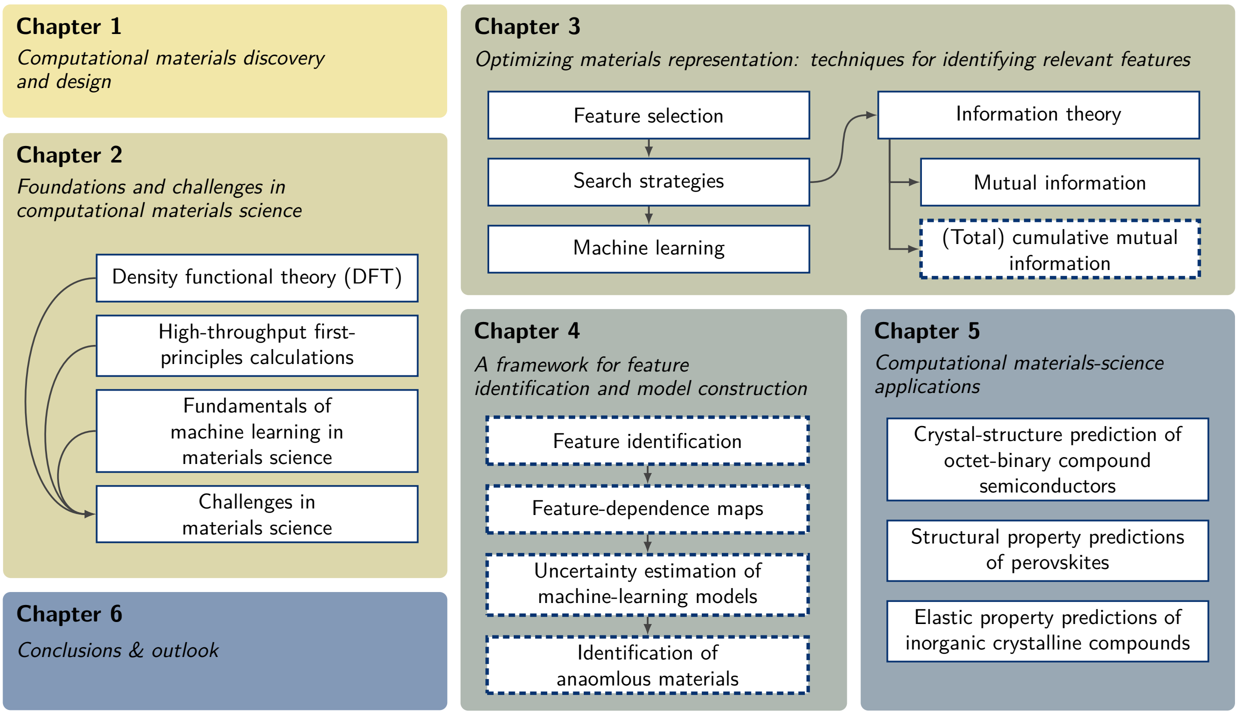

Chapter 1
Computational materials discovery and design
Chapter 2
Foundations and challenges in computational materials science
Density functional theory (DFT)
High-throughput first-principles calculations
Fundamentals of machine learning in materials science
Challenges in materials science
Chapter 6
Conclusions & outlook
Chapter 3
Optimizing materials representation: techniques for identifying relevant features
Feature selection
Search strategies
Machine learning
Information theory
Mutual information
(Total) cumulative mutual information
Chapter 4
A framework for feature identification and model construction
Feature identification
Feature-dependence maps
Uncertainty estimation of machine-learning models
Identification of anaomlous materials
Chapter 5
Computational materials-science applications
Crystal-structure prediction of octet-binary compound semiconductors
Structural property predictions of perovskites
Elastic property predictions of inorganic crystalline compounds

Fig. 1.1. A schematic illustration of the structure of the thesis. The framework and tools proposed in the dashed boxes (which are developed as part of this thesis and are applied in Chapter 5) represent new aspects compared to previous works.

many features, which practically cannot be explored combinatorially by enumeration, a reduction of the number of features is an essential part of an efficient data-driven materials-science pipeline for materials-space exploration, visualization, and understanding of the statistical relationships in the data. Section 3.1 gives a brief overview of existing feature-reduction techniques to either select a small subset of potentially related features to the property of interest [70–72] or to project features onto a lower dimensional manifold of the materials space [73–82]. Feature-reduction techniques are optimization problems and as such are closely related to information-theoretic concepts such as mutual information [83, 84] and the Kullback-Leibler divergence [85]. An information-theoretic approach is outlined in Section 3.2 to identify the features related to the properties of interest by minimizing the Kullback-Leibler divergence between a subset of the features of the data set and the targeted material's property [86–89]. Based on this, an information-theoretic method (TCMI) is developed in Section 3.2.5 to identify both linear and non-linear correlations without assuming any explicit form of the feature-property relationship in the data (cf., [90]). Because information theory is based on a rigorous mathematical framework for feature identification, TCMI has the potential to significantly reduce the number of features required to accurately estimate the properties of interest, while providing a better understanding of the statistical relationships in the data and the generated statistical models.

A general problem is to find the optimal set of features to accurately estimate the property of interest. Because the search is of combinatorial complexity and therefore cannot be solved exhaustively, an efficient optimization strategy is crucial to finding optimal or close to optimal (i.e, sub-optimal) solutions. In Section 3.4, the branch-and-bound algorithm [91–93] is highlighted, a combinatorial optimization search algorithm that uses a so-called feature-selection criterion to enumerate the space of all feature-subset combinations, thereby saving time in discarding subsets whose feature-selection criterion cannot be improved [94–96]. The branch-and-bound algorithm requires a monotonically increasing feature-selection criterion like the information-theoretic approach presented in Section 3.2.5. Because the branch-and-bound algorithm has been proven to be useful in the discovery of non-linear functional dependences [97, 98], an information-theoretic feature-selection method is proposed based on TCMI. Advantages and disadvantages of TCMI and the branch-and-bound algorithm are discussed in Section 3.5 by comparing the feature-subset search with TCMI to established feature-selection on three examples with known as well as empirically identified feature-property relationships.

To actually compare feature subsets in a data-driven framework and relate them to the prediction performance of a machine-learning model, it is beneficial to extend the principles of information theory and the feature-subset search to machine learning. In Chapter 4, the branch-and-bound algorithm is generalized to be applicable to both information-theoretic methods as well as to machine-learning algorithms. Unlike information-theoretic methods, most machine-learning algorithms are sensitive to strongly dependent or multi-collinear related features [99]. Pairwise linear-correlation heat maps and multivariate feature-dependence maps [100] are essentially used in all materials-science applications to identify such feature correlations. In Section 4.2, a new approach is presented to combine feature-dependence maps with the developed feature-identification framework to quickly identify non-linear and multivariate dependent features. As such, it goes beyond pairwise interactions and therefore avoids drawing errorneous conclusions on the relationship between features or between a set of features and the property of interest.

The prediction performance of a machine-learning model is commonly expressed in terms of a single metric such as the root-mean-squared error or the Pearson's coefficient of determination [101]. While a single metric is useful for estimating the goodness-of-fit of a model, it does not provide any information about the error between the model's prediction and the actual value in estimating the property of interest of new materials. For this reason, a quantification of the model's uncertainty is indispensable for estimating the robustness and reliability of machine-learning models, especially when neither knowledge about the underlying relationship nor extensive model validations are available [102–104]. Rather than incorporating an error model into the machine-learning algorithm such as in Bayesian statistics in deep learning [105–107] or Gaussian Process Regression [108, 109], the approach presented in this thesis directly estimates the underlying distribution of model errors either from an ensemble of machine-learning models generated from feature-subset search [52] or from the machine-learning model itself [60–64]. Because both type of methods are sensitive to the choice of training data, a resampling strategy is applied to robustly estimate the model's uncertainty.

In Section 4.4, a heuristic measure is developed, called credibility, to identify materials for which there is insufficient data to reliably estimate the properties of interest. Credibility is calculated by comparing the prediction of a new material to known values of similar materials in the data set.

Chapter 5 links all the presented and developed methods into a model-independent framework for feature identification and model construction. The framework for feature identification and model construction is designed to automatize the model creation and to enable a data-driven identification and characterization of features related to the property of interest. Therefore, three increasingly challenging materials-science applications are reviewed in Chapter 5 to discuss and demonstrate the challenges of identifying relevant features in materials science and to investigate the applicability of the developed feature-identification framework for the quantitative prediction of the crystal structure of octet-binary compound semiconductors (Section 5.2.1), the prediction of structural properties of perovskites (Section 5.2.2), and the prediction of elastic properties of inorganic crystalline compounds (Section 5.2.3).

Finally, Chapter 6 concludes with a discussion of the presented and developed methods, the framework, and yet unanswered questions to be tackled in future research. These include the optimization of the information-theoretic feature-identification methods to large materials data sets, the application of the developed framework to a wider range of materials-science applications, and the accurate modeling of the statistical trends in the data with machine learning.

Chapter 2

Foundations and challenges in materials design

The number of all hypothetical and realizable materials results in a combinatorial explosion of computational and experimental demands. The goal of materials design is therefore to optimize materials properties and to find candidate materials for desired applications, e.g., by using first-principles calculations, experiments, or statistical methods.

First-principles methods determine the microscopic properties of a system based on the Schrödinger equation [69], the fundamental equation in quantum mechanics. Because the Schrödinger equation can be solved exactly only for systems with a small number of electrons ($N < \mathcal{O}(10)$ [110]), solutions of the Schrödinger equation for larger systems are usually approximated. Nevertheless, these approximations are consistent as a whole [29] and have found widespread applications in chemistry and physics to compute total energies, thermodynamic properties, structures, and energy spectra of molecules or crystalline materials [111].

First-principle methods and experimental efforts complement each other in the search for new materials: In contrast to experiments, where each structure of a material needs to be synthesized and tested, first-principles methods can optimize the material's structure while determining the (ground-state) properties of a material. Because not all atomic configurations in the search are stable and thus synthesizable, experiments, in turn, can be used to validate results from first-principles calculations and to provide useful information about the stability of the materials.

The collection, assimilation, and dissemination of first-principles calculations and experimental data to and from materials databases enable entirely new approaches to the analysis, screening, and prediction of novel materials [9]. Statistical methods, such as machine learning, estimate the property of interest from available data. These methods can be used to identify potential trends in the data and to discover new materials. As such, they have the potential to rationalize high-throughput characterization of materials, the prediction of materials properties, and the extraction of qualitative and quantitative rules based on available data. In particular, data-driven approaches can be utilized whenever properties are difficult to measure or to calculate, for properties of interest whose fundamental equations are not (yet) known, or a direct solution of the fundamental equations is unlikely [112]. In these cases, machine learning can be seen as an intermediate step for understanding

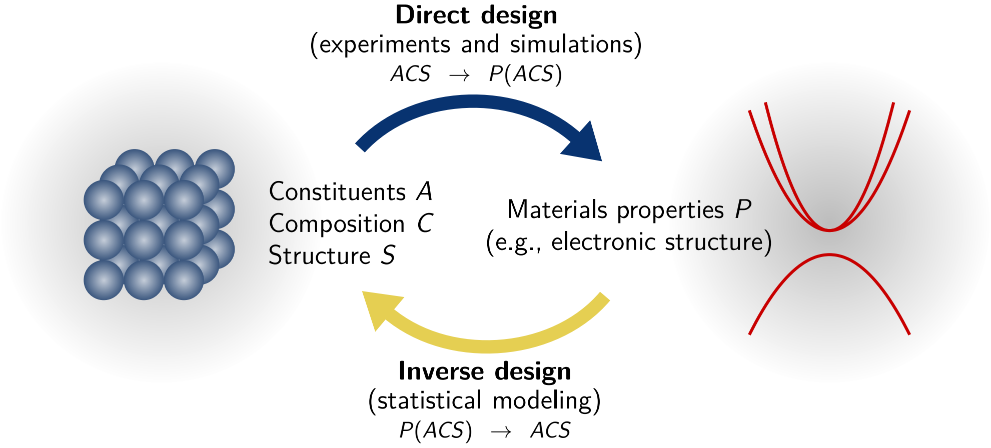


Fig. 2.1. Schematic representation of the direct and inverse design approach in materials science. Adapted from reference [47].

the physical problem at hand until fundamental equations can be derived from a more general physics-based model.

Data-driven approaches are about to become standard tools for scientists and engineers who are experienced in generating large amounts of data through experiment and theory. However, first-principles methods are comparatively demanding and experiments are relatively time intensive. Thus, accurate predictive models are required to reliably estimate the property of interest and to describe the underlying physical relationship from a limited amount of materials data. Ideally, these models are based on available parameters or properties of the material that are faster to compute and easier to obtain than the property of interest itself [55, 65].

2.1 Principles of materials design

There are two different approaches to searching for new materials: materials discovery and materials design. Both approaches identify candidate materials from a wide range of atomic configurations of known and hypothetical materials. The atomic configuration of a material is given by its constituents A (i.e, atoms Z_i and their positions $\vec{r}_i$), composition C (i.e., chemical composition $\vec{s}_i$), and structure S (i.e, lattice constants $\{a, b, c\}$ and angles $\{\alpha, \beta, \gamma\}$ for crystalline solids) [47]. Whereas materials discovery surveys a large space of candidate materials with desired properties of interest, materials design either optimizes the property of interest based on (synthesizable) candidate materials [39] or allows to determine the composition and structure of a material that possess a set of desired properties. By definition, materials discovery always precedes materials design: First, candidate materials need to be identified, before they can be optimized for actual applications.

Direct materials design

In direct materials design, the property of interest P is calculated based on the atomic configuration ACS of the candidate material (cf., Fig. 2.1). Using first-principles methods or experiments for determining the property of interest of candidate materials, the large majority of search methods probe the materials space of atomic configurations either by random sampling [113], evolutionary algorithms [114, 115], thermodynamic principles [116–121], or by molecular-dynamic simulations [122, 123]. Because the exploration of particular atomic configurations scales exponentially with the number of atoms N [124], a direct materials design with these methods may not always be practical. Be it because the search problem cannot be expressed in the required form for applying the particular method or the existence of local optima hinders a global optimization and as such the search for candidate materials with optimal properties.

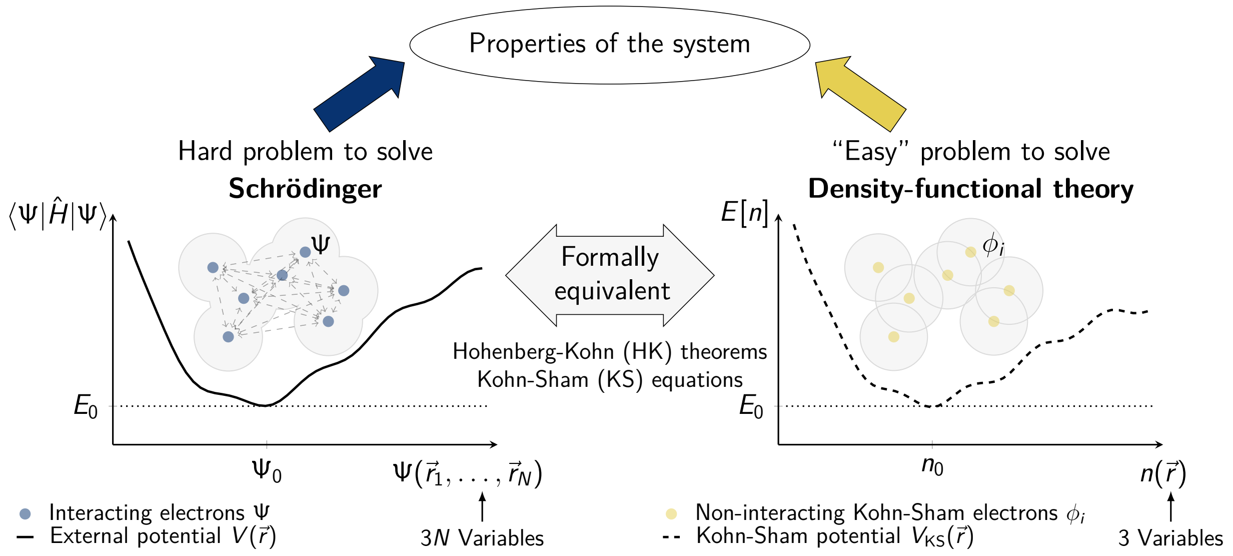
Inverse materials design

Instead of starting from composition and structures to predict the properties of a given material, inverse design [47, 125, 126] seeks to identify which compositions and structures produce materials that possess a set of desired properties (Fig. 2.1). This optimization problem involves two steps: The first step is to determine compositional and structural constraints to identify materials whose properties are close to the property of interest. And the second step is to iteratively refine the set of identified materials to match the specified requirements, either by screening material databases, conducting experiments, or by performing first-principles calculations. In practice, this design principle (re)uses available data to guide the search for candidate materials, thus involving very often (but not necessarily) fewer evaluations of first-principles calculations or experimental measurements than direct material design.

2.2 First-principles calculations

First-principle or *ab initio* methods are based on the atomic species, charges, and positions, i.e., the atomic configuration of a material and its structure [47]. They can be represented as a mapping $P : ACS \mapsto \mathbb{R}^d$ of the atomic configuration $ACS = \{a, b, c; \alpha, \beta, \gamma; \vec{r}_1, \vec{r}_2, \ldots, r_N; \vec{s}_1, \vec{s}_2, \ldots, s_N\}$ (Fig. 2.1) to materials behaviors (magnetism, superconductivity, etc.) and functionalities (chemical reactivity, etc.). Essentially, they determine the properties of materials by approximating the Schrödinger equation [69], which contains all the information to describe the microscopic properties of a system, the allowed energies, stresses, and forces.

A quantum-mechanical system containing N electrons and M nuclei has $4N + 3M$ degrees of freedom resulting from the $3N$ spatial coordinates, $\vec{r} = \{\vec{r}_{i=1\ldots N}\}$, and N spin coordinates, $\vec{\sigma} = \{\vec{\sigma}_{i=1\ldots N}\}$, of the electrons and the $3M$ spatial coordinates, $\vec{R} = \{\vec{R}_{i=1\ldots M}\}$, of the nuclei, respectively. Because the computational requirements of solving the Schrödinger equation increases exponentially with the number of electrons (and atoms), practical implementations involve a series of approximations. Within the Born-Oppenheimer approximation [127], the electronic relaxation to an external perturbation is assumed to be much faster than the ionic motion of the nuclei. As a result, electrons can be considered as basically moving in a constant external field generated by the positively charged nuclei at fixed positions. The electronic and nuclear contributions can be solved independently. For the total energy, the kinetic energy term of the nuclei can be set to zero, whereas the Coulomb repulsion term for the nuclei enters the total energy as a constant. The electronic contribution needs to be further


Properties of the system
Hard problem to solve
Schrödinger
"Easy" problem to solve
Density-functional theory
$\langle\Psi|\hat{H}|\Psi\rangle$
Ψ
E_0
Ψ_0
$\Psi(\vec{r}_1,\ldots,\vec{r}_N)$
Formally equivalent
Hohenberg-Kohn (HK) theorems
Kohn-Sham (KS) equations
$E[n]$
ϕ_i
E_0
n_0
$n(\vec{r})$
● Interacting electrons Ψ
— External potential $V(\vec{r})$
$3N$ Variables
● Non-interacting Kohn-Sham electrons ϕ_i
-- Kohn-Sham potential $V_{\mathrm{KS}}(\vec{r})$
3 Variables

$$\begin{tikzcd}
V(\vec{r}_1,\ldots,\vec{r}_N) \arrow[d] & n(\vec{r}) \arrow[l, dashed, "\mathrm{HK}"'] \arrow[r, dotted, leftrightarrow, "\mathrm{KS}"] & n(\vec{r}) \arrow[r] & V_{\mathrm{KS}}(\vec{r}) \arrow[d] \\
\Psi(\vec{r}_1,\ldots,\vec{r}_N) \arrow[r] & \Psi_0(\vec{r}_1,\ldots,\vec{r}_N) \arrow[u] & & \phi_{i=1,\ldots,N}(\vec{r}) \arrow[ul]
\end{tikzcd}$$

Fig. 2.2. The properties of a material can be determined by solving the Schrödinger equation of the electrons and nuclei (left). A formally equivalent method is the solution of the Kohn-Sham equations of non-interacting independent fictitious Kohn-Sham electrons (right). Although formally exact, approximations in density-functional theory limit the accuracy of the Kohn-Sham orbitals and hence the electron density. Commutative diagrams below each of the plots show the relations between the quantum-mechanical quantities: the wave functions Ψ and Ψ_0, the external potential V, and the electron density n obtained from the Schrödinger equation, on the one hand, and the Kohn-Sham orbitals ϕ_i, the Kohn-Sham V_{KS}, and the electron density n obtained from density functional theory, on the other hand. Adapted from reference [135].

approximated [110]. One such approximation is the Hartree approximation [128, 129], which treats electrons independently, i.e., as non-interacting one-particle electron systems, but violates the Pauli exclusion principle. The Hartree-Fock approximation [130–132] respects the Pauli exclusion principle by using an anti-symmetrized ansatz of the electronic wave function, but neglects correlations between electrons. More accurate approaches further account for the repulsion between the electrons (post-Hartree-Fock methods such as second/fourth order perturbation [133] or coupled cluster theory [134]), which however more and more computationally demanding with increasing number of electrons[1].

Rather than solving the electronic Schrödinger equation directly, density-functional theory (DFT) determines the ground-state properties of a system using a formally equivalent description of the Schrödinger equation as a spatially-dependent electron density of non-interacting fictitious electrons (Fig. 2.2).

[1] The precision of more accurate methods is associated with higher computational requirements. Even without considering the computational requirements of additional methods, a scaling law of $\mathcal{O}(N^3)$ already impedes DFT calculations for large systems ($N > 1000$ atoms).

Based on the Hohenberg-Kohn theorems[2] [25], the formalism of DFT states that the ground-state electronic wave function and resultant microscopic properties of an electron system are uniquely determined by the electron density alone. Within this formalism (as given by the Kohn-Sham equations of DFT [26]) the intractable many-electron problem of interacting electrons can be transformed into a set of N tractable one-particle problems of non-interacting electrons that are moving in an effective external potential (Fig. 2.2). The electronic kinetic energy and the electron-electron interaction of the many-electron problem of interacting electrons can be split into two terms: a non-interacting term of a fictitious electron system and a non-classical correction term, which contains all contributions arising from the interactions, exchange, and Coulomb correlation of electrons (via the so-called exchange-correlation functional[3]). In practice, the exact form of exchange-correlation functional is not known and therefore needs to be approximated, e.g., with the local-density approximation [26]. All the various existing approximations of the exchange-correlation functional can be summarized in the so-called Jacob's ladder [136], a hierarchy where the accuracy of the density-functional calculation increases with increasing computational costs of the exchange-correlation functional.

The electron density of the system can then be obtained by solving the Kohn-Sham equations. By virtue of the Hohenberg-Kohn theorem, all terms in the Kohn-Sham equations can be expressed as explicit functions of the electron density (with the exception of the kinetic term of the fictitious electron system). Therefore, to solve the Kohn-Sham equations, one starts from an initial estimate of the electronic density and solves the Kohn-Sham equations iteratively until convergence.

In the search of new materials, first-principles methods including DFT are typically used in a computational funnel: at each level, selection criteria (stability, costs, environmental sustainability, etc.) rule out materials and progressively more computationally intense methods determine candidate materials in a multi-objective optimization [137].

2.3 Materials databases

The rapid development of computer technologies and the improvement of computational algorithms has resulted in substantial time savings in performing first-principles calculations for hundreds or thousand of atomic configurations. The pace at which the properties of new materials are determined with latest generation of computers enables the systematic variation of atomic configurations, the screening of materials, and the storage of materials data in large and centralized materials databases (Tab. 2.1).

These materials databases grant access to completely new approaches for analyzing, screening, and predicting the properties of novel materials. Because data management places high demands on storage space and computing power, tools and concepts help to organize and standardize the collection, assimilation, and dissemination of materials data to and from materials databases. For example, software tools like pymatgen [151], the atomic simulation environment (ASE) [20], or the automated interactive infrastructure and database (AiiDA) [21–23] facilitate the generation of

[2] The first Hohenberg-Kohn theorem states the ground-state electronic wave function can be uniquely described by the electronic density. The second theorem states that the ground-state energy can be obtained by minimizing the electronic energy with respect to the electron density.

[3] A functional takes functions as arguments rather than a list of variables.

Name and URL	Description	#Materials	Refs
Theoretical and computational structures and properties			
AFLOWLIB (http://aflowlib.org)	High-throughput materials repository for electronic-structure and property calculations with online applications for automating first-principles calculations.	~3 000 000	[16, 17, 33]
Computational Materials Repository (https://cmr.fysik.dtu.dk)	Repository for collecting, storing, and retrieving data from electronic structure codes and property calculations from a diverse set of applications.	~60 000	[32, 138, 139]
Materials Project (https://materialsproject.org)	Online platform for materials exploration on known and predicted materials, which provides open-source analysis tools for materials discovery and design.	~700 000	[13, 15, 35]
NOMAD (https://encyclopedia.nomad-coe.eu/gui/)	Repository to host, organize, and share materials data of a wide range of electronic structure codes through a unified metadata language and data analytics tools.	~10 000 000	[37, 140]
Open Quantum Materials Database (http://oqmd.org)	Database with calculated thermodynamic and structural properties of inorganic crystal structures.	~560 000	[18, 19, 36]
NREL Materials Database (https://materials.nrel.gov)	Computational materials database for renewable energy applications.	~230 000	[34, 141–143]
Citrination (https://citrination.com)	A materials informatics platform combining materials data and analytics tools for materials development and design.	~n/a	[144, 145]
Experimental structures and properties			
Crystallography Open Database (http://crystallography.net)	Crystal structures of organic, inorganic, metal-organics compounds and minerals, excluding biopolymers.	~410 000	[30, 146, 147]
ICSD (https://icsd.fiz-karlsruhe.de)	Database of experimental inorganic, metal-organic, and theoretical structures.	~210 000	[148, 149]
SpringerMaterials (https://materials.springer.com)	Curated, comprehensive, and multidisciplinary collection of materials and chemical properties with extensive coverage of all major topics in materials science and related disciplines.	~300 000	[31]

Tab. 2.1. A list of publicly accessible materials databases and data infrastructures (commercial and non-commercial). A more detailed overview is available in reference [150]. Numbers of materials data reflect the status as of September 2021.

hundreds of new first-principles computations. Managements tools such as fireworks [14] distribute these tasks on high-performance computers with massively parallel architectures, while data concepts such as the FAIR principle [11], in addition, help ensure that materials data are findable, accessible, interoperable, and reusable [10].

To identify materials with the desired properties, a database search is performed based on filter criteria of available properties of a material. These can be observables (e.g., atomic properties), parameters of the system (e.g., environments, temperature, pressure), or materials properties at various length scales and of different degrees of complexity (e.g., thickness of the material, resistivity, elasticity). Because these quantities do not necessarily have to be derived from first-principles calculations, statistical tools such as machine learning can be used to reduce or bypass time-intensive steps in the materials search at the expense of an accurate physical understanding of the problem.

2.4 Machine learning

Data-driven approaches rationalize materials discovery and design through effective methods for generating atomic configurations, managing materials databases and utilizing materials data to direct materials search. These statistical approaches have been used in biology [152] and chemistry [153] for decades. Recently applied to materials science [7, 154], they have shown promise as a tool in modern materials simulations [42, 112, 155–160], e.g., for the crystal-structure prediction of octet-binary compound semiconductors [55, 161], prediction of band gaps and formation energies [66, 162–164], elastic constants [165, 166], superconducting temperatures [167, 168], renewable-energy materials [169], stability of materials [170, 171], or the identification of spin-driven thermoelectric materials [172]. Therefore, first-principles methods are increasingly combined with approximate methods to estimate the properties of a material, such as using machine learning to screen larger materials spaces.

Machine learning, which is a branch of artificial intelligence[4], is capable of building statistical models to estimate a property of interest. The term “machine” means that such models can be created automatically without human intervention. The term “learning” refers to the model improving with more data. Machine learning can be either supervised or unsupervised. Unsupervised machine learning analyzes the data to find groups of similar materials, transforms the data to reduce the number of variables in a model (dimensionality reduction), or can identify statistical outliers in the data. Supervised machine learning instead optimizes the outcomes of a statistical model based on predefined objectives. Both types of machine learning can be used to reduce the complexity of a materials search. Whereas unsupervised learning is often used to visualize the different classes of materials, supervised learning is used to estimate the property of interest based on a large number of materials properties.

Machine learning is primarily used in inverse materials-design approaches, e.g., to screen well-defined chemical-structural classes of materials for desired properties. Provided that the data set is representative in the sense that yet unexplored materials are in the same chemical-structural class

[4] The term “artificial intelligence” was coined by John McCarthy for a conference in 1956 at which the logic theorist [173, 174] was presented as the first program of artificial intelligence. Written by Allen Newell, Herbert A. Simon, and John Shaw it was deliberately designed to perform automated reasoning to prove 38 of the first 52 theorems in Whitehead’s and Russell’s Principia Mathematica.

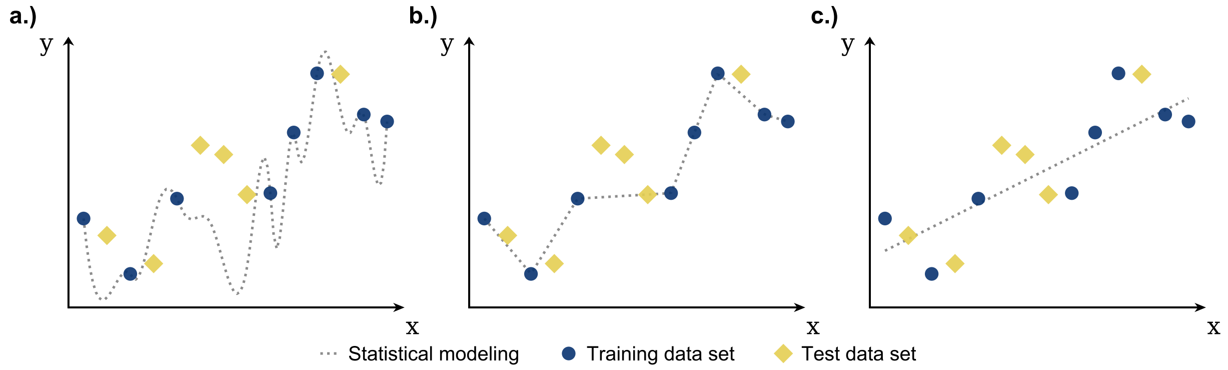


Fig. 2.3. A fit of three different functions (dotted lines) to the same set of training data (dots). While the polynomial (a) and piecewise-linear (b) function exactly reproduce the training data (but not the test data), a machine-learning algorithm like linear regression constructs a statistical model $\hat{f}_{\text{linear}}$ that minimizes the error ε of the fit $y = \hat{f}_{\text{linear}}(x) + \varepsilon$ based on the test data not being used for model construction.

as the materials in the data set, creating a mapping between the properties of a material and the property of interest can be reduced to building a supervised machine-learning model to accurately estimate the property of interest [56]. The mapping can then be used to make predictions, without having to perform expensive experiments or first-principles calculations through the use of tabulated and easily accessible materials properties defined at varying levels of complexity [55, 66].

Mathematically, such a supervised-learning problem can be expressed as a function f that maps a set of input variables $\vec{X} = \{X_1, \ldots, X_d\}$ to the property of interest $\vec{Y}$, i.e., $f : \vec{X} \mapsto Y$. Input variables that determine materials behavior are referred to as features, while combinations or derived features in a statistical model are called the fingerprints [112, 175] or the descriptors [55, 161, 176] of a material.

The function f is exact. In actual materials-science applications though, where it is expected that the fundamental equations are not known, the function f is inherently approximate. First, because calculations and measurements are subject to error. And second, the data set may not include all of the features necessary to relate the essential physics and chemistry to the property of interest. The mapping of f is therefore modeled as a statistical relationship $\hat{f}$ (referred to as the machine-learning model) with respect to a property of interest Y,

$$Y = \hat{f}(\vec{X}) + \varepsilon \, , \tag{2.1}$$

where ε is an error term assumed to be independent of $\vec{X}$ [51, 52]. At the core of the statistical modeling is the approximation of the exact mapping f with the machine-learning model $\hat{f} : \vec{X} \longrightarrow \hat{Y}$: The machine-learning model is optimized to reduce the prediction error (e.g., $|Y - \hat{Y}|$) based on a set of materials data, that have not been used for model construction (Fig. 2.3).

Unlike fitting, which just interpolates between data samples (Fig. 2.3a-b.), machine learning involves fitting and regularization to build a model of the overall statistical trend in the data (Fig. 2.3c.). Because these models often come with inherent assumptions, they have additional parameters that can be tuned to optimize predictions on new data (via a so-called hyper-parameter optimization), such as simplifying the functional form or the number of evaluations, i.e., the complexity of the generated machine-learning model.

2.5 A typical machine-learning pipeline for materials science

The statistical modeling of feature-property relationships needs curated materials data. Provided there are sufficient data from first-principles calculations and experiments, the modeling with machine learning proceeds in three steps: data acquisition, data representation, and data modeling (Fig. 2.4). Each step is interdependent and often requires multiple iterations to build accurate predictive models.

Data acquisition and curation

In the data-acquisition step, materials data are pre-processed. They are checked for accuracy, inconsistencies, or missing materials. Because the generation of materials data from first-principles calculations or experiments is an expensive and time-consuming process, a representative materials data set for the investigated chemical-structural class of materials includes as few materials as possible, while covering a wide range of atomic configurations or compositions. It may be that more materials data are needed at a later stage or that the prediction performance of the generated statistical model is not sufficient. In these cases, experiments or first-principles calculations are performed to expand and iteratively refine the data set.

Materials representation and transformation

In the materials-representation step, each material in the data set is mapped to a set of features that are faster to compute and easier to obtain than the property of interest and allow to reverse the materials-design process [55, 65]. Ideally, these features quantitatively relate the essential physical and chemical properties of atomic configurations or compositions to the property of interest and can be used to uniquely characterized the specified materials.

Which set of features best model the property of interest, is specific to the property of interest and the materials classes under study [56, 177]. A wide range of physical and chemical attributes can be considered for this purpose, such as the structure, composition, and properties of the constituent elements [55, 162, 178–181], or the quantitative representation of the atomic environment [166, 181–187].

Using these attributes, features can be constructed for materials with any number of constituents, including statistics on stoichiometric, elemental, or electronic-structure properties [56, 188, 189]. However, the more features a data set contains, the higher the computational requirements for constructing statistical models. Multi-collinearity, i.e., the interrelationships between features, may further degrade the prediction performance of a machine-learning model. As such, a candidate list of features is often selected based on expertise, intuition, and trial-and-error. Because the materials representation has a great impact on the prediction performance of a machine-learning model (Fig. 2.4), this step requires dedicated methodological approaches such as genetic programming [190], compressed sensing [191], or information science [192] to identify features that are relevant for estimating the property of interest.

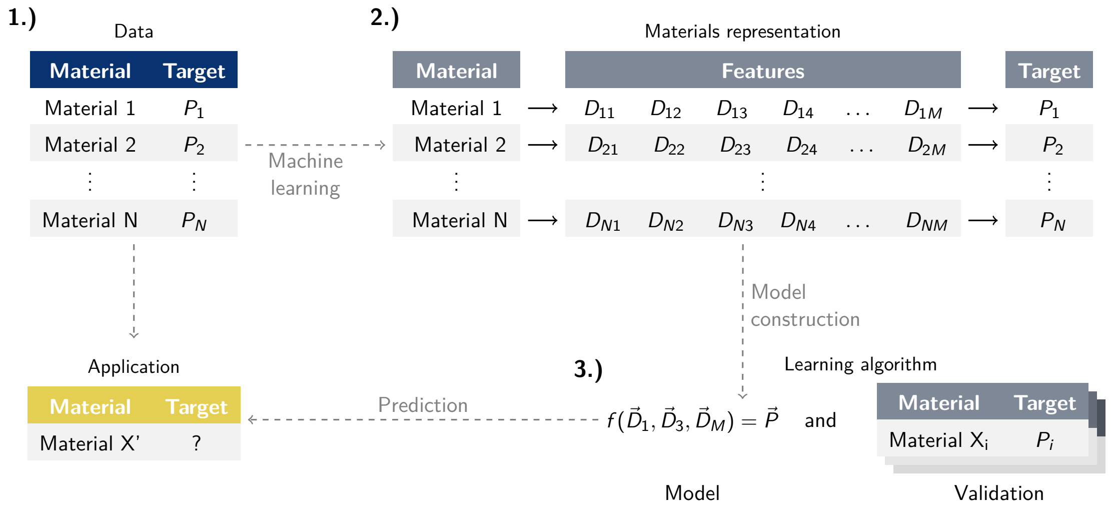


Fig. 2.4. The machine-learning pipeline in materials science starts with available and curated materials data (1), continues with the representation of materials (2), the systematic construction of machine learning models (3), and ends with accurate predictions of desired materials properties $\vec{P}$. Adapted from reference [112].

Data modeling and machine-learning algorithms

In the data-modeling step, a machine-learning model is constructed based on the candidate list of features. The machine-learning model is then used to estimate the property of interest for new materials. The most popular algorithms for generating machine-learning models are linear regression [51], kernel-ridge regression [193, 194], support-vector machines [195, 196], Gaussian-process regression [108, 109], symbolic regression based on compressed sensing [51, 52, 55, 161, 191, 197], decision trees [198, 199], and deep neural networks [12, 105–107, 200]. The actual choice depends on computational requirements, which kind of mapping relates the features to the property of interest, and whether the property of interest is continuous (regression problem) or discrete (classification problem) [51, 52].

An integral part of the data-modeling step is the evaluation of the machine-learning model on data, that have not been used for model construction: either by partitioning the data set into subsets or by applying the model on new data. A common technique to partition the data into subsets (training and validation) is resampling. Resampling [52] builds a model for each training data set and evaluates it on the validation data set. Two well-known resampling methods are cross-validation [52, 201] and (out-of-sample) bootstrapping[5] [203–205].

Both methods provide estimates of the model's predictive ability that would not be available from building the model once using the initial data set. There are different types of cross validation (cf., [201, 206]). The most frequently used cross-validation technique is k-fold cross validation. In k-fold cross validation, data are systematically partitioned into k complementary subsets (sampling

[5] The term "bootstrapping" can be traced back to a newspaper from the Workingman's Advocate in 1834 [202], which serves as a figurative paraphrase for an apparently impossible task of pulling oneself up by the straps of one's boots. In the context of statistics and machine learning, bootstrapping means to iteratively improve the accuracy of a particular estimator or statistical-learning method.

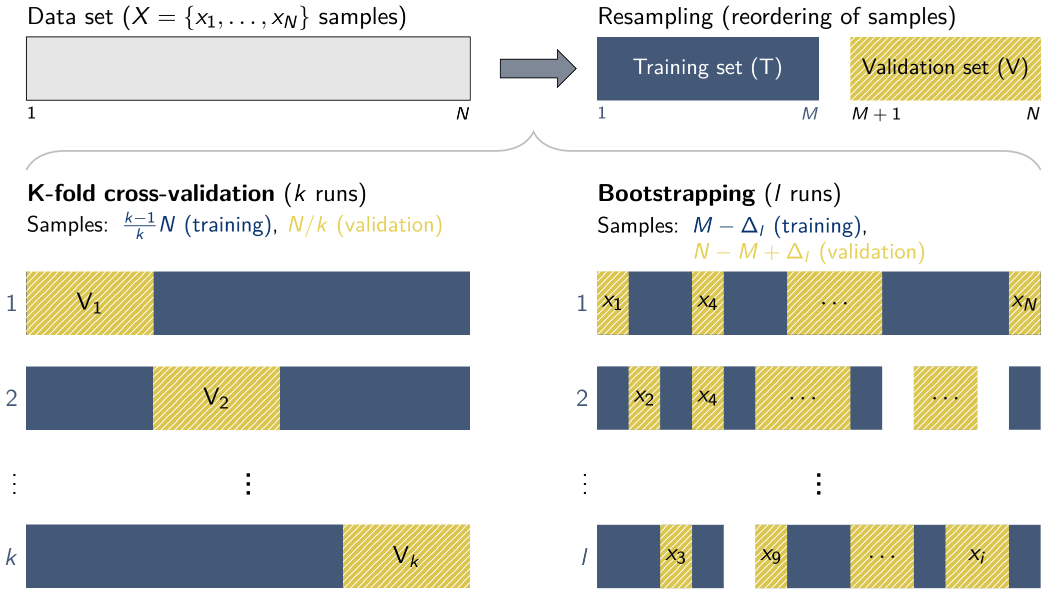

Data set ($X = \{x_1, \dots, x_N\}$ samples)
1
N
Resampling (reordering of samples)
Training set (T)
Validation set (V)
1
M
M + 1
N
K-fold cross-validation (k runs)
Samples: $\frac{k-1}{k}N$ (training), N/k (validation)
1
V1
2
V2
⋮
k
Vk
Bootstrapping (l runs)
Samples: $M - \Delta_l$ (training), $N - M + \Delta_l$ (validation)
1
x1
x4
⋯
xN
2
x2
x4
⋯
⋯
⋮
l
x3
x9
⋯
xi

Fig. 2.5. Common resampling techniques for estimating the prediction performance, i.e., the prediction error, on new data: cross-validation [52, 201] (bottom left) and bootstrapping [203–205] (bottom right). Both techniques split the available data into a training (T) and a validation data set (V). While the training set is used to construct a machine-learning model, the validation set is used to evaluate the model's predictive performance. Training and validation sets can be partitioned, e.g., into k complementary subsets (k-fold cross-validation) or drawn from the data set with replacement (bootstrapping). Due to sampling with replacement, not all samples from the initial data set may be included in the bootstrap partitions.

without replacement), where $k-1$ subsets are used for building the model, and the remaining subset for validating the model's prediction performance. In the bootstrap method, samples are drawn randomly with replacement for building the model, while using the remaining samples for validating the model. However, the bootstrap method (Fig. 2.5b) results in higher estimates of the prediction error than k-fold cross validation (Fig. 2.5a). Furthermore, using it requires the generation of a larger number of partitions to evaluate a model's prediction performance and hence the bootstrap method is more computationally demanding than k-fold cross validation. K-fold cross validation is therefore typically used to estimate the prediction performance of a model by combining (e.g., averaging) the validation results from all partitions. Bootstrapping, in contrast, is used for purposes other than validation; in particular, to estimate a wide range of summary statistics from the set of independent samples of the data set, to quantify the uncertainty of a statistical variable or model, or to compute confidence intervals associated with a particular estimator or statistical-learning method.

2.6 Challenges and limitations of machine learning in materials science

One of the major criticisms of machine learning is the use of machine-learning models as black boxes and the resulting lack of insights in guiding the search for and designing new materials. The fundamental challenge in identifying materials with desired properties is therefore to build an accurate statistical model to describe the underlying physical relationship in terms of features that are related to the property of interest (i.e., the relevant features of the data set).

One of the limitations of applying machine learning to materials science is the availability of data sets large and diverse enough to represent the classes of materials under study. Typically, published data sets are confined to particular atomic configurations or contain only a subset of materials from the search of materials with desired properties [8, 207]. Furthermore, due to the inherently statistical nature of machine learning, there is a high risk of applying these models to the prediction of materials properties for which there was insufficient data for statistical modeling [112, 208]. Applying machine-learning models to new data, therefore requires at least an estimate of the error made in a single prediction [159, 209, 210] or to define a domain of applicability of the machine-learning model [208]. Unfortunately, this pragmatic view on the modeling of feature-property relationships is not widespread in materials science because the estimation of uncertainties in the model predictions (cf., Section 4.3) or the definition of a domain of applicability are far from trivial. A paradigm shift is required to generate these models and uncertainties in a standardized framework. Rather than developing machine-learning models and combining multiple methods to optimally predict the property of interest for each application, a framework using existing and newly developed methods is designed in the following chapters that is independent of the machine-learning method itself.

2.6.1 Feature identification

When machine learning is applied to data sets with many features, the computational cost of the learning algorithm dramatically increases, and in the case of multi-collinear features, the predictions of machine-learning models can degrade significantly [84, 99, 160]. This is often referred to as the "curse of dimensionality" [211]. The curse of dimensionality describes the interplay between the number of samples and features, the relationships between the features, and the complexity of the machine-learning model. The smaller the data set and the more features it contains, the stronger the impact of the curse of dimensionality on the prediction performance of the generated machine-learning models. The focus of the thesis is therefore to identify the set of features that are related to the property of interest. This problem is addressed by reducing the number of features in a data set prior to statistical modeling.

Despite the large diversity of approaches in the literature, there is no systematic route to identify relevant features and to construct accurate machine-learning models from a large set of features. Although there are first approaches of automated methods without the need to curate a list of relevant features for the prediction of material properties [15, 212, 213], they require prior knowledge of the material's structure. Other methods [39, 214, 215] are designed to predict the material's structure, but are only applicable to a single or a limited number of materials classes. Finally, many statistical-modeling algorithms (such as compressed-sensing implementations [55, 197] or regression-tree-based approaches [216]) implicitly select features that are, however, difficult to relate to the property of interest and can lead to erroneous scientific conclusions when used with many features. This thesis therefore investigates possible approaches based on information theory, but also develops methods that use both information theory and machine learning in a common search strategy to identify the relevant features for estimating the property of interest.

2.6.2 Uncertainty estimation

Machine-learning models are inherently statistical. Further, implicit model assumptions or the limited flexibility of the machine-learning models may negatively impact the prediction performance. Because these models estimate the property of interest pointwise based on a set of features, but commonly the prediction performance is only estimated globally, it is not clear how well they approximate the actual value of the property of interest of a new material. The materials-science examples in this thesis show that predictions of machine-learning models can differ significantly from the actual value. Globally estimated prediction performances, e.g., with resampling methods, are therefore not suitable for assessing individual predictions with machine learning.

Uncertainty estimates provide an interval for the error made in each individual predictions. Gaussian process regression [108, 109] and deep neural networks [105–107] are two commonly used machine-learning algorithms that provide uncertainty estimates either based on Bayesian probabilities or on distributions from the internal parameters of the model. However, these methods only estimate the uncertainty in the model's prediction, but not the difference between the model's prediction and the actual value of the property of interest. Moreover, with these methods it is difficult to identify individual materials that cannot be accurately predicted by the machine-learning model. In this thesis, conformal prediction [62–64] is used to determine precise levels of confidence for machine-learning predictions within a probabilistic tolerance and a heuristic is developed to identify individual materials which cannot be predicted well by the machine-learning models.

Chapter 3

Optimizing materials representation: techniques for identifying relevant features

The set of features for predictive modeling of materials data from first-principles calculations or experiments is usually selected by knowledge, intuition, and many trials; often without demonstrating that these features are actually related to the property of interest [155, 217–220]. Ghiringhelli, Vybiral *et al.* [55, 161] therefore drew attention to the critical role of the features: By formulating the identification of feature combinations related to the property of interest as a compressed-sensing problem [191, 219, 221–224], Ghiringhelli, Vybiral *et al.* [55] used the least absolute shrinkage and selection operator (LASSO) [225] to pre-select candidate-feature combinations and to construct mathematical models with symbolic regression and the lowest number of candidate features (LASSO+ℓ_1 regularization followed by an ℓ_0-regularized symbolic regression). The number of feature combinations and the correlations within, however, limit this approach to data sets with a maximum of about 10–15 features, if no pre-selection of promising candidate-feature combinations is to be made [55]. In Refs. [176, 197] the same idea was employed by using the sure-independence screening and sparsifying operator (SISSO) [197]. SISSO iteratively selects candidate-feature combinations, effectively reducing the residual error to the property of interest (cf., Appendix A.1). A symbolic regression model is then constructed based on these candidate-feature combinations through a linear least-squares fit. In contrast to LASSO, SISSO can deal with (inter-)correlated features and billions of candidate-feature combinations. However, since both compressed-sensing methods rely on combining features, they can quickly become computationally demanding, even for a small number of initial features of the data set.

Given the complexity of feature-property relationships and the infinite number of possible materials, so far no general and systematic approach has been established to reduce the set of initial features $\vec{X}$ by identifying the relevant set of features $\vec{X}' \subseteq \vec{X}$ to the property of interest Y.

This chapter, therefore, addresses the representation of materials and the challenges in developing a feature-selection workflow for the identification of features that are related to the properties of interest (i.e., the relevant features of the data set). The fundamental challenge is to find the features related to a certain property or function using a score, which allows the identified features to be ranked and ordered by relevance prior to applying machine learning for generating statistical and predictive models. The chapter starts with an outline of dimensionality-reduction techniques [226, 227], i.e., techniques to reduce the number of features of the data set prior to relating the features to the property of interest. These techniques are independent of a mathematical model and thus can reduce the number of features and the computational requirements of subsequent analyses, as in machine learning. Dimensionality reduction is carried out either through the creation and utilization of combination of features[1] (feature extraction [227]) or the selection of a subset of features directly from the set of the data (feature selection [71, 72]).

Feature selection can be seen as a combination of a search technique for proposing new feature subsets and a feature-selection measure for scoring the subsets. Various feature-selection algorithms have been proposed [228–230] and several feature-selection measures have been extensively explored [231–235], but none of these methods is applicable to real-valued features such as encountered in materials science without introducing additional parameters in the relevance quantification of feature subsets. Because feature selection is an optimization problem that resembles that of information-theoretic concepts [236], a non-parametric, deterministic information-theoretic method is developed, called total cumulative mutual information (TCMI) [90]. TCMI is designed to quantify the cumulative mutual dependence between a set of features and the property of interest for real-valued features.

To identify the subset of features related to the property of interest, TCMI needs to be applied to each feature-subset combination in the data set. The score obtained by TCMI can then be used to rank the feature subsets in descending order of their strength of cumulative mutual dependence, with the strongest cumulative mutual dependence determining the most relevant set of features. There are several search strategies to identify the optimal feature subset [92, 229, 237–239], ranging from exact solvers such as an exhaustive search, to heuristic, and stochastic solvers [72, 91–96, 229, 236, 237, 240–243]. Exact solvers are generally impractical for data sets with a large number of features (due to the exponential time needed to solve the optimization problem). Sub-optimal solvers (greedy, heuristic, or stochastic algorithms), by contrast, may not find the optimal subset of features related to the property of interest. Branch-and-bound [91–96] can be used as an optimal or sub-optimal solver [92, 239]. TCMI is therefore used with branch-and-bound to identify the features related to the property of interest, and its performance is compared with existing methods for identifying relevant features.

3.1 Dimensionality reduction

The identification and characterization of materials-property relationships between the features of the data and the target property is one of the fundamental challenges in computational materials science (Section 2.6). Especially in high-dimensional materials spaces [244–246], which cannot practically be explored exhaustively, a reduction of the initial number of features can often alleviate the challenging phenomenon of the “curse of dimensionality” [211] and speed up the model construction [247] either by selecting a small subset of features or by projecting features onto a lower dimensional materials space. Such dimensionality-reduction techniques are an essential part of a successful data-mining pipeline for efficient feature-space exploration, visualization, and an understanding of the statistical relationships present in the data.

[1] So-called derived or transformed features.

3.1.1 Feature extraction

Describing the relationship of a specific property or function in terms of a set of available physical parameters or materials properties (i.e., the features of the data set) is central to creating accurate predictive models. However, it is often not clear which features are related to the property of interest, nor is it feasible to manually enumerate all features for the construction of machine-learning models with the lowest prediction error.

There are not many machine-learning algorithms that can identify relevant features for the statistical modeling. Kernel-ridge regression [193, 194] and Gaussian-process regression [108, 109], for instance, cannot. The class of methods that transform the data for effective machine learning and implicitly identify candidate features from the initial set of features, is called representation learning [248]. Two examples are deep neural networks [12, 105–107, 200] and symbolic regression (whether via stochastic optimization [51, 52] or compressed sensing [55, 161, 197]). For this reason, numerous available physical parameters and materials properties are assembled and used to construct machine-learning models.

This approach has two major drawbacks: First, only a small number of features may actually be related to the property of interest, and second, many features may be statistically interrelated, resulting in both sensitive and unnecessarily complex machine-learning models and higher demands on computational resources. One approach to reduce the number of features is feature extraction[2]. Feature extraction projects the high-dimensional problem onto a lower dimensional (latent) materials space (Fig. 3.1). It involves the generation of a new set of transformed features based on combinations of the initial features of the data set. Many machine-learning algorithms implicitly perform feature extraction during model construction. Common supervised feature-extraction methods are the aforementioned deep neural networks and symbolic regression, but also gradient-boosting decision trees [249–252]. Common unsupervised feature extraction methods are principal component analysis [73–75]), multi-dimensional scaling [76, 77, 80, 82], and manifold learning [78–82]. Since the newly created features are expected to retain the relationships of the initial set of features, feature extraction is often used for data visualization [80–82, 175] and for building predictive models from the reduced representation instead of the full initial set of features (Fig. 3.1).

3.1.2 Feature selection

A materials representation is based on available training data as well as a performance measure and may contain features or combinations of features not related to the property of interest. Whereas feature extraction transforms the initial set of features into a new reduced feature set, feature selection [71, 72] identifies a subset of features of the data set with respect to a property of interest without changing them. Feature selection thus has the advantage over feature extraction of quantifying the relevance of individual physical parameters or materials properties of the initial set of features relative to the property of interest [70]. Feature selection is based on the terms of relevance and redundancy [72, 253, 254]: whereas redundant features are related to other features in the data set and hence can

[2] Feature extraction is also referred to as feature construction [236] or dimensionality reduction [227].

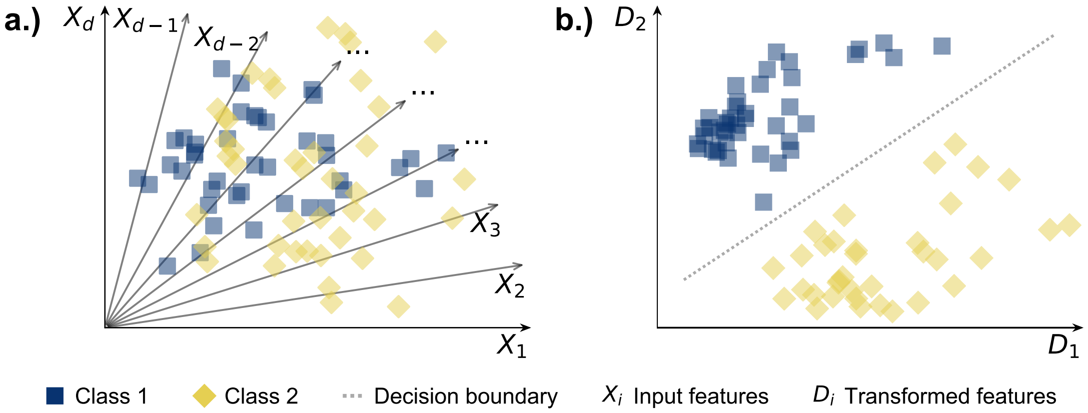


Fig. 3.1. Dimensionality reduction reduces the high-dimensional materials space (a.) to a lower dimensional materials space (b.). Shown is a classification problem using feature extraction, which transforms the input features $\vec{X} = \{X_1, \ldots X_d\}$ into a combination of features set $\vec{D} = \{D_1(x_1, \ldots, X_d), D_2(x_1, \ldots, X_d)\}$ and thus improves the separability of the two materials classes.

be replaced with other features in the model, relevant features cannot be removed without degrading the prediction performance of a machine-learning model[3].

Feature-selection methods reduce the risk of modeling spurious relationships between the features and the property of interest by discarding features that are not related to the property of interest. They work on problems with a large number of features and lead to an efficient representation of the data set with reduced computational costs of machine-learning algorithms and less complex predictive models [236]. Formally, the problem of feature selection is to find an optimal subset $\vec{X}^* \subseteq \vec{X}$ of input features $\vec{X} = \{X_1, \ldots, X_d\}$ with cardinality d subject to maximizing a feature-selection criterion $\mathbb{Q}$ with respect to a property of interest Y,

$$\mathbb{Q}(Y; \vec{X}^*) = \max_{\vec{X}' \subseteq \vec{X}, |\vec{X}'| \leq d} \mathbb{Q}(Y; \vec{X}') \quad \Leftrightarrow \quad \vec{X}^* = \operatorname*{arg\,max}_{\vec{X}' \subseteq \vec{X}, |\vec{X}'| \leq d} \mathbb{Q}(Y; \vec{X}') \; . \tag{3.1}$$

As such, feature selection combines a search strategy for exploring feature subsets with an evaluation measure for scoring the feature subsets.

Search strategies

The simplest approach to test each possible set of optimality from the initial set of features $\vec{X}$ is an exhaustive search. An exhaustive search strategy explores the complete set of feature combinations[4] $\sum_{k=1}^{d} \binom{d}{k} = 2^d - 1$, and as such is prohibitive for high-dimensional data sets due to cost and time constraints of computer resources[5] [72]. Therefore, many sub-optimal feature-selection methods have been proposed to find feature subsets close to the optimal representation while balancing accuracy and speed. For example, an intuitive approach adopted by the sequential forward selection [95,

[3] The set of relevant features is also called the Markov blanket of a data set [255].

[4] The empty subset is excluded in feature selection.

[5] The problem of searching for the optimal feature subset is non-deterministic polynomial-time ($\mathcal{NP}$)-hard [256, 257].

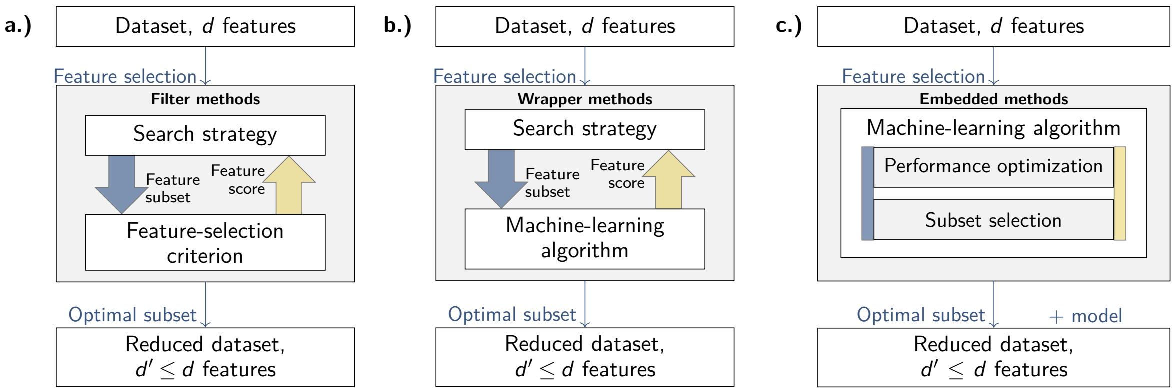


Fig. 3.2. The three classes of feature-selection methods differ in how the relevance of feature subsets is estimated. While filters (a.) use criteria that are independent of the machine-learning algorithm to assess the relevance of a feature subset, wrappers evaluate the relevance of a feature subset based on a predefined machine-learning algorithm (b.) and embedded methods simultaneously optimize materials representation and prediction performance of a machine-learning model (c.), e.g., through regularization [51, 225, 261].

229, 240] or backward elimination [241] starts with either an empty or a full subset of features and gradually adds or removes one feature at a time, respectively. Other strategies, such as greedy or randomized search, generate new feature subsets based on both input features and those already selected [190, 237] and are widely used for the simulation of physical processes in metallurgy and biology [159, 214, 258, 259].

An optimal search strategy that has not yet received much attention, is branch-and-bound [91–96]. Branch-and-bound combines both optimality and speed [92, 239] and guarantees to find the global or local optimum (i.e., by limiting the search depth or space to the k-best performing feature subsets). Because branch-and-bound implicitly performs an exhaustive search, it is guaranteed to find the optimal feature subset. In contrast to sub-optimal strategies, the branch-and-bound algorithm requires a monotonically increasing feature-selection criterion for finding the optimal feature subset by combinatorially enumerating the set of all features and stopping at feature subsets, whose feature-selection criterion cannot be improved (cf., Section 3.4).

Feature-selection criteria and classification of feature-selection methods

Different feature-selection criteria impose different requirements on the optimality condition of a feature subset, which is why the optimal feature subset of one feature-selection criterion may differ from that of another criterion and are therefore often difficult to compare [236]. All feature-selection algorithms can be categorized into three groups, depending on how the relevance of a feature subset is estimated [72, 260]: filters, wrappers, and embedded methods (Fig. 3.2).

Filter methods Filter methods [236] identify and evaluate feature subsets solely on the basis of available data and independent of a machine-learning algorithm. They are used as a processing step to quantify the relevance of feature subsets and to rank them by their strength prior to statistical modeling (Fig. 3.2a). Because filter methods use different optimization criteria than machine-learning models generated from these subsets, supervised machine-learning models may not always reach the lowest possible error in estimating the property of interest [236]. Examples include similarity [260], statistical [101, 262], distance [228, 231, 232], dependency [233], consistency [234], and information measures [230, 235].

Wrapper methods Wrapper methods [72] evaluate the relevance of a feature's subset based on the prediction performance of a supervised machine-learning model generated with this subset (Fig. 3.2b). As a result, wrapper methods identify the best-performing feature subset for a supervised machine-learning algorithm based on a given search strategy [228, 236, 260]. Common wrapper methods are recursive feature elimination [263] and genetic algorithms [122, 238, 264].

Embedded methods Lastly, embedded methods [265] simultaneously perform feature selection and modeling by optimizing the objective function of the machine-learning algorithm (Fig. 3.2c). The most common machine-learning algorithms used as embedded methods are the least-absolute shrinkage and selection operator [225, 261] (and recently the sure-independence screening and sparsifying operator [197]) and decision-tree algorithms such as random forest [266] or gradient boosting [249–252, 267, 268] (cf., Appendix A).

3.2 Materials representation with information theory

The construction of an optimal set of features requires feature-selection methods that are able to identify a wide range of relationships in the data. Ideally, such feature-selection methods estimate the relevance of a set of features from a limited number of data samples and rank the features according to their relevance for predicting the property of interest [236].

Targeted to best identifying the set of features for predicting the property of interest, embedded filter-selection methods automatically perform feature selection during model construction, while wrapper methods evaluate the relevance of a set of features tailored to a specific machine-learning algorithm (Fig. 3.2). In data sets with a large number of features [244–246], however, wrapper and embedded methods are computationally demanding[6]. In contrast, filter methods decouple the feature-subset search from the model generation, often providing a means to identify feature subsets based on a feature-selection criterion within a rigorous mathematical theory of feature relevance. Filter methods can therefore in principal be applied to data sets with a large number of features, in contrast to wrapper or embedded methods, which are mainly computationally limited (i.e., they are more expensive than filter methods) by the underlying machine-learning algorithm for evaluating the relevance of feature subsets.

Not all filter-based feature-selection criteria are suitable for materials-science problems, especially when multiple features are jointly related to the property of interest. For example, feature-selection criteria such as Pearson's R or Spearman's rank ρ correlation [101, 269], and distance-correlation measures [270–272] are limited to bivariate relationships and can only identify specific types of dependencies (e.g., monotonic relationships). In contrast, feature-selection criteria based on kernel

[6] This is often referred to as the "curse of dimensionality" [211].

$$\begin{tikzcd}
\vec{X} \arrow[r, dashed, "\exists f"] \arrow[d, "p"'] & Y \arrow[d, "p"] \\
p(\vec{x}) \arrow[r, shift right] \arrow[r, leftarrow, shift left, "g"] & p(y)
\end{tikzcd}$$

$$p : \mathbb{R}^d \to I, \;\; I = [0, 1]$$

$$g : I \to I, \begin{cases} p(y) = \int p(\vec{x}, y)\ d\vec{x} \\ p(\vec{x}) = \int p(\vec{x}, y)\ dy \end{cases}$$

$$\vec{X} = \{X_1, \ldots X_d\}, \;\; \vec{x} \in \vec{X}, \;\; y \in Y$$

Fig. 3.3. Commutative diagram of a materials representation between features $\vec{X}$ and a property of interest Y and its alternate description via marginal probability distributions, $p(\vec{x})$ and $p(y)$, related by the map g and defined by their joint probability distribution $p(\vec{x}, y)$.

density estimation [273, 274] or k-nearest neighbor estimation [275] are intrinsically dependent on the scale of the features and implicitly assume that relationships between features and the property of interest are smooth and differentiable. The prevailing filter-based feature-selection criterion is mutual information (Sec. 3.2.3). Mutual information [83, 84] describes the features and the target property as probability distributions (Fig. 3.3) to quantify the statistical dependence between a set of features and the property of interest [235]. Mutual information is based on the concept of entropy from information theory [83, 84] and has found applications in many areas of science ranging from quantum information [276] and the understanding of black holes in physics [277] to genetic structures in biology [278], electronic structures or chemical reactivity in chemistry [279], and machine learning [280, 281].

Most of the feature-selection criteria including mutual information are only defined for discrete and categorical features such as quantum numbers or materials classes. However, in many applications (such as in materials science) most features are real-valued, continuous quantities[7]. Although extensions have been proposed [282–290], essentially all approaches depend on the scale of the features, making it difficult to assess and to compare the relevance of different feature subsets: a rescaling of the features into standard SI units, for example, can completely result in a different relevance score of feature (subsets) than using the features as-is without any pre-processing of the data set[8]. Therefore, based on a variant, namely cumulative mutual information [291], a non-parametric information-theoretic feature-selection criterion is developed in the following that is applicable to continuous features and can be applied to materials data (Secs. 3.2.5 and 3.3.3).

3.2.1 Entropy, mutual information, and Kullback-Leibler divergence

Entropy and mutual information are used in communication theory for data compression and compressed sensing [191, 221, 223, 224, 292], but more recently have found applications in materials

[7] Continuous features are real-valued quantities as opposed to discrete features, which are ordinal. Both are numerical features for which a distance metric can be defined. Categorical and nominal features must first be converted to a numerical feature before they can be used in data-science applications.

[8] Consider an equal-width binning of continuous features into discrete features. Depending on the scaling of the feature, it is possible that all values of this feature are assigned to one and the same bin only, thereby resulting in a feature with zero relevance.

science [192, 293, 294]. By maximizing the entropy[9] [295, 296], an optimal representation of the target property can be realized by identifying a set of features, whose mutual dependence is maximal with respect to a property of interest Y (Fig. 3.3).

3.2.2 Kullback-Leibler divergence

The principle of maximum entropy can be turned into a score, where mutual information and related information-theoretic methods such as Kullback-Leibler divergence[10] D_{KL} [85, 297] quantify the relevance of a feature subset in terms of the strength of the mutual dependence in a normalized range[11]. Mathematically, this is done by expressing the features and the property of interest as probability distributions and searching for a set of features, whose (joint) probability distribution makes the fewest assumptions about the shape and distribution of the property of interest (Fig. 3.3). Using Kullback-Leibler divergence, for instance, the relevance for each feature subset $\vec{X}$ can be quantified by measuring the dissimilarity between two distributions, $U(\vec{X},Y)$ and $V(\vec{X},Y)$,

$$D_{\mathrm{KL}}(U(\vec{X},Y)||V(\vec{X},Y)) = \sum_{y\in Y}\sum_{\vec{x}\in X} U(\vec{x},y)\log\frac{U(\vec{x},y)}{V(\vec{x},y)} \geq 0\ . \tag{3.2}$$

Kullback-Leibler divergence D_{KL} is zero if and only if the two distributions are identical,

$$D_{\mathrm{KL}}(U(\vec{X},Y)||V(\vec{X},Y)) = 0 \quad \Leftrightarrow \quad U(\vec{X},Y) \equiv V(\vec{X},Y)\ , \tag{3.3}$$

and is positive otherwise. Using the joint probability distribution as U and the marginal distribution as V, Kullback-Leibler divergence provides a simple means to compare and rank the relevance (=mutual dependence) of different feature subsets with respect to a property of interest: if Kullback-Leibler divergence is zero ($D_{\mathrm{KL}} = 0$), subsets of features are totally unrelated to the property of interest. If Kullback-Leibler divergence is positive ($D_{\mathrm{KL}} > 0$), its value is increasingly larger the stronger the features are related to the property of interest.

3.2.3 Mutual information

Mutual information is a special case of Kullback-Leibler divergence and relates the joint probability distribution $U(\vec{X},Y) \equiv p(\vec{x},y)$ of discrete variables (=features and target) to the product of their marginal distributions $V(\vec{X},Y) \equiv p(\vec{x})p(y)$ [83, 84],

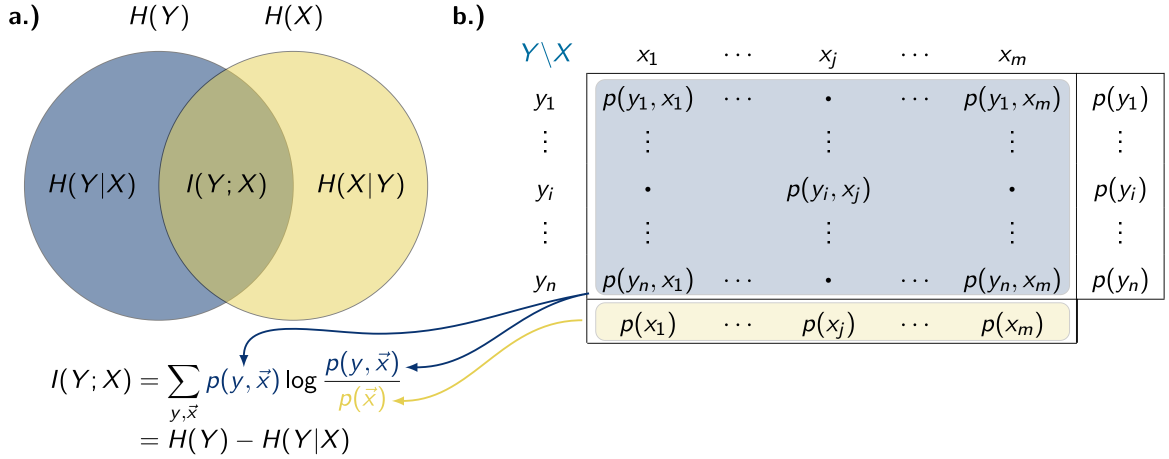
$$I(Y;\vec{X}) = \sum_{y\in Y}\sum_{\vec{x}\in X} p(y,\vec{x})\log\frac{p(y,\vec{x})}{p(y)p(\vec{x})} \equiv D_{\mathrm{KL}}(p(y,\vec{x})||p(y)p(\vec{x}))\ . \tag{3.4}$$

The joint probability distribution is the probability of the co-occurrence of the feature values (y, $\vec{x}$) $\in$ $(Y,\vec{X})$ (Fig. 3.4). The marginal distribution is the probability distribution of a single variable

[9] Jaynes [295] showed that the concept of entropy in statistical mechanics is equivalent to the concept of entropy in information theory.

[10] In information theory, Kullback-Leibler divergence is often referred to as relative entropy [84].

[11] The Kullback-Leibler divergence is in the range $[0, 1]$, where "zero" indicates statistical independence, "one" functional dependence, and anything in between the relative strength of dependence between the features and the property of interest [235].



$Y\backslash X$	x_1	$\cdots$	x_j	$\cdots$	x_m	
y_1	$p(y_1, x_1)$	$\cdots$	•	$\cdots$	$p(y_1, x_m)$	$p(y_1)$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_i	•		$p(y_i, x_j)$		•	$p(y_i)$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_n	$p(y_n, x_1)$	$\cdots$	•	$\cdots$	$p(y_n, x_m)$	$p(y_n)$
	$p(x_1)$	$\cdots$	$p(x_j)$	$\cdots$	$p(x_m)$	

$$
\begin{aligned}
I(Y;X) &= \sum_{y,\vec{x}} p(y,\vec{x}) \log \frac{p(y,\vec{x})}{p(\vec{x})} \\
&= H(Y) - H(Y|X)
\end{aligned}
$$

Fig. 3.4. Schematic representation of all information-theoretical quantities required for the calculation of mutual information of two variables $\{Y, X\}$. Panel a.) depicts the relationship between Shannon entropies, $H(X)$ and $H(Y)$, conditional Shannon entropies, $H(Y|\vec{X})$ and $H(\vec{X}|Y)$, and mutual information $I(Y;\vec{X})$ in a Venn-like diagram [300, 301], whereas b.) displays X in Y as a contingency table to directly compute the marginal probability distributions, $p(y)$ and $p(x)$, and joint probability distribution $p(y,x)$, from X and Y.

in the subset $\{\vec{X}, Y\}$, integrating out all other variables from the joint probability distribution. Under the assumption of independence, i.e., when the variables $\vec{X}$ and Y are statistically uncorrelated, the product of the marginal and joint probability distributions are identical and mutual information is zero, $I(Y;X) = 0$. The greater the dependence between $\vec{X}$ and Y, the more different are the joint and marginal probability distributions and the higher is mutual information, $I(Y;X) \geq 0$. Mutual information increases monotonically with the cardinality of the feature's subset $\vec{X}' \subseteq \vec{X}$,

$$
I(Y;\vec{X}' \setminus X) \leq I(Y;\vec{X}') \quad \forall X \in \vec{X} \ . \tag{3.5}
$$

One immediate consequence of Eq. 3.4 is the invariance of mutual information under coordinate transformations ($\vec{X} \longrightarrow \vec{X}'$ and $Y \longrightarrow Y'$) such as translations and re-parameterization that preserve the order of $\vec{X}$ and Y [235, 297, 298]. Mutual information is therefore ideally suited if features are dimensional quantities and whose rescaling should not affect the mutual dependence to the property of interest. Mutual information is also invariant under the addition of variables $\vec{Z}$ that are unrelated with $\vec{X}$ and Y, $p(\vec{X}, \vec{Z}, Y) = p(\vec{X}, Y)p(\vec{Z})$ [299]. As such, it provides a reliable mutual-dependence estimate, even when only a small number of features are actually mutually related to the property of interest in the data set.

Mutual information, $I(Y;\vec{X}) = H(Y) - H(Y|\vec{X})$, is linearly related to the Shannon entropy $H(Y)$ and conditional Shannon entropy $H(Y|\vec{X})$ [302] (Fig. 3.4). The Shannon entropy $H(Y)$ is a measure of the uncertainty on the occurrence of a feature value y whose probability $p(y)$ is described by $y \in Y$,

$$
H(Y) = -\sum_{y \in Y} p(y) \log p(y) \ . \tag{3.6}
$$

Conditional Shannon entropy quantifies the amount of uncertainty about the value of $y \in Y$ in the presence of the distribution $\vec{X}$,

$$H(Y|\vec{X}) = -\sum_{y \in Y} \sum_{\vec{x} \in X} p(y, \vec{x}) \log p(y|\vec{x}) \ , \tag{3.7}$$

with $p(y|\vec{x}) = p(y, \vec{x})/p(\vec{x})$ as the conditional probability of y given $\vec{x} \in \vec{X}$. Whereas Shannon entropy is bounded from below ($H(Y) \geq 0$), conditional Shannon entropy has values in a range bounded by $H(Y)$, $0 \leq H(Y|\vec{X}) \leq H(Y)$. Conditional Shannon entropy is zero, $H(Y|\vec{X}) = 0$, if variables $\vec{X}$ and Y are completely dependent and is maximal, $H(Y|\vec{X}) = H(Y)$, if variables are statistically independent (=totally unrelated) of each other. Thus, mutual information is bounded from below by $I(Y; Y) = 0$ and restricted to the interval $0 \leq I(Y; \vec{X}) \leq H(Y)$. In order to use mutual information for comparing the feature relevance of a feature set $\vec{X}$ to describe the property of interest Y, mutual information is normalized[12] [298],

$$D(Y; \vec{X}) = \frac{I(Y; \vec{X})}{H(Y)} = \frac{H(Y) - H(Y|\vec{X})}{H(Y)} \ . \tag{3.8}$$

The normalized mutual information (Eq. 3.8), hereafter referred to as fraction of information[13] can be used to evaluate the relevance of different feature subsets with respect to the same property of interest. This dependence score is in the range $[0, 1]$, where "zero" and "one" represent statistical independence and functional dependence, respectively (i.e., the larger the score, the stronger the mutual dependence and the higher the relevance of a subset).

Unfortunately, mutual information and fraction of information are only defined for discrete features[14]. Although mutual information can be generalized to continuous features by using differential entropy [84],

$$I(Y; \vec{X}) = \int_{y \in Y} \int_{x \in \vec{X}} p(y, \vec{x}) \log \frac{p(y, \vec{x})}{p(y)p(\vec{x})} \, d\vec{x} \, dy \ , \tag{3.9}$$

the existence of the probability distributions for $\vec{X}$ and Y are still needed to determine the mutual dependence between a set of features and the property of interest given $I(Y; \vec{X})$.

3.2.4 From probability to cumulative distributions

Probability distributions for continuous features are generally difficult to compute and therefore require approximations. A plethora of approaches have been presented in the literature to estimate continuous features by generalizing mutual information: multivariate maximal correlation analysis [282], maximal information coefficient [283], universal dependency analysis [284, 285], and mutual-information-based feature-selection algorithms originally developed for discrete data [286–290]. However, all approaches quantize continuous features based on clustering [275, 291, 303, 308],

[12] There are multiple possibilities to normalize mutual information [303, 304].

[13] Depending on the research field, fraction of information is also known as coefficient of constraint [305], uncertainty coefficient [306], or proficiency [307].

[14] Discrete features are ordinal quantities as opposed to continuous features, which are real-valued. Both are numerical features for which a distance metric can be defined. Categorical and nominal features must first be converted to a numerical feature before they can be used in data-science applications.

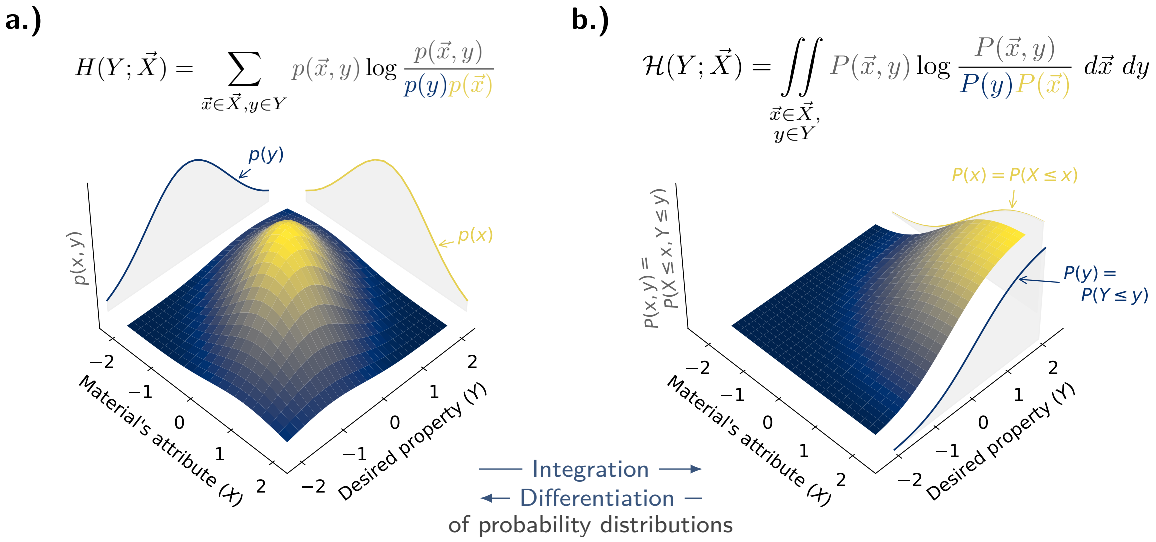


Fig. 3.5. A schematic diagram of the isomorphism between probability (a.) and cumulative distributions (b.) and their respective information-theoretical measures of dependence. Left: mutual information (Section 3.2.3). Right: cumulative mutual information (Section 3.2.5).

discretization [286, 309–312], or density estimation [313–317] and implicitly introduce additional parameters to transform continuous into discrete features. However, all of these parameters need to be redetermined depending on the number of features and the subset of features. In practice, such approaches are therefore extremely dependent on the applied parameter set and are sensitive to the scale of the features[15]. As a result, approaches generalizing mutual information based on quantifying continuous features are unstable in the identification of mutual dependent features and hence are difficult to apply to materials-science data.

Because probability distributions from continuous features cannot be easily obtained, the basic idea of cumulative mutual information [291] and the dependence measure developed in this thesis (total cumulative mutual information, TCMI) is to express the features of the data set as cumulative and residual cumulative distributions (Fig. 3.5). The cumulative distribution P (and residual cumulative distribution $P' \approx 1 - P$) of a variable X evaluated at x describe the probability that X takes on a value less than or equal to x (or a value greater than or equal to x, respectively),

$$P(x) := P(X \le x)\,, \qquad\qquad P'(x) := P(X \ge x) = 1 - P(X < x)\,. \tag{3.10}$$

Cumulative and residual cumulative distributions are defined for both continuous and discrete variables. They are monotonically increasing and decreasing, respectively, i.e., $P(x_1) \le P(x_2)$ or $P'(x_1) \ge P'(x_2)$, $\forall x_1 \le x_2$, with limits

$$\begin{matrix} \lim\limits_{x\to-\infty} P(x) = 0 \\ \lim\limits_{x\to\infty} P(x) = 1 \end{matrix}\,, \qquad\qquad \begin{matrix} \lim\limits_{x\to-\infty} P'(x) = 1 \\ \lim\limits_{x\to\infty} P'(x) = 0 \end{matrix}\,. \tag{3.11}$$

[15] Sensitivity in the feature range implies that a rescaling of a feature, e.g., changing the unit, leads to a completely different solution of the algorithm.

Cumulative distributions are more regular and less sensitive to statistical noise than probability distributions [318, 319]. If the derivatives exist, the probability distribution of continuous and discrete variables can be determined from the cumulative distribution by differentiation,

$$p(x) = \frac{dP(x)}{dx} \quad \Leftrightarrow \quad P(x) = \int_{-\infty}^{x} p(x')\, dx' \,, \tag{3.12a}$$

$$p(x) = 1 - \frac{dP'(x)}{dx} \quad \Leftrightarrow \quad P'(x) = \int_{x}^{\infty} p(x')\, dx' \,. \tag{3.12b}$$

Similar to probability distributions, cumulative and residual cumulative distributions are invariant under a change of variables and re-parameterizations, but only under those that preserve the order of the feature values in $\vec{X}$ and Y. For instance, cumulative distribution are invariant under positive monotonic transformations $T : \mathbb{R} \longrightarrow \mathbb{R}$,

$$\begin{gathered} P(x) = P(T(x)) \;\vee\; \bar{P}(x) = \bar{P}(T(x)) \\ \forall x \in X : x \mapsto T(x) \quad \text{such that} \quad x_1 < x_2 \Rightarrow T(x_1) < T(x_2) \,. \end{gathered} \tag{3.13}$$

such as translations or scalings of the features, while in all other cases of invertible or non-invertible maps [320] the order of the original elements of the variables is changed and with it the cumulative distribution.

3.2.5 Cumulative mutual information

Based on cumulative distributions, cumulative mutual information (Fig. 3.5b.) quantifies the inherent mutual dependence expressed in the joint cumulative distribution $P(\vec{x}, y) := P(\vec{X} \leq \vec{x}, Y \leq y)$ of variables $\vec{X}$ and Y relative to the product of their marginal cumulative distributions $P(\vec{x})$ and $P(y)$,

$$\begin{aligned} \mathcal{J}(Y;\vec{X}) &= \int_{y\in Y} \int_{\vec{x}\in\vec{X}} P(y,\vec{x}) \log \frac{P(y,\vec{x})}{P(y)P(\vec{x})}\, d\vec{x}\, dy \\ &= D_{\mathrm{KL}}(P(y,\vec{x})||P(y)P(\vec{x})) \,. \end{aligned} \tag{3.14}$$

Cumulative mutual information, $\mathcal{J}(Y;\vec{X})$, is monotonically increasing with increasing mutual dependence of $\vec{X}$ and Y and is zero, if and only if $\vec{X}$ and Y are statistically uncorrelated (i.e., under the independence assumption of random variables). Similar to mutual information (Eq. 3.4), cumulative mutual information (Eq. 3.14) determines the degree of mutual dependence (=relevance) as the reduction in the uncertainty of Y given $\vec{X}$, i.e., $\mathcal{J}(Y;\vec{X}) = \mathcal{H}(Y) - \mathcal{H}(Y|\vec{X})$. Cumulative mutual information is linearly related to cumulative entropy $\mathcal{H}(Y)$ [318, 319, 321–323] and conditional cumulative entropy $\mathcal{H}(Y|\vec{X})$,

$$\mathcal{H}(Y) = -\int_{y\in Y} \int_{\vec{x}\in\vec{X}} P(y,\vec{x}) \log P(y)\, d\vec{x}\, dy \tag{3.15a}$$

$$\mathcal{H}(Y|\vec{X}) = -\int_{y\in Y} \int_{\vec{x}\in\vec{X}} P(y,\vec{x}) \log P(y|\vec{x})\, d\vec{x}\, dy \,, \tag{3.15b}$$

where $P(y|\vec{x}) = P(y, \vec{x})/P(\vec{x})$ is the conditional cumulative distribution of $y \leq Y$ given $\vec{x} : \forall x_i \leq X_i$, $X_i \in \vec{X}$. The conditional entropy has values in the range of $0 \leq \mathcal{H}(Y|\vec{X}) \leq \mathcal{H}(Y)$, where the minimum value, $\mathcal{H}(Y|\vec{X}) = 0$, is reached only if the variables $\vec{X}$ and Y are completely dependent and $\mathcal{H}(Y|\vec{X}) = \mathcal{H}(Y)$ if the variables $\vec{X}$ and Y are independent.

Bounds restrict cumulative mutual information to a closed interval $0 \leq \mathcal{I}(Y;\vec{X}) \leq \mathcal{H}(Y)$ with an upper bound dependent on Y. Using the normalized variant, hereafter referred to as fraction of cumulative information,

$$\mathcal{D}(Y;\vec{X}) := \frac{\mathcal{I}(Y;\vec{X})}{\mathcal{H}(Y)} = \frac{\mathcal{H}(Y) - \mathcal{H}(Y|\vec{X})}{\mathcal{H}(Y)} , \quad 0 \leq \mathcal{D}(Y;\vec{X}) \leq 1 , \tag{3.16}$$

the relevance of different feature subsets with respect to the same property of interest can be determined without requiring to introduce additional parameters or to discretize continuous features.

3.3 Empirical estimation of cumulative mutual information

Mutual and cumulative mutual information (Eqs. 3.4 and 3.14) quantify the relevance of a subset of features based on the assumption of smooth and differentiable probability or cumulative distributions, respectively. Due to the limited availability of data, however, the exact functional shape of the probability or cumulative distributions is not directly accessible and hence must be inferred from the data. The estimation [324, 325] of these distributions from a limited amount of data, from the so-called empirical estimates $\widehat{\mathscr{E}}$, is therefore a crucial step in quantifying the mutual dependence of a set of features $X \in \vec{X}$ and the property of interest Y.

From a statistical point of view, an empirical estimate $\widehat{\mathscr{E}}$ consists of a set of independent and identically distributed (i.i.d.) drawn samples $\vec{Z} = \{(y_1, x_1), (y_2, x_2), \ldots, (y_N, x_N)\}$, $(y_i, x_i) \in (Y, X)$ from an unknown population of X and Y. The goal is to estimate a probability $\widehat{p}$ or cumulative distribution $\widehat{P}$ that is most likely to have generated the set of available samples $\vec{Z}$. Using the maximum likelihood estimate [326, 327], the probability distribution $\widehat{p}(Z = z)$ or cumulative probability distribution $\widehat{P}(Z \leq z)$ can be obtained by counting the frequency of occurring feature values,

$$\widehat{p}(Z = z) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{z_i = z} \quad \text{or} \quad \widehat{P}(Z \leq z) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{z_i \leq z} , \quad z \in Z, Z \in \{Y, X\} , \tag{3.17}$$

where $\mathbb{1}_A$ denotes the indicator function that is one if A is true, and zero otherwise. The maximum likelihood estimate has a number of attractive limiting properties: for example, it is invariant under transformations and converges pointwise to the actual value of $\mathscr{E}$, $\widehat{\mathscr{E}}(z) \longrightarrow \mathscr{E}(z)$, for every value of $z \in Z$ as $N \longrightarrow \infty$ [328, 329]. As such, it provides an unbiased estimate with respect to the underlying population as the sample size increases to infinity [327].

3.3.1 Empirical estimators

It should be noted that empirical estimators have so far been defined in the literature only for mutual information. Empirical estimators for cumulative (conditional) entropy and mutual information therefore need to be derived in order to use them for TCMI in Section 3.3.3.

Empirical cumulative entropy

The maximum likelihood estimate of the cumulative entropy $\mathcal{H}(Y)$ (Eq. 3.15a) is called the empirical cumulative entropy [291, 318]. It can be obtained by calculating the empirical cumulative distributions $\widehat{P}$ according to Eq. 3.17,

$$\widehat{P}(y_i) = \frac{|\{y \leq y_i \mid y \in Y\}|}{N} \,. \tag{3.18}$$

Empirical cumulative entropy can be calculated from the set of sample data with linear time complexity $\mathcal{O}(N)$,

$$\widehat{\mathcal{H}}(Y) = -\sum_{i=1}^{N-1} \Delta y_i \widehat{P}(y_i) \log \widehat{P}(y_i) = -\sum_{i=1}^{N-1} (y_{i+1} - y_i)\, \widehat{P}(y_i) \log \widehat{P}(y_i) \,, \tag{3.19}$$

where y_i denotes all the values of the property of interest Y occurring in the data set $y_0 < y_1 < \cdots < y_N$ in sorted order of Y[16].

Empirical conditional cumulative entropy

Similar to empirical cumulative entropy (Eq. 3.15b), conditional cumulative entropy can be estimated by

$$\begin{aligned} \widehat{\mathcal{H}}(Y;\vec{X}) &= -\sum_{i=1}^{N-1}\sum_{j=1}^{N-1} \Delta y_i \Delta \vec{x}_j \widehat{P}(y_i,\vec{x}_j) \log \widehat{P}(y_i|\vec{x}_j) \\ &= -\sum_{i=1}^{N-1}\sum_{j1=1}^{N-1} \cdots \sum_{jd=1}^{N-1} (y_{i+1} - y_i)\,(x_{j1+1}^{(1)} - x_{j1}^{(1)}) \ldots (x_{jd+1}^{(d)} - x_{jd}^{(d)}) \widehat{P}(y_i,\vec{x}_j) \log \widehat{P}(y_i|\vec{x}_j) \,, \end{aligned} \tag{3.20}$$

where $\widehat{P}(y_i,\vec{x}_j)$ denotes the joint cumulative distribution of $y_i \in Y$, $\vec{x}_j \in \vec{X}$, $\vec{X} = \{X_1, \ldots X_d\}$, and $x_i^{(k)} \in X_k$ is the i component of the k-th feature of the data set ($k = 1, \ldots d$). In contrast to the empirical cumulative entropy, the empirical conditional cumulative entropy has exponential time complexity $\mathcal{O}(N^d)$ and thus becomes computationally demanding for data sets with a large number of features d and data samples N.

Empirical cumulative mutual information

Based on the empirical and conditional cumulative entropy (Eqs. 3.19 and 3.20), the empirical estimate of cumulative mutual information $\widehat{\mathcal{I}}(Y;\vec{X})$ (cf., Eq. 3.14) used for TCMI can be derived as

$$\widehat{\mathcal{I}}(Y;\vec{X}) = \sum_{i=1}^{n-1} \Delta y_i \widehat{P}(y_i) \log \widehat{P}(y_i) - \frac{1}{n} \sum_{i,j=1}^{n-1} \Delta y_i \widehat{P}(y_i,\vec{x}_j) \log \widehat{P}(y_i|\vec{x}_j) \,, \tag{3.21}$$

[16] The order of samples $y_1 \leq \ldots \leq y_N$, $y \in Y$ is also called the order statistics of Y.

where $\Delta y_i = |y_{i+1} - y_i|$ is defined as the difference between two consecutive values of the property of interest Y in sorted order. Similarly, the empirical estimate for the fraction of cumulative information (cf., Eq. 3.16) is given by

$$\widehat{\mathcal{D}}(Y;\vec{X}) = 1 - \frac{1}{n}\left[\sum_{i,j=1}^{n-1} \Delta y_i \widehat{P}(y_i, \vec{x}_j) \log \widehat{P}(y_i|\vec{x}_j) \Bigg/ \sum_{i=1}^{n-1} \Delta y_i \widehat{P}(y_i) \log \widehat{P}(y_i)\right] . \tag{3.22}$$

3.3.2 Adjustment of empirical estimators for small sample sizes

Empirical estimators for mutual information are known to assign stronger dependences for larger subsets of features than for smaller ones regardless of the underlying relationship [330]. Especially for small data sets with a large number of features, empirical estimators for mutual information never reach their theoretical maximum (functional dependence) or minimum (statistical independence) [304, 331]. This has the consequence of incorrectly identifying features as relevant that are not related to Y. A correction of empirical estimators based on mutual information is usually done by subtracting the actual value of the empirical estimator $\widehat{\mathcal{E}}$ with respect to the average $\widehat{\mathcal{E}}_0$ of the estimator,

$$\widehat{\mathcal{E}}^*(Y;\vec{X}) = \widehat{\mathcal{E}}(Y;\vec{X}) - \widehat{\mathcal{E}}_0(Y;\vec{X}) . \tag{3.23}$$

The average $\widehat{\mathcal{E}}_0$ vanishes for a large number of samples $\widehat{\mathcal{E}}_0(Y;\vec{X}) \longrightarrow 0$ as $n \longrightarrow \infty$ in case there is an exact functional dependence between $\vec{X}$ and Y [332]. However, $\widehat{\mathcal{E}}_0$ is generally greater than zero when the number of data samples is limited and can become as large as the empirical estimator $\widehat{\mathcal{E}}$ when the number of data samples is very small. $\widehat{\mathcal{E}}_0$ can therefore be interpreted as a correction term for comparing the empirical estimates of different feature subsets on a common baseline: In general, if the value of the correction term is large, more data samples are needed to reliably estimate a dependence with respect to a property of interest. If the value of correction term is small, the adjusted empirical estimator either indicates a strong mutual dependence between a set of features $\vec{X}$ and a property of interest Y (high $\widehat{\mathcal{E}}^*(Y;\vec{X})$) or a weak mutual dependence, if the features of the data set are not suited to estimate Y (low $\widehat{\mathcal{E}}^*(Y;\vec{X})$).

A non-zero expected value of an empirical estimator is equivalent to the mean, if the empirical estimator is computed multiple times based on random permutations of all features independently for each data sample,

$$\widehat{\mathcal{E}}_0(Y;\vec{X}) := \frac{1}{|\mathcal{M}|} \sum_{M \in \mathcal{M}} \widehat{\mathcal{E}}(Y_M;\vec{X}_M) , \tag{3.24}$$

where $M \in \mathcal{M}$ is a specific realization of such a permutation. The underlying intuition is that the actual value of the empirical estimator $\widehat{\mathcal{E}}$ may be caused by spurious (random) dependences. Therefore, by considering all random permutations of all features independently for each data sample, the spurious contribution of the empirical estimator can be factored out and an adjusted unbiased empirical estimator obtained. The permutations can be computed by enumeration, which however is impractical. An alternative description is provided by a hypergeometric model of randomness[17] [330, 334]. Such a model describes the permutation of variables as probability distributions, where the average can be calculated separately for each sample in the data set.

[17] Also known as permutation model [333].

$$\begin{array}{c|ccccc|c}
\mathbf{Y\backslash X} & \tilde{X}_1 & \cdots & \tilde{X}_j & \cdots & \tilde{X}_c & \\
\hline
\tilde{Y}_1 & n_{11} & \cdots & \cdot & \cdots & n_{1c} & a_1 \\
\vdots & \vdots & & \vdots & & \vdots & \vdots \\
\tilde{Y}_i & \cdot & & n_{ij} & & \cdot & a_i \\
\vdots & \vdots & & \vdots & & \vdots & \vdots \\
\tilde{Y}_r & n_{r1} & \cdots & \cdot & \cdots & n_{rc} & a_r \\
\hline
 & b_1 & \cdots & b_j & \cdots & b_c &
\end{array}$$

Fig. 3.6. An $r \times c$ cumulative contingency table $\mathscr{M}$ related to a specific realization of the joint cumulative distribution $P(y_i, x_j)$ given row marginal a_i and column marginal b_j of two variables $\tilde{X}$ and $\tilde{Y}$ with row marginals, $a_i = \sum_{j=1}^{c} n_{ij}$, and column marginals, $b_j = \sum_{i=1}^{r} n_{ij}$. The two cumulative marginal sum vectors $a = [a_i]$ and $b = [b_j]$ are constant and satisfy the fixed marginals condition, $\sum_{i=1}^{r} a_i = \sum_{j=1}^{c} b_j = N$, where N is the number of data samples.

A hypergeometric model of randomness has already been discussed in the literature for mutual information [294, 304, 330, 334]. In the following, the hypergeometric model of randomness is extended to cumulative mutual information. Because the derivation is almost verbatim to the derivation of the hypergeometric model for mutual information, detailed proofs will be omitted.

The expected value of cumulative mutual information between all permutations of two variables X and Y with $|Y_i| = a_i$, $i = 1, \ldots, r$ and $|X_j| = b_j$, $j = 1, \ldots, c$ equals

$$\hat{\mathscr{I}}_0(Y;X|M) = \hat{\mathscr{I}}_0(a, b|M = [n_{ij}]_{j=1\cdots c}^{i=1\cdots r}) = -\sum_{i=1}^{r-1}\sum_{j=1}^{c} \Delta y_i(M)\frac{n_{ij}}{n}\log\frac{n_{ij}}{b_j}\ . \tag{3.25}$$

Equation 3.25 can be interpreted as an $r \times c$ cumulative contingency table, $M = [n_{ij}]_{j=1\cdots c}^{i=1\cdots r}$ (Fig. 3.6), with n_{ij} being a specific realization of the joint cumulative distribution $P(y_i, x_j)$ given row marginal a_i and column marginal b_j.

By rearranging the sums in Eq. 3.25 and expressing the sum over the entire permutation of variable values as a sum over all permutations of possible values of n_{ij}, the expected value of cumulative mutual information can be written as

$$\begin{aligned}
\hat{\mathscr{I}}_0(Y;X) &= -\sum_{M\in\mathscr{M}}\sum_{i=1}^{r-1}\sum_{j=1}^{c} \Delta y_i(M)\frac{n_{ij}}{n}\log\frac{n_{ij}}{b_j}\mathscr{P}(Y;X|M) \\
&= -\sum_{i=1}^{r-1}\sum_{j=1}^{c}\sum_{n_{ij}} \Delta y_i(n_{ij}, a_i, b_j|M)\cdot\frac{n_{ij}}{n}\log\frac{n_{ij}}{b_j}\mathscr{P}(n_{ij}, a_i, b_j|M)\ ,
\end{aligned} \tag{3.26}$$

where $\mathscr{P}(n_{ij}, a_i, b_j|M)$ is the probability to encounter an associative cumulative contingency table subject to fixed marginals. This probability, with the cell at the i-th row and j-th column equals to n_{ij}, is identical to the hypergeometric distribution,

$$\mathscr{P}(n_{ij}, a_i, b_j|M) = \mathscr{P}(b_j - n_{ij}, r-1, r-i, b_j-1) = \binom{r-i}{b_j - n_{ij}}\binom{i-1}{n_{ij}-1}\Big/\binom{r-1}{b_j-1}\ . \tag{3.27}$$

The hypergeometric distribution describes the probability of $b_j - n_{ij}$ successes in $b_j - 1$ draws without replacement where the finite population consists of $r-1$ elements, of which $r-i$ are classified as successes and $\mathcal{P}(n_{ij}, a_i, b_j|M)$ is evaluated in the range $\max(0, i+b_j-r) \leq n_{ij} \leq \min(i, b_j)$. Likewise, the difference $\Delta y_i(M)$ between two consecutive values of the property of interest can be described by a binomial distribution,

$$\Delta y_i(n_{ij}, a_i, b_j|M) = \frac{1}{\mathcal{N}} \sum_{k=1}^{k_{\max}} \binom{r-k-1}{b_j-2} \left(y_{(i+k)} - y_{(i)} \right) , \tag{3.28}$$

where the upper limit is given by $k_{\max} = \min(n - b_j + 1, r - i)$ and $\mathcal{N}$ is the normalization constant:

$$\mathcal{N} = \sum_{k=1}^{k_{\max}} \binom{r-k-1}{b_j-2} . \tag{3.29}$$

Putting all these terms together, the expected cumulative mutual information (Eq. 3.25) under the assumption of the hypergeometric model of randomness is

$$\begin{aligned} \hat{\mathcal{I}}_0(Y;X) = -\sum_{i=1}^{r-1} \sum_{j=1}^{c} \sum_{n_{ij}} \Delta y_i(M|n_{ij}, a_i, b_j) \frac{n_{ij}}{n} \log\left(\frac{n_{ij}}{b_j}\right) \\ \cdot \frac{(r-i)!(i-1)!(b_j-1)!(r-b_j)!}{(b_j-n_{ij})!(r-i-b_j+n_{ij})!(n_{ij}-1)!(i-n_{ij})!(r-1)!} . \end{aligned} \tag{3.30}$$

The empirical estimator for cumulative mutual information can then be defined as

$$\widehat{\mathcal{I}}^*(Y;\vec{X}) = \widehat{\mathcal{I}}(Y;\vec{X}) - \widehat{\mathcal{I}}_0(Y;\vec{X}) , \tag{3.31}$$

$$\widehat{\mathcal{D}}^*(Y;\vec{X}) = \frac{\widehat{\mathcal{I}}^*(Y;\vec{X})}{\widehat{\mathcal{H}}(Y)} = \widehat{\mathcal{D}}(Y;\vec{X}) - \widehat{\mathcal{D}}_0(Y;\vec{X}) \tag{3.32}$$

where $\widehat{\mathcal{I}}^*(Y;\vec{X})$ is the adjusted empirical cumulative mutual information, $\widehat{\mathcal{D}}^*(Y;\vec{X})$ is the adjusted fraction of empirical cumulative information, and $\widehat{\mathcal{I}}_0(Y;\vec{X})$ is the expected cumulative mutual information under the independence assumption of random variables.

3.3.3 Total cumulative mutual information

The empirical estimation of (cumulative) mutual information (Eq. 3.22), and especially its adjusted variant (Eq. 3.23), provides a non-parametric deterministic measure to estimate the relevance of a set of features that is statistically related to the property of interest.

The information-theoretic dependence measure developed here is based on the empirical estimation of cumulative mutual information. Referred to as total cumulative mutual information (TCMI), it computes the empirical estimate of cumulative mutual information using the cumulative ($\widehat{P}(Z \leq z)$) and residual cumulative distributions ($\widehat{P}'(Z \geq z) = 1 - \widehat{P}(Z < z)$, Eq. 3.10) to quantify the strength of dependence, i.e., relevance of a feature set, in terms of two empirical dependence measures (Eq. 3.22):

$\widehat{\mathscr{D}}^*(Y;X)$ defined on the empirical cumulative distribution $\widehat{P}$ and $\widehat{\mathscr{D}}'^*(Y;X)$ defined on its residual $\widehat{P}'$ (Eq. 3.10),

$$\widehat{\mathscr{D}}^*(Y;\vec{X}) = \widehat{\mathscr{D}}(Y;\vec{X}) - \widehat{\mathscr{D}}_0(Y;\vec{X}) \,, \qquad \widehat{\mathscr{D}}(Y;\vec{X}) = \left[\widehat{\mathscr{H}}(Y) - \widehat{\mathscr{H}}(Y|\vec{X})\right] / \widehat{\mathscr{H}}(Y) \tag{3.33a}$$

$$\widehat{\mathscr{D}}'^*(Y;\vec{X}) = \widehat{\mathscr{D}}'(Y;\vec{X}) - \widehat{\mathscr{D}}_0'(Y;\vec{X}) \,, \qquad \widehat{\mathscr{D}}'(Y;\vec{X}) = \left[\widehat{\mathscr{H}}'(Y) - \widehat{\mathscr{H}}'(Y|\vec{X})\right] / \widehat{\mathscr{H}}'(Y) \,. \tag{3.33b}$$

Both measures, $\hat{\mathscr{D}}^*(Y;X)$ and $\hat{\mathscr{D}}'^*(Y;X)$, estimate the dependence and provide lower and upper bounds on the mutual dependence of a feature subset (cf., Eq. 3.10). As the sample size increases to infinity, both measures converge to the same value. However, due to the limited number of data samples and the necessary correction of the dependence measures (Section 3.3), $\hat{\mathscr{D}}^*(Y;X)$ and $\hat{\mathscr{D}}'^*(Y;X)$ are different in practice. TCMI is therefore defined as the average mutual dependence of a feature subset with respect to the property of interest as given by

$$\langle \widehat{\mathscr{D}}^*_{\mathrm{TCMI}}(Y;\vec{X}) \rangle := \frac{\hat{\mathscr{D}}^*(Y;\vec{X}) + \hat{\mathscr{D}}'^*(Y;\vec{X})}{2} \,. \tag{3.34}$$

TCMI shares the same properties as cumulative mutual information (cf., Sec. 3.2.5): it is a deterministic, non-parametric measure which is invariant under positive monotonic transformations (Eq. 3.13) and increases monotonically with the cardinality of the feature's subset (cf., Eq. 3.5). It can therefore be applied to continuous features without having to introduce additional parameters, that would otherwise affect the dependence estimates when features are rescaled. As TCMI is further corrected using a baseline (Sec. 3.3.2), the features related to the property of interest can be determined independently of the sample size, the actual functional form of the feature-property relationship, or the number of features in a data set. The exponential computational complexity in the calculation of the dependence score (Sec. 3.3) may limit the current implementation of TCMI to data sets with very few samples or features. TCMI's applicability to data sets with larger numbers of features will be examined in the following chapters. The focus of Section 3.5 provides a first comparison between TCMI and other feature-identification methods on simple data sets. In Chapter 5, TCMI is finally investigated for its application to materials-science problems.

3.4 The branch-and-bound algorithm

The identification of the strongest mutual dependence (i.e., highest relevance) of feature subsets with respect to a property of interest, requires to estimate the mutual dependence of all feature-subset combinations of a data set by a search strategy that is efficient in practice. Optimal feature-selection search strategies (Eq. 3.1) are combinatorial optimization problems for constructing efficient representations independent of a machine-learning model. In particular, they enable parallel experimentations with existing feature-identification methods to identify the optimal feature set $\vec{X}^* = \{X_1^*, \ldots, X_r^*\} \subseteq \vec{X}$,

$$\vec{X}^* = \operatorname*{arg\,max}_{\vec{X}' \in \vec{X}} \mathbb{Q}(Y;\vec{X}') \,, \qquad \text{(Eq. 3.1 of Section 3.1.2)}$$

where $\vec{X} = \{X_1, \ldots, X_d\}$ is the set of initial features and $\mathbb{Q} : \vec{X} \longrightarrow Y$ is the feature-selection criterion that establishes a mapping between a subset of features $\vec{X}' \subseteq \vec{X}$ and the property of interest Y.

The most widely used algorithm for addressing combinatorial optimization problems is branch and bound. Branch and bound [91, 92] implicitly enumerates the space of all possible feature subsets (the branching aspect), thereby saving time in discarding subsets whose feature-selection criterion cannot be improved [94–96] (the bounding aspect). The branch-and-bound algorithm requires a feature selection criterion $\mathbb{Q}$ that is independent of the ordering of the feature subset and increases monotonically as the number of features in a feature subset $\vec{X}' \rightarrow \vec{X}$ gradually increases [91, 92],

$$\mathbb{Q}(Y;\vec{X}') \leq \mathbb{Q}(Y;\vec{X}) \ , \qquad \text{if } \vec{X}' \subseteq \vec{X} \ . \tag{3.35}$$

As an explicit enumeration is usually not possible due to the exponentially increasing number of potential feature subsets for data sets with many features, an essential part in the branch-and-bound algorithm is the use of bounds to search parts of the feature space only implicitly. For example, if feature subsets $\vec{X}_1, \vec{X}_2, \cdots, \vec{X}_k$ are obtained by sequentially adding features from the set of features $\vec{X}$ one by one,

$$\vec{X}_1 \subseteq \vec{X}_2 \subseteq \cdots \subseteq \vec{X}_k \ , \qquad \vec{X}_k \subseteq \vec{X} \ , \tag{3.36}$$

the monotonically increasing feature-selection criterion $\mathbb{Q}$ and monotonically decreasing bounding criterion $\overline{\mathbb{Q}}$ [94, 237] ensure that

$$\overline{\mathbb{Q}}(\vec{X}_1) \geq \overline{\mathbb{Q}}(\vec{X}_2) \geq \cdots \geq \overline{\mathbb{Q}}(\vec{X}_k) \quad \text{as compared to} \quad \mathbb{Q}(\vec{X}_1) \leq \mathbb{Q}(\vec{X}_2) \leq \cdots \leq \mathbb{Q}(\vec{X}_k) \ . \tag{3.37}$$

The combinatorial ($\mathcal{NP}$)-hard optimization problem [256, 257] can then be turned into a $\mathcal{P}$-hard (convex) optimization problem, if the bounding criterion is the maximum value of a feature-selection criterion found so far in the search, i.e., $\overline{\mathbb{Q}}(Y;\vec{X}^*) \equiv \max_{\vec{X}^* \subseteq \vec{X}} \mathbb{Q}(Y;\vec{X}^*)$, such that

$$\overline{\mathbb{Q}}(Y;\vec{X}') < \overline{\mathbb{Q}}(Y;\vec{X}^*) \quad \Rightarrow \quad \overline{\mathbb{Q}}(Y;\vec{X}') < \mathbb{Q}(Y;\vec{X}^*) \ , \quad \forall |\vec{X}'| \geq |\vec{X}^*| \ . \tag{3.38}$$

Equation 3.38 implies that all feature subsets with more features than $\vec{X}'$ have a smaller value in the feature-selection criterion $\mathbb{Q}$ than $\vec{X}^*$,

$$\mathbb{Q}(Y;\vec{X}' \cup \vec{X}'') \leq \mathbb{Q}(Y;\vec{X}^*) \qquad \forall \vec{X}'', \vec{X}' \subseteq \vec{X} \ , \tag{3.39}$$

As a result of the suboptimality test (Eqs. 3.37 and 3.39), all feature subsets with a smaller value in the bounding criterion than the current best value of the feature-selection criterion can be discarded in the search [91–93, 95]. In this way, only promising candidate feature subsets are evaluated for optimality, thereby saving time in discarding subsets whose feature-selection criterion cannot be improved and are guaranteed to be sub-optimal[18] [92, 237].

To solve the combinatorial optimization problem, branch-and-bound iteratively builds a search tree T of feature subsets $\vec{X}' \subseteq \vec{X}$ with increasing subset cardinality [94, 96] (Fig. 3.7 and Alg. 3.1). Initially the tree contains only the empty subset, $\vec{S}_0 = \emptyset$. At each iteration (steps 1 to 7 in the example of Fig. 3.7), unexplored feature-subset combinations are generated by augmenting the subset $\vec{S}_k \subseteq \vec{X}$ with one feature $X \in \vec{X}$ at a time, e.g., $\vec{S}_1 = \vec{S}_0 \oplus X$, and then are added to T (branching step).

[18] Other feature-selection techniques based on artificial intelligence guarantee the optimum solution only in case of an exhaustive search [237].

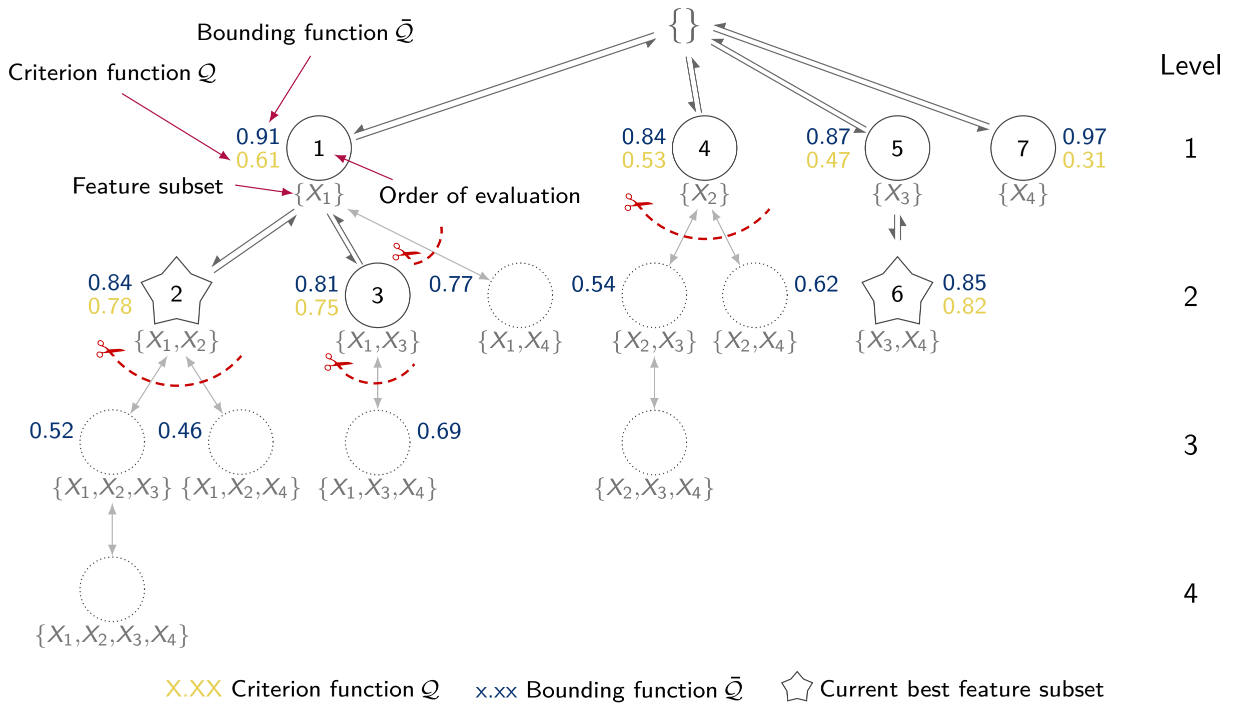


Fig. 3.7. Example of a depth-first tree search strategy of the branch-and-bound algorithm [91, 92, 94–96] to find the optimal subset of features. Shown is the tree traversal going from top to down and left to right (step 1 to 7), the feature-selection bounding criterion $\bar{\mathbb{Q}}$ (top right or left of the circles), the feature-selection criterion $\mathbb{Q}$ (bottom right or left of the circles), and feature subsets (labels at the bottom of the circles). The current-best feature subset in the iterative search is represented by a star. The sequence of exploration of feature subsets is displayed inside the circles. Anytime the bounding criterion in some internal nodes is less than the current-best feature-selection criterion, i.e., $\forall \vec{X}_k, \bar{\mathbb{Q}}(Y;\vec{X}_k) \leq \mathbb{Q}(Y;\vec{X}^*)$, sub-trees can be pruned (indicated by the scissors symbol) and computations be skipped (depicted as dotted circles of feature subsets).

The expansion and combination of the feature subsets in the search space can be represented as paths (connection of lines) and nodes (circles) of a search tree (Fig. 3.7). While traversing the tree from the root (the empty subset) down to terminal nodes (the feature subsets) from left to right, the feature-selection criterion $\mathbb{Q}$, the bounding criterion $\overline{\mathbb{Q}}$, and the currently best subset $\vec{X}^* := \arg\max_k \mathbb{Q}(Y;\vec{X}_k)$ are calculated for each of the feature subsets $\vec{X}_k$. Anytime the feature-selection criterion of a subset is found to be larger than the currently best subset $\vec{X}^*$, the bounding criterion is updated and the current values of $\mathbb{Q}(Y;\vec{X}_k)$, $\overline{\mathbb{Q}}(Y;\vec{X}_k)$, and $\vec{X}^* := \vec{X}_k$ are stored (selection step). Whenever the bounding criterion is found to be less than the current-best feature-selection criterion, i.e., $\forall \vec{X}_k, \overline{\mathbb{Q}}(Y;\vec{X}_k) \leq \mathbb{Q}(Y;\vec{X}^*)$, sub-trees can be pruned and computations can be skipped (bounding step) [94, 96]. In step 1 of Figure 3.7, for example, the branch-and-bound algorithm searches for the best pair of features that have a higher feature-selection criterion $\mathbb{Q}$ than $\vec{S}_1$. The feature subset $\vec{S}_2 = \{X_1, X_2\}$ has the highest feature-selection criterion of all pairs of features. Its value $\mathbb{Q}_2$ and bounding function $\bar{\mathbb{Q}}_2$ are therefore used to prune further feature subsets. Because all subsets with three features have lower bounding criteria than the current best feature-selection criterion $\mathbb{Q}_2$ (feature subset $\vec{S}_2$), these feature subsets can be ignored by the branch-and-bound algorithm. At this step, the

```
Data: Features $\vec{X}$, target $Y$
Result: Optimal features $\vec{X}^* \subseteq \vec{X}$
function branch_and_bound($Y$, $\vec{X}$):
    $\vec{S}_0 = \varnothing$;
    subsets = $\{\vec{S}_0\}$;
    optimal = $\vec{S}_0$;
    bound = $\mathbb{Q}(Y; \text{optimal})$;
    while subsets do
        $\vec{S}_{k-1}$ = subsets.$pop(0)$ ;
        for $X_i \in \vec{X} \setminus \vec{S}_{k-1}$ do
            $\vec{S}_k = \vec{S}_{k-1} \oplus X_i$;
            Compute $\mathbb{Q}(Y; \vec{S}_k)$ and $\overline{\mathbb{Q}}(Y; \vec{S}_k)$;
            // Check suboptimality condition (I) (Eqs. 3.37 and 3.39)
            if $\overline{\mathbb{Q}}(Y; \vec{S}_k)$ < bound then
                continue;
            // Check suboptimality condition (II)
            if $\mathbb{Q}(Y; \vec{S}_k) < \overline{\mathbb{Q}}(Y; \vec{S}_{k-1})$ then
                subsets = subsets $\cup \{\vec{S}_k\}$;
                // Update optimal feature subset and bound
                if $\mathbb{Q}(Y; \vec{S}_k) > \mathbb{Q}(Y; \text{optimal})$ then
                    optimal = $\vec{S}_k$;
                    bound = $\mathbb{Q}(Y; \vec{S}_k)$;
    return optimal;
```

Algorithm 3.1: A pseudo-code listing of the branch-and-bound algorithm [94, 96].

feature subset with all four features does not need to be evaluated as its bounding criterion is the lowest of all feature subsets.

Then, the algorithm returns to the previous feature subset and evaluates the next unexplored feature-subset combination for expansion. The same applies if all combinations have been evaluated for a given feature subset. In step 3 (Fig. 3.7), for example, the algorithm backtracks and evaluates $\mathbb{Q}$ and $\bar{\mathbb{Q}}$ for the feature subset $\vec{S}_3 = \{X_1, X_3\}$. Again, the feature subset $\{X_1, X_3, X_4\}$ has a lower $\bar{\mathbb{Q}}$ than $\mathbb{Q}^* \equiv \mathbb{Q}_2$ and thus can be ignored. The algorithm backtracks and evaluates $\bar{\mathbb{Q}}$ of the remaining feature subset $\{X_1, X_4\}$. Because the bounding criterion is lower than $\mathbb{Q}^*$, it cannot be an optimal solution of the feature-subset search, and is henceforth pruned. In step 4, the algorithm backtracks, evaluates $\mathbb{Q}$ and $\bar{\mathbb{Q}}$ for the feature subset $\vec{S}_4 = \{X_2\}$ and skips further calculations as all n-tuples containing the feature X_2 have a lower $\bar{\mathbb{Q}}$ than $\mathbb{Q}^*$. The algorithm backtracks again and computes all n-tuples of feature subsets containing the feature X_3 (step 5). In step 6, the branch-and-bound algorithm finds a subset ($S_6 = \{X_3, X_6\}$) with a higher $\bar{\mathbb{Q}}$ and $\mathbb{Q}$, that now becomes the optimal feature subset $X^* \equiv S_6$

of the search, while the feature-selection $\mathbb{Q}^* \equiv \mathbb{Q}_6$ and bounding criterion $\bar{\mathbb{Q}}^* \equiv \bar{\mathbb{Q}}_6$ are updated and used for subsequent evaluations. Finally, $\bar{\mathbb{Q}}$ and $\mathbb{Q}$ are computed for the remaining feature subset $S_7 = \{X_4\}$ and tested for optimality.

Once the entire tree has been examined (step 7, Fig. 3.7), the search terminates and the best feature subset as well as a ranking of feature subsets in descending order of the feature-selection criterion value is returned. Since feature subsets are discarded only if Equation 3.39 is fulfilled, the feature subset with the highest feature selection-criterion that best solves the combinatorial optimization problem (Eq. 3.1) is returned (called the optimal feature subset, here: the feature subset S_6), while all other feature subsets with the highest value of the feature-selection criterion within each path (i.e., the terminal nodes) are referred to as sub-optimal feature subsets (in the example these are the feature subsets S_2, S_3, S_4, and S_7).

The computational complexity of the branch-and-bound algorithm is largely determined by two factors: the branching factor b and the depth D of the tree [96]. The branching factor is the maximum number of generated feature-subset combinations at each level l of the tree and can be estimated by the central binomial coefficient $b \leq \max_{l=1,\dots,D} \binom{d}{l} \approx \binom{d}{d/2}$, if $\vec{X}$ has d features. The depth D of the tree is given by the largest cardinality of the feature subset, represented as the longest path in the tree T from the root to a terminal node. Thus, any branch-and-bound algorithm has worst-case $\mathcal{O}(Mb^d)$ computational time complexity, where M is the time needed to evaluate the feature-selection criterion for a feature-subset combination in the tree. However, due to the suboptimality condition (Eq. 3.39) the algorithm is extremely efficient in practice [92].

In summary, the search strategy, the branching, and the pruning rules are crucial for the efficiency of the branch-and-bound algorithm [94, 96]. For example, a common branching strategy keeps track of explored feature subsets, so that feature subsets are only evaluated once and more promising feature subsets are evaluated first. The pruning rules ensure that regions of the search space are discarded, whenever the feature-selection criterion cannot be improved (Eq. 3.39), while the search strategy determines the order in which feature subsets are explored. For example, feature subsets with the highest feature-selection criterion can be evaluated first (best-first search strategy) or only the n-most promising feature subsets (beam search), which, however, leads to a sub-optimal solution of feature subsets [237]. Another possibility is to evaluate all feature subsets at the same level of the tree (breath-first search strategy), i.e., same feature-subset cardinality, or evaluate a feature subset until all combinations are exhausted and the next feature subset is explored (depth-first search strategy). All strategies differ in terms of their time complexity and memory requirements [226] and, with the exception of the beam search, lead to the same optimal feature subset.

3.5 Comparison of feature-selection methods

The success of a feature-selection workflow in materials science hinges on the reliable identification and characterization of relationships between a set of features and a property of interest to build predictive machine-learning models for targeted materials-science applications (Section 2.6). Because typical data sets in computational materials science can have hundreds to thousands of features, the following feature-selection workflow is designed to approach the task as a search problem (Eq. 3.1) and

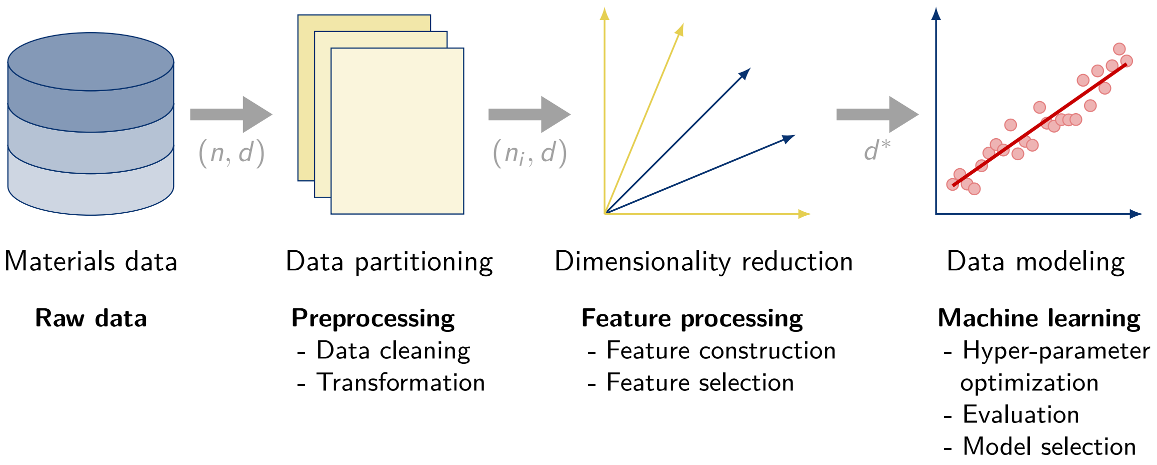


Fig. 3.8. Idealized stages of a feature-selection workflow for materials science. First, materials data of n samples and d features $\vec{X} = \{X_1, \dots, X_d\}$ are preprocessed and partitioned into smaller sets of $n_i \leq n$ samples. Second, dimensionality reduction is performed on the partitioned data sets and the most stable and relevant features $\vec{X}^* \subseteq \vec{X}$ of size $d^* = |\vec{X}^*|$ are identified. In the last step, a machine-learning model is created from the identified features $\vec{X}^*$ and applied to predict the materials-science property of interest.

to construct a materials representation with reduced computational complexity for predictive modeling. This workflow provides a ranked list of feature subsets from a systematic search (Section 3.1.2) and evaluates the relevance of these subsets (Section 3.1.2) based on a specified feature-selection criterion [71, 236].

3.5.1 The workflow

A general feature-selection workflow for materials science comprises the identification of relevant features, a stability analysis of the identified features, and the final creation of a predictive machine-learning model (Fig. 3.8). The goal of such a feature-selection workflow is to identify a set of features related to the property of interest Y. It involves capturing the underlying statistical relationships in the data and to create statistical models whose prediction performance is similar to or even better than building a model using all the initial features of a data set. Currently, most of the works [55, 161–168, 171, 172, 177, 179, 335] either construct and select features heuristically by expertise, intuition, or trial-and-error without actually systematically analyzing their relevance. Consequently, a pipeline of statistical tools is proposed to evaluate the relevance of feature subsets, not limited to materials-science data sets.

As it is computationally prohibitive for non-deterministic polynomial ($\mathcal{NP}$)-hard materials-science problems to search the whole space of possible feature subsets, one usually has to settle for approximations or search heuristics, which are generally not valid for every data set [336]. Therefore, it is suggested to apply branch-and-bound [91–96] that combines both optimality and speed [92, 239] and has proven to be efficient in the discovery of non-linear functional dependences [97, 294]. Branch-and-bound guarantees to find the optimal subset of features from a feature-selection criterion without evaluating all possible subsets (Section 3.1.2). It requires a monotonically increasing feature-selection criterion [92] such as total cumulative mutual information (TCMI, Section 3.3.3).

In the following, TCMI with branch-and-bound is compared with common feature-selection techniques. Three examples are discussed where there is sufficient knowledge to empirically relate the features of the data set to the targeted materials property (concrete data set) or the relevant features are directly accessible (bivariate normal distribution, Friedman regression data set). The comparison and benchmarking include filter methods such as Pearson's R [101] and Spearman's rank ρ correlation [269], wrapper methods such as recursive feature elimination [263] and embedded methods using machine-learning algorithms such as random forests [266], gradient boosting [249–252, 267, 268], and the sure-independence screening and sparsifying operator [197].

3.5.2 Feature-selection methods

Pearson's and Spearman's coefficient of determination Pearson's R and Spearman's rank ρ correlation quantify the strength of linear and monotonic correlations between two variables (e.g., between two features or one feature and the property of interest). As it can be assumed that relationships between materials properties are rarely linear or bivariate, the Pearson's $R^2 : (Y, X) \to [0, 1]$ and Spearman's $\rho^2 : (Y, X) \longrightarrow [0, 1]$ coefficient of determination may not be able to identify all relevant features of a data set. Besides, to have a feature-selection criterion, one has also to define a threshold for pruning and decide which feature to prune[19]. Both of these are arbitrary. Pearson's R and Spearman's rank ρ correlation are therefore only used to analyze the data set for non-linear pairwise monotonically related variables.

Total cumulative mutual information (TCMI) TCMI is an information-theoretic dependence measure that satisfies the monotonicity condition of the branch-and-bound algorithm (Eq. 3.35). Its score can therefore be used with branch-and-bound to efficiently identify the strongest mutual-dependent feature subsets to the property of interest. Because TCMI relates the strength of a dependence to the dependence of the same set of features under the independence assumption of random variables (Eq. 3.32), its expected value can be utilized to define a bounding criterion by estimating the adjusted variants of empirical (cumulative) mutual information by an upper bound, i.e., $\widehat{\mathcal{I}}^*(Y;\vec{X}) \le \overline{\mathcal{I}}^*(Y;\vec{X})$ and $\widehat{\mathcal{D}}^*(Y;\vec{X}) \le \overline{\mathcal{D}}^*(Y;\vec{X})$, where

$$\overline{\mathcal{I}}^*(Y;\vec{X}) = 1 - \widehat{\mathcal{I}}_0(Y;\vec{X}) \,, \tag{3.40a}$$

$$\overline{\mathcal{D}}^*(Y;\vec{X}) = \frac{\overline{\mathcal{I}}^*(Y;\vec{X})}{\widehat{\mathcal{H}}(Y)} \equiv 1 - \overline{\mathcal{D}}_0(Y;\vec{X}) \,. \tag{3.40b}$$

Feature-selection and bounding criterion can then be instantiated with

$$\mathbb{Q}(Y;\vec{X}) \equiv \langle \widehat{\mathcal{D}}^*_{\mathrm{TCMI}}(Y;\vec{X}) \rangle \,, \qquad \overline{\mathbb{Q}}(Y;\vec{X}) \equiv 1 - \langle \overline{\mathcal{D}}^*_{\mathrm{TCMI}}(Y;\vec{X}) \rangle \,, \tag{3.41}$$

[19] It is to be noted that alternatives to Pearson's R and Spearman's rank ρ correlation, such as the normalized mutual information [84], also require the specification of a threshold and are therefore not suitable for feature identification.

where $\langle\overline{\mathcal{D}}^*_{\mathrm{TCMI}}(Y;\vec{X})\rangle$ denotes the estimated residual empirical cumulative mutual information under the assumption of independent random variables (Eq. 3.40),

$$\langle\overline{\mathcal{D}}^*_{\mathrm{TCMI}}(Y;\vec{X})\rangle := \frac{\hat{\mathcal{D}}_0(Y;\vec{X}) + \hat{\mathcal{D}}'_0(Y;\vec{X})}{2}\ , \quad 0 \leq \langle\overline{\mathcal{D}}^*_{\mathrm{TCMI}}(Y;\vec{X})\rangle \leq 1\ . \tag{3.42}$$

By using branch-and-bound, an optimal feature subset $\vec{X}^* \subseteq \vec{X}$ is identified that has the largest joint mutual dependence with Y. The joint mutual dependence is intrinsically related to the relevance of a feature subset (Secs. 3.2.2 and 3.3.3). As there are possibly many feature subsets that are statistically equivalent in their joint mutual dependence, i.e., they have the score $\langle\widehat{\mathcal{D}}^*_{\mathrm{TCMI}}(Y;\vec{X}')\rangle$ close to the optimal subset $\vec{X}^*$, all these subsets might be equally suited to estimate Y. Therefore, in the experiments below, all features of the topmost feature subsets are considered relevant, if the difference between the score of the subset $\langle\widehat{\mathcal{D}}^*_{\mathrm{TCMI}}(Y;\vec{X}')\rangle$ and the optimal feature subset $\langle\widehat{\mathcal{D}}^*_{\mathrm{TCMI}}(Y;\vec{X}^*)\rangle$ is smaller than 0.01, i.e.,

$$\Delta_{\mathrm{TCMI}}(Y;\vec{X}^*,\vec{X}') = \langle\widehat{\mathcal{D}}^*_{\mathrm{TCMI}}(Y;\vec{X}^*)\rangle - \langle\widehat{\mathcal{D}}^*_{\mathrm{TCMI}}(Y;\vec{X}')\rangle \leq 0.01\ . \tag{3.43}$$

The value 0.01 can be considered as a convergence threshold (here the score must increase by at least more than 1% to be considered a new optimal feature subset). A high value results in terminated the search for the (sub-)optimal feature set too early, a low value leads to an unnecessarily long evaluation time.

Optimal feature subsets may be only weakly related to the property of interest. Therefore feature subsets are considered relevant, only if the value $\langle\widehat{\mathcal{D}}^*_{\mathrm{TCMI}}\rangle$ is larger than or equal to the midpoint of the total range of the score, $\langle\widehat{\mathcal{D}}^*_{\mathrm{TCMI}}(Y,\vec{X}^*)\rangle \geq 0.5$. This threshold is based on the assumption that the correction term is smaller than the joint mutual dependence (Sec. 3.3.2) and is used in the following to decide whether an actual mutual relationship can be reliably determined or not[20].

Recursive feature elimination (RFECV) Recursive feature elimination [263] is a greedy feature-selection method used to eliminate multivariate dependences between features [236, 337]. It is based on backward elimination to first assign weights to features by constructing a machine-learning model (e.g., support-vector machines [195, 196], random forests [266], linear regression [51]), to then remove the features with the lowest weights, and finally to recursively consider smaller subsets of features until the model with the lowest prediction error is identified. In the examples below, recursive feature elimination is combined with the random-forest algorithm [266] to find the optimal feature subset for a machine-learning model. The optimal subset of relevant features is obtained with 10-fold cross-validation (Section 2.5).

Gradient-boosting decision trees with permutation feature importance (FS-GBDT) Ensemble machine-learning methods such as random forest [266] and gradient boosting [249–252] are machine-learning models based on decision trees [198, 199] (cf., Appendix A.2). They estimate a

[20] Instead of using a threshold, one could also sequentially create a statistical model from the identified feature subsets in decreasing order of the score and simply select the feature subset that generates the statistical model with the highest prediction performance (cf., Sec. 5.2).

feature's relevance by using an inbuilt feature-importance measure similar to TCMI and the baseline adjustment for information-theoretic measures (Eq. 3.23). A widely-known feature-importance measure, called permutation feature importance, randomly permutes each feature $X \in \vec{X}$ and compares the resulting model performance before and after permuting X [266, 338]. A feature is considered relevant, if the model performance decreases significantly with the permuted feature, i.e., when the feature is related to the property of interest. A feature is considered redundant if the prediction performance of the model remains unaffected, i.e., when the property of interest is independent of the feature or the feature is multi-collinear related to another feature in the feature subset. In the examples below, permutation feature importance is computed with gradient boosting [252, 267, 268] using scikit-learn [44] and the `rfpimp`[21] package [100].

Feature selection using the sure-independence screening and sparsifying operator (FS–SISSO) The sure-independence screening and sparsifying operator (SISSO) [197] combines symbolic regression with compressed sensing [191] to create billions of candidate feature combinations prior to selecting those that are suitable for estimating the property of interest (cf., Appendix A.1). Originally designed for generating deterministic symbolic-regression models based on algebraic operations of analytical functions, SISSO has since been used to relate the selected features or feature combinations of the model to the property of interest [340–344]. However, the relevance of features or feature combinations may be tightly coupled to the created SISSO model and the chosen parameter settings (cf., Sec. 3.1.2). Therefore, a hyper-parameter optimization is performed to identify the relevant features of a data set by finding the model with the optimal parameter settings (cf., Section 2.4). In all experiments, relevant features are determined with 10-fold cross-validation using all of the available algebraic operators (cf., Appendix A.1). The number of symbolic-regression terms is determined by applying the operator set recursively up to three times ($rung \leq 3$)[22], while keeping only those terms with a maximum of $maxcomplexity = 5$[22] features and $subs\_sis = 300$[22] feature combinations in each iteration (cf., Appendix A.1).

3.5.3 Examples

Practical examples from materials science face the problem of statistically modeling a relationship between a set of features and a targeted property without knowing the underlying process that generated the data. Idealized examples by contrast are often oversimplified and therefore have difficulties in capturing the complexity of materials data. Unfortunately, there are no materials-science data sets that can be used to compare different feature-selection methods on the basis of a common ground truth. Therefore, three examples focusing on different aspects of feature identification are discussed to evaluate the feature-selection methods for their applicability to materials-science applications. In the first example, the feature-selection methods have to identify the two out of ten features to accurately model a bivariate normal distribution. The second example focuses on the identification of interrelated relevant features. The third example examines methods for selecting features based on experimentally collected materials data where the underlying relationships between

[21] The `eli5` package [339] provides similar functionality and can be used instead of the `rfpimp` package [100].

[22] The name refer to the setting defined in the Fortran code of the SISSO paper [197].

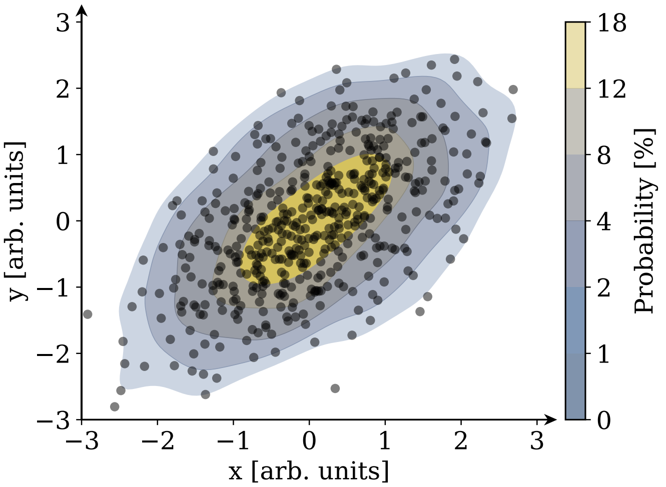


Fig. 3.9. Bivariate normal probability distribution with mean $\vec{\mu} = \left[0; 0\right]$ and covariance matrix $\mathbf{\Sigma} = \left[1\ 0.5; 0.5\ 1\right]$. A scatter plot with 500 data samples is shown with contour lines of equal probability densities $(\vec{x}; \vec{\mu}, \Sigma) \in \{0.01, 0.02, 0.05, 0.08, 0.13\}$.

the features and the property of interest are not directly known, but sufficient knowledge exists to empirically relate the features of the data set to the targeted materials property.

Bivariate normal distribution

The identification of a bivariate normal distribution in a high-dimensional space focuses on the ability of TCMI and feature-selection techniques to identify non-linear relationships even when other unrelated features are present in the data set. For this purpose, $n = 500$ samples are drawn from a bivariate normal distribution (Fig. 3.9a.) with zero mean $\vec{\mu} = \left[0; 0\right]$ and covariance matrix $\mathbf{\Sigma} = \left[1\ 0.5; 0.5\ 1\right]$,

$$\mathcal{N}(\vec{x}; \vec{\mu}, \mathbf{\Sigma}) = \frac{1}{\sqrt{4\pi^2 \det \mathbf{\Sigma}}} \exp\left[-\frac{1}{2}(\vec{x} - \vec{\mu})^{\intercal} \Sigma^{-1} (\vec{x} - \vec{\mu})\right] , \qquad \vec{x} = \left[X_1; X_2\right] . \tag{3.44}$$

Likewise, n samples each from a univariate normal and scalar exponential, logistic, triangular, uniform, Laplace, Rayleigh, Poisson, and Weibull distribution are drawn and added to the list of available features, all with zero mean $\mu = 0$ and identity covariance matrix $\Sigma = 1$. In total, the data set consists of 11 features $\vec{X} = \{X_1, X_2, \ldots X_{11}\}$, of which only $\vec{x} = \{X_1, X_2\}$ are related to $Y = \mathcal{N}(\vec{x}; \vec{\mu}, \mathbf{\Sigma})$. Tests are performed on 50, 100, 200 and 500 data samples in order to investigate the dependence of the feature-selection criterion on the number of data samples. Results are reported in Tab. 3.1.

In terms of Pearson's or Spearman's rank correlation coefficient, none of the features $X_i \in \vec{X}$ have coefficients of determinations higher than 1% with respect to the bivariate normal distribution Y, i.e., $R^2(Y, X_i) < 0.01$ or $\rho^2(Y, X_i) < 0.01$. As there are no features in the data set that are pairwise non-linear monotonically related to Y, the challenge of feature-selection methods is therefore to identify the features X_1 and X_2 from their non-linear relationship to Y.

#	Dependence	Features										
	Measure	$\mathbf{X_1}$	$\mathbf{X_2}$	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
50 samples	TCMI	✓*				✓*						
	RFECV	✓	✓									
	FS-GBDT		✓		✓							
	FS-SISSO	✓	✓					✓				
100 samples	TCMI		✓*					✓*				
	RFECV	✓	✓									
	FS-GBDT	✓	✓			✓		✓	✓			
	FS-SISSO	✓	✓					✓		✓	✓	
200 samples	TCMI	✓	✓									
	RFECV	✓	✓									
	FS-GBDT	✓	✓			✓	✓		✓			
	FS-SISSO	✓	✓					✓				
500 samples	TCMI	✓	✓									
	RFECV	✓	✓									
	FS-GBDT	✓	✓									
	FS-SISSO	✓	✓									

Tab. 3.1. A table of identified feature (subsets) for the prediction of the bivariate normal distribution obtained by total cumulative mutual information (TCMI, Section 3.3.3), recursive feature elimination (RFECV) [263], gradient boosting (FS-GBDT) [249–252], and the sure-independence screening and sparsifying operator (FS-SISSO) [197] for different number of data samples. The relevant features of the bivariate distribution are typeset in bold ($\mathbf{X_1}$ and $\mathbf{X_2}$). Features of the data set also include the normal X_3, exponential X_4, logistic X_5, triangular X_6, uniform X_7, Laplace X_8, Rayleigh X_9, Poisson X_{10}, and Weibull X_{11} distribution. Features marked with an asterisk (✓*) have a TCMI value smaller than $\langle \widehat{\mathcal{D}}^*_{\mathrm{TCMI}} \rangle < 0.5$.

TCMI, RFECV, FS-GBDT, and FS-SISSO correctly identify the relevant feature subset $\vec{X}^* = \{X_1, X_2\}$ of the bivariate normal distribution as the number of data samples increases to 500 (Tab. 3.1). Using only 50 data samples, RFECV is the only method that correctly identifies X_1 and X_2 as relevant. Between 50 and 500 data samples, FS-GBDT as well as FS-SISSO identify a superset $\vec{X}'$ of the relevant feature subset $\vec{X}^* = \{X_1, X_2\}$. TCMI requires slightly more data samples than any of the other feature-selection methods (more than 100 data samples) as its score is intrinsically linked to the strength of the mutual dependence between the features and the property of interest: TCMI starts with a score close to independence for 50 data samples and increases slowly to a score of $\langle \widehat{\mathcal{D}}^*_{\mathrm{TCMI}} \rangle = 0.6$ for 500 data samples, indicating that more samples than 500 are needed for TCMI to actually reliably identify the relevant features of this data set.

Friedman regression data set

The second feature-selection problem addresses the identification of relevant features in the presence of multi-collinear features by using the Friedman regression data set [347]. The data set has five

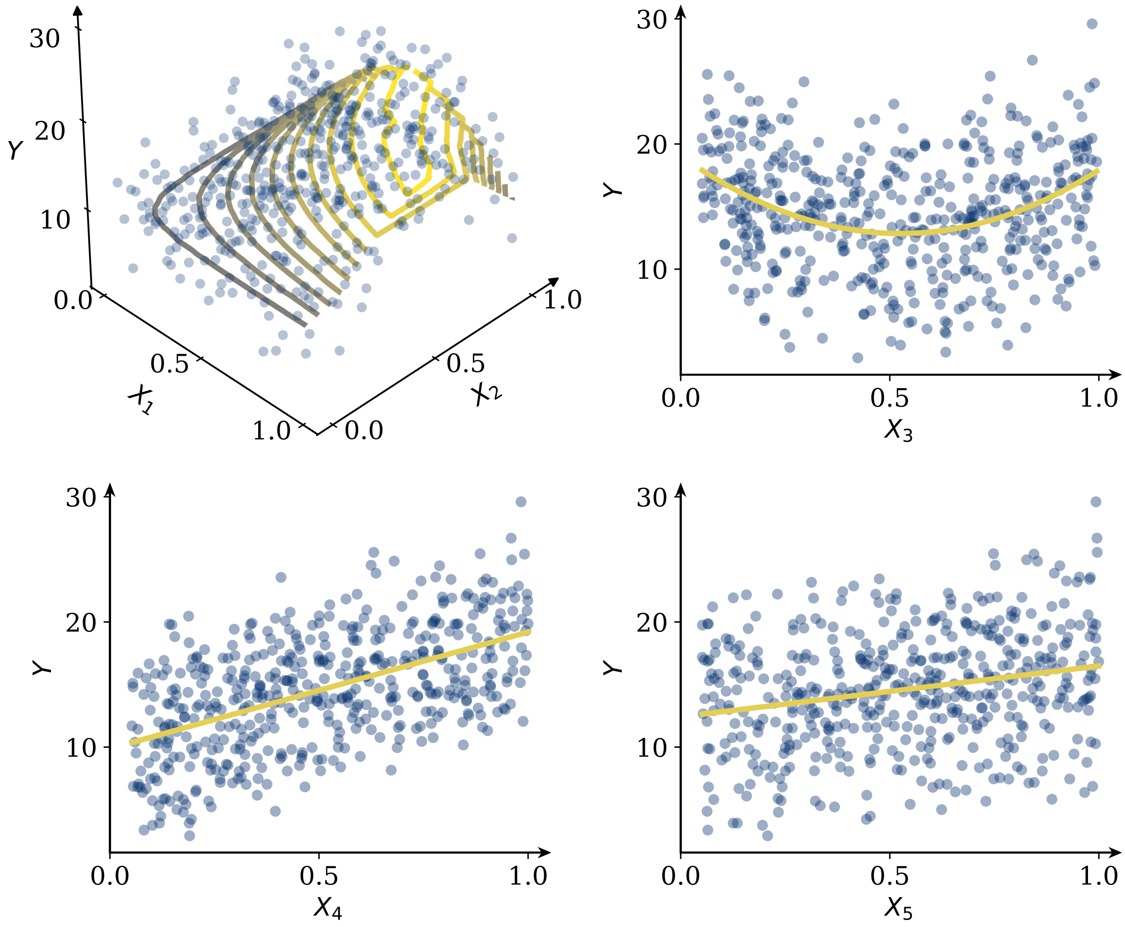


Fig. 3.10. Graphical representation of the Friedman data set [347]. Shown are the cross-sections of the target function (Eq. 3.45) with the five relevant features $X_1, \ldots, X_5$ as contours or lines and the distribution of the 500 data samples in the six-dimensional hypercube $X_i \in [0, 1]$.

relevant features $X_1, \ldots, X_5$ that are randomly generated from univariate uniform distributions of which three features are non-linearly and two features are linearly related to the response $Y = f(\vec{X})$,

$$Y = 10 \sin(\pi X_1 X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \epsilon \ . \tag{3.45}$$

Here, ϵ is an error term (cf., Eq. 2.1) that is modeled as the standard normal deviate $\mathcal{N}(0, 1)$. The function is evaluated on the hypercube, $X_i \in [0, 1]$ (Fig. 3.10). Originally, five univariate uniform distributions $X_6, \ldots, X_{10}$ unrelated to Y were added to increase the number of the features in the data set. In this example, Eq. 3.45 is considered with four redundant features $X_{11}, \ldots, X_{14}$ in addition, which are strongly correlated with $X_1, \ldots, X_4$ and were generated by $g(X_i) = X_i + \mathcal{N}(0, 0.025)$.

Again results show that none of the features are highly pairwise non-linear monotonically related with the property of interest in terms of Pearson's or Spearman's coefficient of determination ($R^2(Y, X_i), \rho^2(Y, X_i) \leq 0.4$) and that TCMI, RFECV, FS-GBDT, and FS-SISSO correctly identify all of the relevant features as the number of data samples increases to 500 (Tab. 3.1). More specifically, TCMI

#	**Dependence**								**Features**						
	Measure	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	$\mathbf{X_5}$	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}
50 samples	TCMI						✓*		✓*	✓*					
	RFECV												✓		✓
	FS-GBDT	✓	✓		✓										
	FS-SISSO	✓	✓	✓		✓							✓	✓	✓
100 samples	TCMI				✓	✓									✓
	RFECV		✓		✓	✓						✓	✓	✓	✓
	FS-GBDT	✓	✓	✓	✓	✓						✓	✓		✓
	FS-SISSO	✓	✓	✓		✓									✓
200 samples	TCMI		✓		✓	✓							✓		✓
	RFECV	✓	✓	✓	✓	✓						✓	✓		✓
	FS-GBDT	✓	✓	✓	✓	✓						✓	✓	✓	✓
	FS-SISSO	✓	✓	✓	✓	✓						✓			
500 samples	TCMI	✓	✓	✓	✓	✓						✓	✓	✓	✓
	RFECV	✓	✓	✓	✓	✓						✓	✓	✓	✓
	FS-GBDT	✓	✓	✓	✓	✓						✓	✓	✓	✓
	FS-SISSO	✓	✓	✓	✓	✓						✓			✓

Tab. 3.2. Identified features for predicting the response function (Eq. 3.45) of the Friedman regression data set [347] obtained by total cumulative mutual information (TCMI, Section 3.3.3), recursive feature elimination (RFECV) [263], gradient boosting (FS-GBDT) [249–252], and the sure-independence screening and sparsifying operator (FS-SISSO) [197] for different number of data samples. Features marked with an asterisk (✓*) are from a feature subset with a TCMI value smaller than $\langle \widehat{\mathscr{D}}^{*}_{\text{TCMI}} \rangle < 0.5$. The relevant features of the Friedman regression data set are typeset in bold ($X_1 - X_5$). This data set also has four redundant features ($X_{11} - X_{14}$) and five irrelevant features ($X_6 - X_{10}$).

identifies less relevant features with the same number of data samples as compared to the other feature-selection techniques. Incidentally, repeating the experiment with a different randomization of the data samples also shows that TCMI is unstable with less than 100 data samples, whereas RFECV, FS-GBDT and FS-SISSO are unstable in the feature selection with less than 500 data samples. In particular with data sets less than 200 data samples, they identify redundant features and sometimes one of the unrelated features $X_6, \ldots, X_9$ as relevant. This feature-selection instability is due to the fact that the identification task is not unique and the features X_{1i}, $i \in \{1, 2, 3, 4\}$ are collinear related to X_i. As such, collinear features are interchangeable in their relevance and only one of the two features is needed to effectively estimate the target property Y. Whereas this issue plays no role in the statistical modeling of the feature-property relationships, it poses serious problems in the identification of relevant features and therefore may require to remove multi-collinear features prior to machine learning, e.g., with feature-dependence maps [100] possibly in combination with minimal-redundancy-maximal-relevance schemes [230].

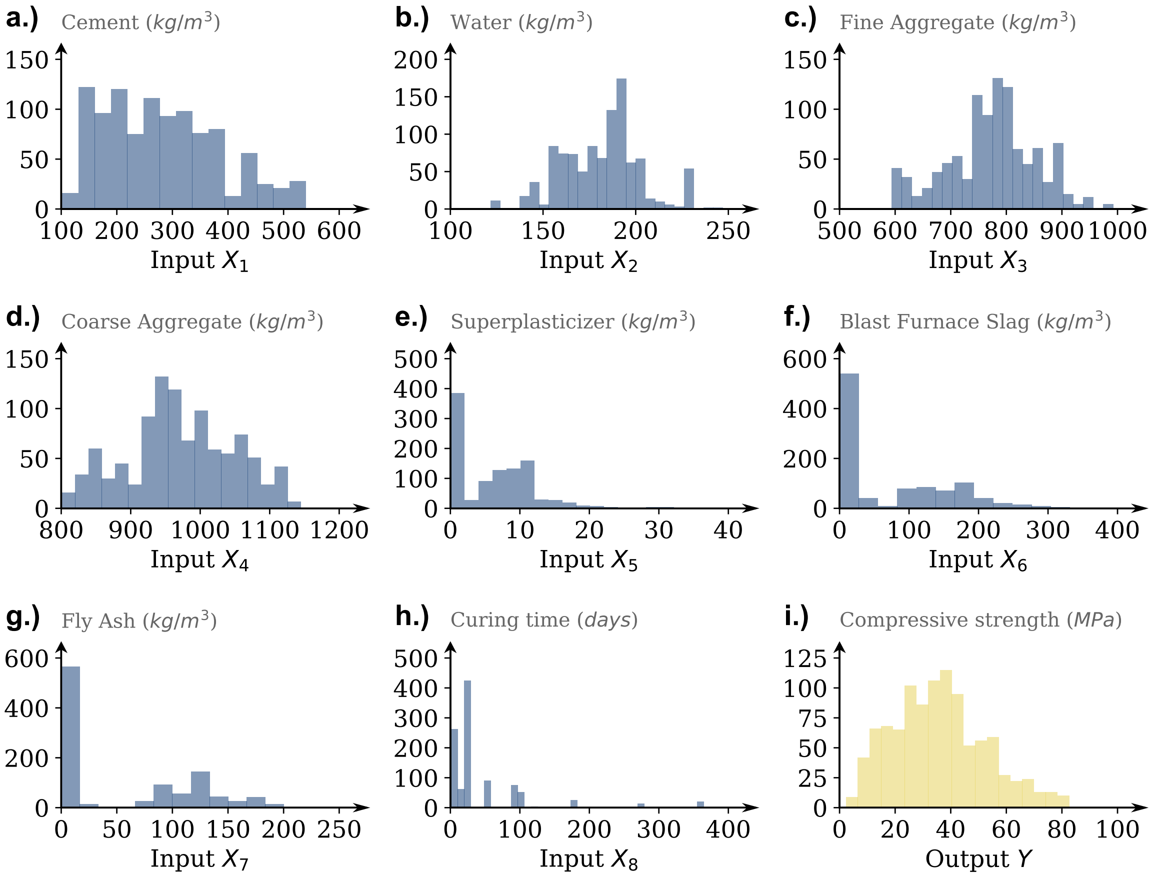


Fig. 3.11. Histogram (frequency counts) of the statistical distributions of the features $X_1 \dots X_8$ and the target property Y in the concrete compressive strength data set [350, 351].

High-performance concrete

In this example, the behavior of concrete structures under external loads is studied using the compressive strength, the most fundamental property of concrete. Concrete consists of Portland cement, water, and an aggregate (sand, gravel, or crushed stones) and is one of the most widely used and versatile building materials in the world [348]. The making of high-performance concrete includes supplementary cementitious materials, such as fly ash (a coal combustion product), blast furnace slag (a by-product of iron and steel production), and chemical admixtures to improve the flow characteristics of concrete, such as superplasticizers. As high-performance concrete is a composite material that exhibits a strongly non-linear relationship with compressive strength, it is of great importance to build a predictive model to estimate the compressive strength for specific applications.

A total of eight features $\vec{X} = \{X_1, \dots, X_8\}$ (Fig. 3.11) and $n = 1030$ data samples from 17 different sources were compiled and tested to estimate the compressive strength of high-performance concrete [350, 351]. Although the underlying relationships between the ingredients and the compressive strength are not known, this data set is relatively well understood: cement X_1, water X_2, blast-furnace slag X_6, and curing time X_8 are the most relevant features in estimating the compressive strength

#	Dependence Measure	Features							
		X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
50 samples	TCMI	✓*	✓*		✓*	✓*		✓*	
	RFECV	✓		✓		✓			✓
	FS-GBDT	✓	✓	✓	✓		✓		✓
	FS-SISSO	✓	✓	✓	✓	✓	✓		✓
100 samples	TCMI	✓		✓	✓	✓	✓		✓
	RFECV	✓	✓	✓					✓
	FS-GBDT	✓	✓	✓		✓	✓		✓
	FS-SISSO	✓	✓	✓	✓		✓		✓
200 samples	TCMI	✓	✓	✓	✓	✓	✓	✓	✓
	RFECV	✓	✓	✓		✓			✓
	FS-GBDT	✓	✓	✓	✓	✓			✓
	FS-SISSO	✓	✓	✓	✓	✓	✓		✓
500 samples	TCMI	✓	✓	✓	✓	✓	✓	✓	✓
	RFECV	✓	✓	✓	✓	✓	✓	✓	✓
	FS-GBDT	✓	✓	✓	✓	✓	✓	✓	✓
	FS-SISSO	✓†	✓†				✓†		✓†

Tab. 3.3. Identified features for the prediction of the compressive strength of high performance concrete [350, 351] obtained by total cumulative mutual information (TCMI, Section 3.3.3), recursive feature elimination (RFECV) [263], gradient boosting (FS-GBDT) [249–252], and the sure-independence screening and sparsifying operator (FS-SISSO) [197] for different number of data samples. The features in the dataset are: cement X_1, water X_2, fine aggregate (sand) X_3, coarse aggregate (gravel, crushed stone) X_4, superplasticizer X_5, blast-furnace slag X_6, fly ash X_7, and curing time X_8. Features marked with an asterisk (✓*) are from a feature subset with a TCMI value smaller than $\langle \widehat{\mathcal{D}}^*_{\text{TCMI}} \rangle < 0.5$. Features marked with a dagger (✓†) are from a SISSO model f_{SISSO} with a Pearson's coefficient of determination of $R^2(Y, f_{\text{SISSO}}(\vec{X}^*)) < 0.9$.

of high-performance concrete [348, 350, 351]. Furthermore, lower water content in relation to the cement content leads to a stronger concrete [352]. However, studies have also shown that all features are strongly multi-collinear interrelated [348]. It is thus not surprising that TCMI, RFECV, and FS-GBDT identify all features of the data set as relevant as the number of data samples increases to 500 (Tab. 3.3), whereas FS-SISSO emphasizes the relevance of cement, water, slag, and curing time. Feature identification therefore faces the challenge that multi-collinear features can seriously impact the performance of these methods (i.e., larger feature subsets must be evaluated) and these also may prevent the identification of feature subsets with fewer features than present in the data set (as in the case of TCMI, RFECV, and FS-GBDT).

Using machine-learning algorithms for feature identification also show that a feature identification may be effected by materials which cannot be reliably estimated by a global statistical model such as with symbolic regression. When searching for relevant features, it is then seen that the prediction performance of a global model decreases significantly as the number of data samples increases. This is apparent, for example, for FS-SISSO: the prediction performance of the generated SISSO model degrades significantly from $R^2(Y, f_{\text{SISSO}}(\vec{X}^*)) = 0.92$ for 200 data samples to $R^2(Y, f_{\text{SISSO}}(\vec{X}^*)) = 0.8$ for 500 data samples. The reason is that the top 10% of materials with the largest prediction errors (which cannot be accurately represented by the SISSO model) are characterized by an average 2.4 times higher age, 1.6 times higher slag, and 1.6 times lower fly ash content than the other materials in the data set. Since a hyperparameter optimization is performed for FS-SISSO (by varying the number of symbolic regression terms and the number of features in each term), a symbolic regression model is selected for 500 data samples that has a smaller number of terms and features than for 200 data samples. Applying a feature-selection method to real (material) data sets therefore faces the challenge that, due to different groups or classes of material, a hyperparameter optimization during feature identification potentially leads to models with lower prediction performance the more data samples are used. Therefore, when different groups or classes of materials are likely to be present in a data set and the feature-identification method needs to be hyperoptimized, local models or dedicated methods using data-mining tools such as subgroup discovery [353–356] must be utilized when searching for the relevant features of the property of interest.

3.6 Discussion

Feature identification is an important part of a successful data-mining pipeline to identify the features that are related to the property of interest [228]. These features can then be used to build less complex statistical models with lower computational costs for statistical estimation than building a model from the full set of initial features of a data set.

Since it is difficult to identify the features related to the property of interest based on a non-linear combination with feature extraction [227], feature selection [71, 72] is used to identify the relevant features from the initial set of features (Section 3.1). Feature-selection strategies can be categorized into three groups, depending on how the relevance of a feature subset is estimated [72, 260]: filters, wrappers, and embedded methods (Fig. 3.2). Targeted to best identifying the set of features for the property of interest, embedded methods [265] simultaneously perform feature selection during model construction, while wrapper methods [72] evaluate the relevance of a set of features tailored to a specific machine-learning algorithm. In data sets with a large number of features [211, 244–246], however, both embedded and wrapper methods are computationally demanding and are sensitive to multi-collinear features [99, 357] (Section 3.5). In contrast, filter methods can identify feature subsets independently of a machine-learning model, thus providing the freedom to choose a machine-learning algorithm that best works for the prediction of a property of interest [86–89] (Section 3.2).

Typical data sets can have hundreds to thousands of features. For most problems, the number of features for an optimal feature subset varies considerably with the number of data samples and the property of interest [51, 52]. Therefore, a manual identification of possible candidate feature subsets is often not feasible. Branch-and-bound [91–94, 96] has been proven to be efficient in the search of candidate feature subsets [97, 294] and guarantees to find the optimal set of features, without evaluating all possible subsets (Section 3.1.2). It requires a feature-selection criterion that satisfies the monotonicity condition in [92]. Because most evaluation metrics for machine learning are non-monotonically increasing and the search for relevant features resembles that of information-theoretic concepts, total cumulative mutual information (TCMI, Section 3.3.3) was developed as a monotonic increasing feature-selection criterion. TCMI combined with branch-and-bound finds the optimal set of features by quantifying the mutual dependence between a set of features and the property of interest. Applied to each feature-subset combination in the data set, it can be used to rank the feature subsets in descending order of their strength of mutual dependence.

Three examples focusing on different aspects of feature identification were discussed and TCMI was compared with existing methods for identifying relevant features based on machine learning (RFECV, FS-GBDT, FS-SISSO). In all three examples, the relationship between the features and the property of interest were either sufficiently known through empirically relating the features of the data set to the targeted materials property (concrete data set) or the relevant features were directly accessible (bivariate normal distribution, Friedman regression data set). Feature-selection methods couldcan therefore be compared and tested for their applicability based on a common ground truth. All tested feature-selection methods produce consistent results as the number of data samples increases. However, as the examples have shown, machine learning-based feature-selection methods are sensitive to multi-collinear features (Tabs. 3.2 and 3.3): RFECV, FS-GBDT, and FS-SISSO may select features that are not relevant at all (bivariate distribution, Tab. 3.1) and FS-SISSO has the tendency to omit redundant features, especially when the data set is small (Friedman regression data set Tab. 3.2, concrete data set Tab. 3.3). The investigated feature-selection methods based on machine learning therefore provide no reliable indicator of the relevance of all features in a data set.

In contrast, TCMI is intrinsically linked to the mutual dependence of the features and the property of interest and performs similarly well to established feature-selection methods (Section 3.5). In particular, the score of TCMI provides an indicator of the quality of the identified feature (subsets) and enables an assessment of when more data samples or better descriptive features are needed (Sec. 3.3.2). Especially when the cardinality of the feature subsets is high and the relevance scores are low, TCMI indicates that regardless of the machine-learning algorithm, better descriptive features are needed for a given application.

In some cases like in the bivariate normal distribution, TCMI identifies the relevant features of the data set with a smaller amount of data than FS-GBDT or FS-SISSO, but requires more data samples than RFECV. However, in the Friedman regression data set, TCMI is the only method that falsely assigns relevance to unrelated features with only 50 data samples, albeit with a low dependence score. In the concrete data set, TCMI first identifies a set of features with a low dependence score, but then identifies all features as relevant with less number of data samples than RFECV or FS-GBDT. TCMI successively identifies more features as relevant as the number of data samples increases, but requires larger amount of data to identify the same number of relevant features than the investigated feature-selection methods based on machine learning (RFECV, FS-GBDT, FS-SISSO). In the worst case, this results in machine-learning models, which potentially have larger prediction errors than models created with features from feature-selection methods using machine learning. In contrast, RFECV, FS-GBDT, and FS-SISSO have been shown to identify the related features to the property of interest with the least number of data samples. However, identified relevant features in RFECV, FS-GBDT, and FS-SISSO vary considerably (Tabs. 3.2 and 3.3), especially when there are multi-collinear features [99] and the data sets are small (Tabs. 3.1 and 3.3). The number of data samples and the interrelationships between features therefore largely determine the applicability of feature-selection methods. A framework is therefore developed in the next chapter to reliably relate the relevance of a feature subset to the prediction performance of statistical models independent of the applied feature-selection algorithm.

Chapter 4

A framework for feature identification and model construction

Feature identification is a multi-faceted combinatorial optimization problem of the search strategy, the feature-selection method, and the assessment of the relevance of features and feature subsets [260]. Therefore, a framework is developed to provide a relevance and uncertainty estimation of possible feature subsets on the basis of a limited number of materials within a probabilistic tolerance. The framework is explained in the following and its application is discussed in the next Chapter (Chapter 5). The extensive and systematic search for features related to the property of interest within an automated framework for machine learning is a combination of available methods and new approaches

- for applying feature identification independently of the feature-selection algorithm,
- for visualizing multivariate statistical relationships of multivariate features,
- for estimating the uncertainty of statistical models using prediction intervals for each new prediction of the models,
- for identifying so-called anomalous materials whose property of interest cannot be accurately estimated by a statistical model based on the set of available data samples.

As discussed in Section 3.1.2: there are three types of feature-selection methods: filter, wrapper, and embedded methods. All methods have their own advantages and disadvantages in identifying the relevant features for a specific application. Filter methods, on the one hand, have the advantage of identifying features without making assumptions about the actual feature-property relationships in the data. However, machine-learning models built on the identified sets of features from filter methods can have very little predictive power. First, because feature identification and model construction are two completely different tasks [243, 358–360]. And second, because the feature sets identified may not contain all the features that a machine learning algorithm needs to create the most accurate model from the set of data (cf., Section 3.5).

Wrapper and embedded methods, on the other hand, identify the related features to the properties of interest based on an evaluation measure of a machine-learning model (e.g., from the Pearson's coefficient of determination, root-mean-squared error, etc.). Therefore, wrapper and embedded

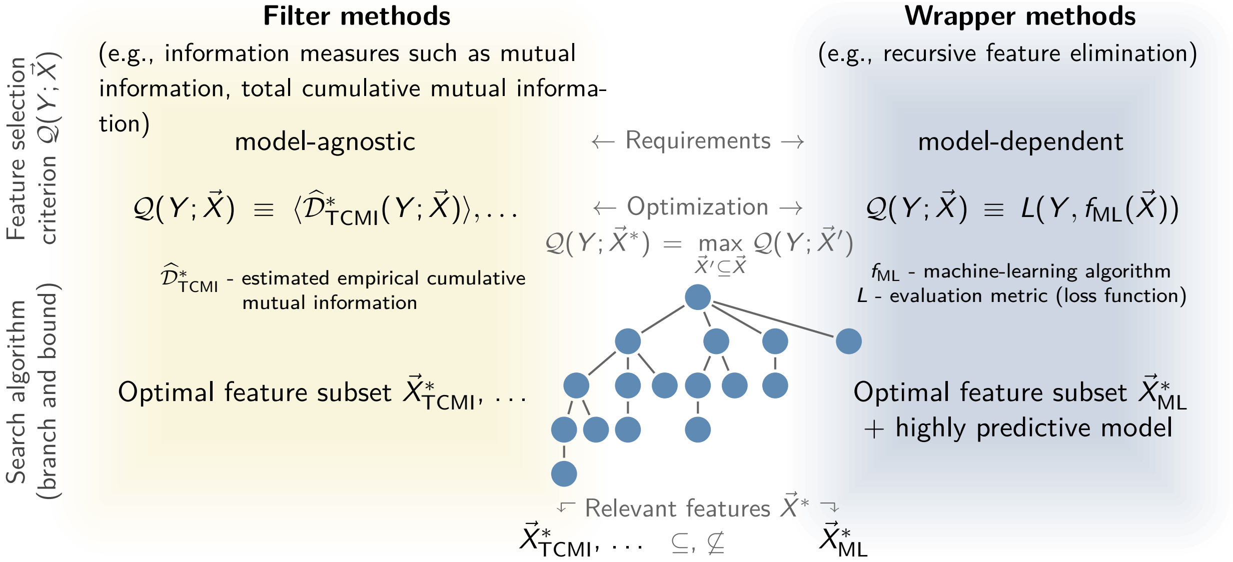

Filter methods
(e.g., information measures such as mutual information, total cumulative mutual information)
model-agnostic
$\mathcal{Q}(Y;\vec{X}) \equiv \langle\widehat{\mathcal{D}}^*_{\mathrm{TCMI}}(Y;\vec{X})\rangle, \ldots$
$\widehat{\mathcal{D}}^*_{\mathrm{TCMI}}$ - estimated empirical cumulative mutual information
Optimal feature subset $\vec{X}^*_{\mathrm{TCMI}}, \ldots$
Wrapper methods
(e.g., recursive feature elimination)
model-dependent
$\mathcal{Q}(Y;\vec{X}) \equiv L(Y, f_{\mathrm{ML}}(\vec{X}))$
f_{ML} - machine-learning algorithm
L - evaluation metric (loss function)
Optimal feature subset $\vec{X}^*_{\mathrm{ML}}$
+ highly predictive model
← Requirements →
← Optimization →
$\mathcal{Q}(Y;\vec{X}^*) = \max_{\vec{X}' \subseteq \vec{X}} \mathcal{Q}(Y;\vec{X}')$
Feature selection criterion $\mathcal{Q}(Y;\vec{X})$
Search algorithm (branch and bound)
Relevant features $\vec{X}^*$
$\vec{X}^*_{\mathrm{TCMI}}, \ldots \quad \subseteq, \not\subseteq \quad \vec{X}^*_{\mathrm{ML}}$

Fig. 4.1. A flow diagram of the proposed feature-selection workflow with filter and wrapper methods using the same search strategy, namely the branch-and-bound algorithm [91, 92] (Section 3.4), for identifying relevant feature subsets for the prediction of the properties of interest.

methods are more efficient than filter methods in terms of identifying feature-property relationships for the statistical modeling of the data. However, the limited availability of data and multivariate features may complicate the identification of related features to the properties of interest with machine learning (Section 2.6). Especially when the data sets contain a very small number of samples[1], while having a large number of features (as is the case in most of materials science), machine-learning algorithms may therefore provide insufficient results.

To actually compare identified feature subsets in a common framework applicable to filter and wrapper methods, in this thesis, a search strategy is proposed that uses the same search strategy independently of the feature-selection criterion (Fig. 4.1). As the branch-and-bound search algorithm (Section 3.4) is in general not limited to information-based measures [361], it is adapted for use with a machine-learning algorithm to identify the set of features related to a property of interest.

A machine-learning-based feature selection with the branch-and-bound algorithm has the advantage of constructing machine-learning models with a successively increasing number of features and time complexity in the evaluation. As the evaluation with machine-learning models is performed from simple to more complex feature subsets, optimal feature subsets can be found quickly with accurate predictive capabilities, avoiding potential problems with (non-linear) interrelated features [59, 362] or high-dimensional materials spaces [211, 244–246]. The identified feature subsets can then be used to assess and optimize the materials representation (Chapter 3), to visualize non-linear statistical relationships in the data (Section 4.2), or to build an ensemble of statistically equivalent models to robustly and reliably predict the property of interest within a probabilistic tolerance (Sections 4.3 and 4.4).

[1] The size of the data set is estimated by its ratio to the number of features. An exact number of samples is therefore difficult to estimate. As a rough estimate, data sets with less than 1000 data samples can always be considered small.

The prediction performance of a machine-learning model is commonly expressed in terms of a single metric (such as the root-mean-squared error, mean-absolute error, etc.) using resampling techniques (Section 2.5) such as cross-validation [52, 201] or bootstrapping [203–205]. However, a single metric, while useful for estimating the goodness-of-fit of the model, does not provide any information about the actual error between the model's prediction and the property of interest [363]. Hence, it can be expected that the model's prediction may vary significantly from the actual value of the property of interest in regions for which there are not enough data (cf., [208]). Consequently, an uncertainty estimate is desirable and even necessary in many materials-science applications [104] to evaluate the consistency of the model with the data [364], to increase the reliability and credibility of the model's estimations [102, 103], and to identify regions of the materials space that are underrepresented or are promising to explore [363, 365]. There are a number of *ad hoc* approaches for generating uncertainty estimates. For instance, Gaussian process regression [108, 109] and deep neural networks [105–107] are two commonly used machine-learning algorithms that provide uncertainty estimations either based on Bayesian probabilities or on distributions from the internal parameters of the model.

A general methodology applicable to any machine-learning algorithm for estimating the uncertainty of individual statistical-model predictions is conformal prediction (cf., Section 4.3.1). Conformal prediction [62–64] uses the same set of data as for statistical modeling to determine precise levels of confidence for new predictions within a probabilistic tolerance – and can therefore be employed not only to identify regions of the materials space that are underrepresented or cannot be accurately modeled by the specified machine-learning models, but also to identify materials which are potentially subject to some other mechanism of the materials behavior (cf., Section 4.4).

This chapter lays out the foundations for developing a common framework (Section 5.1) for identifying relevant features and building data-driven statistical models (Section 5.2). It introduces and discusses practical algorithms to measure the relevance of feature subsets with information measures or machine-learning algorithms (Alg. 4.1), to quantify multivariate feature relationships (Section 4.2), to estimate the uncertainty in the predictive modeling (Alg. 4.2), and to identify materials whose property of interest are difficult to estimate based on a generated statistical model (Alg. 4.3). Section 4.1 starts by extending the branch-and-bound algorithm to non-monotonic increasing feature-selection criteria (cf., Section 3.4). In order to visualize multivariate feature relationships, feature-dependence maps are introduced in Section 4.2 and combined with the developed model-agnostic feature-identification method as described in Section 4.1. The uncertainty of the predictive modeling, discussed in Section 4.3, is based on a conformal prediction [60–64, 366] to estimate the range of a prediction within a probabilistic tolerance. Based on that, materials can be identified whose property of interest cannot be accurately predicted by a statistical model. To this end, a heuristic is devised in Section 4.4 to determine how different a material is compared to a set of other materials in a data set. The complete framework is then discussed in more detail in the next Chapter (Chapter 5) along with practical applications from materials science (Section 5.2).

4.1 A general feature-identification workflow for materials science

In materials-science applications with hundreds and thousands of features, dimensional-reduction techniques (Section 3.1) are often a necessary step in modeling complex relationships in the data. There are applications where the focus of dimensionality reduction (cf., Section 3.1) is primarily on reducing the computational cost and the prediction error of the machine-learning models to accelerate the screening of new materials. These applications include recommendation systems[2] and high-throughput approaches in which machine-learning algorithms are used in a calculation funnel, where at each level selection criteria (stability, costs, environmental sustainability, etc.) rule out materials and progressively more computationally or experimental intense methods are used to determine candidate materials in a multi-objective optimization [137]. The search for the relevant features for a particular application or a machine-learning algorithm with dimensional reduction (more specifically feature selection, cf., Section 3.1.2) instead requires relating the (constructed) features to the property of interest via a mathematical expression or a statistical model based on the features of a data set. The crucial point is that a feature identification prior to statistical modeling should not significantly degrade the prediction performance of a machine-learning model, as otherwise existing and potentially important feature-property relationships may not be captured and the machine-learning model may not be applicable to new data.

As with all other feature-selection criteria for filter methods, information measures identify relevant features independent of the machine-learning model (Fig. 4.1). However, the identified features may not necessarily be related to the prediction performance (or errors) of the generated machine-learning models (especially if the machine-learning algorithm makes assumptions about the data that are not reflected in the filter method). In addition, information measures may require more data than machine-learning algorithms to identify the same features related to the properties of interest (cf., Section 3.5). Since in data-driven materials-science applications the amount of data is often limited and the feature relationships are multifaceted and intricate, information measures – despite providing a non-parametric and deterministic characterization of statistical relationships in the data on the basis of a rigorous mathematical theory of feature relevance (Section 3.2) – may not be practical. In contrast, wrapper methods using machine-learning algorithms are capable of estimating the properties of interest even with small amounts of data and of performing dimensionality reduction, while optimizing the machine-learning model for the specific materials-science application under consideration [55, 59, 161, 191, 197, 221, 223, 224, 292]. However, machine-learning algorithms are typically heuristic in that they usually do not guarantee that all features required in the statistical modeling are actually statistically related to the property of interest (cf., Section 3.5). This is because the relevance of strongly related features can be spread across multi-collinear features and therefore features of lower relevance may appear stronger related to the property of interest for the machine-learning algorithm than the actual related features of the data set [99].

Multi-collinear features are frequently encountered in computational materials science: most often, features are constructed from elemental or structural properties of a compound, which are implicitly determined by the Kohn-Sham equations using the atomic species, charges, and positions

[2] Recommendation systems based on active learning [102–104] can identify candidate materials even if the feature representation is not optimal or machine-learning models are not highly accurate [58, 103, 367, 368].

as the only physically relevant input variables (cf., Section 2.2). It is therefore desirable to apply a feature-identification method that closely links the identification of related features to the prediction performance of the machine-learning model, even in the presence of multi-collinear features. As with TCMI, this is an optimization problem that can be addressed by utilizing the same search search strategy as in Section 3.1.2 in a common framework for feature selection, henceforth referred to as the feature-identification framework for computational materials science.

4.1.1 The algorithm

The foundation of the feature-identification framework for computational materials science is a generalization of the branch-and-bound algorithm to non-monotonic feature selection criteria (cf., Section 3.4). The proposed novel approach to generalize branch-and-bound algorithm is applicable to any information-theoretic method and machine-learning algorithm for identifying the subset of features related to the property of interest (cf., Section 5.2). Based on the branch-and-bound algorithm (cf., Alg. 3.1), the generalized algorithm evaluates all possible feature combinations, starting with an empty set and successively enlarging the feature subset for as long as the feature-selection criterion increases within a probabilistic tolerance. The feature-selection criterion of the original branch-and-bound algorithm must be a monotonically increasing function (Eq. 3.35). However, many of the objective criteria, such as for machine-learning algorithms, are non-monotonically increasing, which leads to the pruning of potential solutions in the feature-subset search [237, 361]. Therefore, the presented generalized branch-and-bound algorithm introduces a tolerance threshold for the feature-selection criterion to continue the search beyond otherwise infeasible subsets (i.e., those which have a lower value of the bounding criterion than the feature-selection of the current best feature subset in the search, cf., Section 3.4).

The idea of introducing a tolerance parameter for the branch-and-bound algorithm is not new. Under the concept of approximate monotonicity, Foroutan & Sklansky [361] proposed a modified branch-and-bound algorithm that uses a tolerance parameter to compensate for the deviations from the monotonicity of the feature-selection criterion. Based on this, Siedlecki & Sklansky [237] suggested to compute the expected value of the feature-selection criterion as a function of the iteratively enlarged feature-subsets to avoid early pruning of feature subsets and to continue searching in those parts of the search space that would otherwise be skipped by using the original branch-and-bound algorithm. In both approaches, however, the tolerance threshold is an externally adjustable parameter that is often not known in advance. In contrast to setting a fixed threshold of the tolerance parameter [237, 361], the presented algorithm (hereafter referred to as the tolerance-based branch-and-bound algorithm, TB3) automatically adjust the tolerance parameter in the search by employing resampling techniques (Section 2.5) and statistical hypothesis tests.

In the context of a machine-learning-based feature selection, the branch-and-bound algorithm generalized to non-monotonic feature-selection criteria can be stated as follows (cf., Eq. 3.1): Given a d-dimensional feature set $\vec{X} = \{X_1, \ldots, X_d\}$, a target property Y, a user-defined objective function $\mathscr{G}$, and a machine-learning algorithm $\hat{f} : \vec{X} \mapsto Y$, the TB3-algorithm uses the objective function $\mathscr{G}$ to define a lower bound of $\mathscr{G}$ based on the Student's t-test [369] at a given confidence level α. The

TB3-algorithm first computes the mean of the objective function $\mathcal{G}$ (similar to Siedlecki & Sklansky [237]),

$$\langle \mathcal{G}(Y, \hat{f}(\vec{X})) \rangle := \frac{1}{|\mathcal{R}|} \sum_{R \in \mathcal{R}} \mathcal{G}(Y_{(R)}, \hat{f}(\vec{X}_{(R)})) \,, \tag{4.1}$$

obtained by applying a resampling technique $\mathcal{R}$ to partition the data set into multiple sets of i.i.d. samples, $\{Y_{(R)}, \vec{X}_{(R)}\} \subseteq \{Y, \vec{X}\}$. Based on that, the TB3-algorithm computes the confidence interval of $\mathcal{G}$,

$$\mathcal{G}^{\alpha}_{\pm}(Y, \hat{f}(\vec{X})) := \langle \mathcal{G}(Y, \hat{f}(\vec{X})) \rangle \pm \frac{t_{1-\alpha/2}}{\sqrt{|\mathcal{R}|}} \, \mathrm{sdev}_{\mathcal{G}}(Y, \vec{X}) \,, \tag{4.2}$$

where "t" is the Student's t distribution from the Student's t-test [369], $\alpha \in [0, 1]$ is the confidence level, and "$\mathrm{sdev}_{\mathcal{G}}$" is the sample standard deviation[3] of $\mathcal{G}$,

$$\mathrm{sdev}_{\mathcal{G}}(Y, \vec{X}) = \sqrt{\frac{1}{|\mathcal{R}| - 1} \sum_{R=1}^{|\mathcal{R}|} \Big(\mathcal{G}(Y_{(R)}, \hat{f}(\vec{X}_{(R)})) - \langle \mathcal{G}(Y, \hat{f}(\vec{X})) \rangle \Big)^2} \,. \tag{4.3}$$

The confidence level mainly affects the computational complexity of the algorithm. The higher the confidence level, the more feature subsets are evaluated in the search and the more likely the optimal feature subset is found. Accordingly, the difference between the original branch-and-bound algorithm and the TB3-algorithm presented here is that the TB3-algorithm guarantees that the optimal feature subset is found, but only within a probabilistic tolerance α.

The features related to a property of interest are then determined by searching for the optimal subset of features $\vec{X}^{*} \subseteq \vec{X}$ by maximizing[4] the feature-selection criterion $\mathcal{G}$ with respect to a machine-learning algorithm $\hat{f}$ (cf., Section 3.4),

$$\langle \mathcal{G}(Y, \hat{f}(\vec{X}^{*})) \rangle = \max_{\vec{X}' \subseteq \vec{X}, |\vec{X}'| \leq d} \langle \mathcal{G}(Y, \hat{f}(\vec{X}')) \rangle \quad \Rightarrow \quad \vec{X}^{*} = \operatorname*{arg\,max}_{\vec{X}' \subseteq \vec{X}, |\vec{X}'| \leq d} \mathcal{G}(Y; \hat{f}(\vec{X}')) \,. \tag{4.4}$$

Equation 4.4 can be solved with the branch-and-bound algorithm (Alg. 3.1), where the feature-selection and bounding criterion for the TB3-algorithm are defined as follows:

$$\mathbb{Q}(Y; \vec{X}) \equiv \langle \mathcal{G}(Y, \hat{f}(\vec{X})) \rangle \,, \qquad \overline{\mathbb{Q}}(Y; \vec{X}) \equiv \mathcal{G}^{\alpha}_{-}(Y, \hat{f}(\vec{X})) \,. \tag{4.5}$$

Similar to the suboptimality test of the branch-and-bound algorithm (Eq. 3.39), a tolerance-based suboptimality condition for the TB3-algorithm can be derived. The tolerance-based suboptimality

[3] The confidence interval is estimated based on the (corrected) sample standard deviation, usually denoted as $s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{N-1}}$. Unlike the population standard deviation (population = all possible observations), the sample standard deviation is calculated on a subset of samples of the population, where x_i, $i = 1, \ldots, N$ are the observed values and $\bar{x}$ is the mean of these observations. Its deviation is greater than that of the standard deviation of the population, but converges to this value the more data samples N are used for s.

[4] Most optimization problems in machine learning minimize a cost function (usually some kind of error). Here, the feature-selection criterion is maximized to find the strongest relationship between a set of features and the property of interest. Maximization problems can always be recast into minimization problems and vice versa depending on how the optimization function is defined.

condition under approximate monotonicity combines the feature-selection criterion and bounding criterion into one condition,

$$\mathbb{Q}(Y;\vec{X}') - \overline{\mathbb{Q}}(Y;\vec{X}') \geq \mathbb{Q}(Y;\vec{X}'') - \overline{\mathbb{Q}}(Y;\vec{X}'') \;, \qquad \vec{X}' \subseteq \vec{X}'' \subseteq \vec{X} \;, \tag{4.6}$$

where the respective parent feature subset $\vec{X}''$ is used instead of the current best subset $\vec{X}^{*}$ in the search (cf., Eq. 3.37).

The corresponding algorithm (Alg. 4.1) proceeds by repeated sampling from the training data (Section 2.5), the construction of the machine-learning models, as well as the calculation of the mean $\langle \mathcal{G} \rangle$ (feature-selection criterion) and the lower confidence bound $\mathcal{G}_{-}^{\alpha}$ (bounding function) of the model's prediction performance with samples that were not used for model construction (line 9). There are two tests for the tolerance-based suboptimality condition (Eq. 4.6) to ensure that Eq. 3.39 holds (line 14 and 16). The first test prunes feature subsets whose mean of the objective function ($\mathbb{Q} = \langle \mathcal{G} \rangle$) of the current feature subset $\vec{X}''$ is lower than the lower confidence bound ($\overline{\mathbb{Q}} = \mathcal{G}_{-}^{\alpha}$) of its parent feature subset $\vec{X}'$,

$$\mathbb{Q}(Y;\vec{X}'') < \overline{\mathbb{Q}}(Y;\vec{X}') \;, \qquad \vec{X}' \subseteq \vec{X}'' \subseteq \vec{X} \;, \tag{4.7}$$

i.e., when subsets are statistically different at a confidence level of α as given by the Student's t-test [369]. The second test then prunes feature subsets $\vec{X}''$ whose lower confidence bound either is smaller than $\overline{\mathbb{Q}}$ of its parent feature subset $\vec{X}'$,

$$\overline{\mathbb{Q}}(Y;\vec{X}'') < \overline{\mathbb{Q}}(Y;\vec{X}') \;, \tag{4.8}$$

or its mean objective function ($\mathbb{Q} = \overline{\mathcal{G}}$) is equal to its parent feature subset within a specified convergence threshold ϵ,

$$\left| \mathbb{Q}(Y;\vec{X}'') - \mathbb{Q}(Y;\vec{X}') \right| < \epsilon \;, \qquad \epsilon > 0 \;. \tag{4.9}$$

Both tests ensure that as the number of features increases, the feature-selection criterion converges to its maximum value (i.e., its sample standard deviation becomes smaller) and the feature subsets become more strongly related to the property of interest.

Unlike the original branch-and-bound algorithm (Section 3.4 and Alg. 3.1), there is no upper bound for the objective function $\langle \mathcal{G} \rangle$. Therefore the TB3-algorithm (Eq. 4.4) requires more feature-subset evaluations than the original branch-and-bound algorithm. This has to be taken into account when searching for the relevant features in high-dimensional materials spaces or when many features are only weakly related to the properties of interest. Tests have shown that the TB3-algorithm can in principle be applied to thousands of features (cf., Section 5.2.3). However, an efficient and flexible machine-learning algorithm is needed to quickly sift through the still large space of possible combinations features subsets.

There are effectively two parameters for the TB3-algorithm: the confidence level α and the convergence threshold ϵ. While the confidence level α specifies the probability with which the optimal feature subset should be identified, the convergence threshold ϵ is used to discard feature subsets that contribute only marginally to the prediction performance of the model, thus saving computational resources. Setting the convergence threshold ϵ to a reasonable choice can therefore dramatically

```
Data: Features $\vec{X}$, target $Y$
Input: Resampling technique $\mathscr{R}$, objective function $\mathscr{G}$, machine-learning algorithm $\hat{f}$,
       confidence $\alpha$, threshold $\epsilon$ + parameters of $\mathscr{R}$ as well of $\hat{f}$
Result: Optimal features $\vec{X}^* \subseteq \vec{X}$
function tolerance_based_branch_and_bound($Y$, $\vec{X}$, $\mathscr{R}$, $\mathscr{G}$, $\hat{f}$, $\alpha$, $\epsilon$):
    $\vec{S}_0 = \varnothing$;
    subsets = $\{\vec{S}_0\}$;
    optimal = $\vec{S}_0$;
    while subsets do
        $\vec{S}_{k-1} \in$ subsets;                              // Select subset from list
        subsets = subsets $\setminus \vec{S}_{k-1}$;                  // Remove subset from list
        for $X_i \in \vec{X} \setminus \vec{S}_{k-1}$ do
            $\vec{S}_k = \vec{S}_{k-1} \cup \{X_i\}$;
            // Estimate tolerance parameter for objective function (Eq. 4.2)
            Compute $\mathbb{Q}(Y;\vec{S}_{k-1}) \equiv \langle \mathscr{G}(Y, \hat{f}(\vec{S}_{k-1})) \rangle$ and $\overline{\mathbb{Q}}(Y;\vec{S}_{k-1}) \equiv \mathscr{G}^{\alpha}_{-}(Y, \hat{f}(\vec{S}_{k-1}))$;
            Compute $\mathbb{Q}(Y;\vec{S}_k)$ and $\overline{\mathbb{Q}}(Y;\vec{S}_k)$;
            // Check suboptimality condition (I)
            if $\mathbb{Q}(Y,\vec{S}_k) < \overline{\mathbb{Q}}(Y;\vec{S}_{k-1})$ then
                // Check suboptimality condition (II)
                if $\overline{\mathbb{Q}}(Y;\vec{S}_k) > \overline{\mathbb{Q}}(Y;\vec{S}_{k-1})$ and $|\mathbb{Q}(Y;\vec{S}_k) - \mathbb{Q}(Y;\vec{S}_{k-1})| > \epsilon$ then
                    subsets = subsets $\cup \{\vec{S}_k\}$;
                    // Update optimal feature subset
                    if $\mathbb{Q}(Y;\vec{S}_k) > \mathbb{Q}(Y;$optimal$)$ then
                        optimal = $\vec{S}_k$;
    return optimal;
```

Algorithm 4.1: A pseudo-code listing of the TB3-algorithm presented in this thesis.

lower the computational requirements. In general, the convergence threshold can be set to zero, but a reasonable choice is to set ϵ to the allowed tolerance of the objective function $\langle \mathscr{G} \rangle$, i.e., to terminate the feature-subset search once the mean value of the objective function $\langle \mathscr{G} \rangle$ does not improve by a value of ϵ.

4.1.2 The framework

Based on the feature-subset search within a probabilistic tolerance, the proposed feature-identification framework can be used with any objective function from an information measure or an evaluation metric from the data-science community such as the Pearson's coefficient of determination R^2 [101] or the root-mean-squared error (RMSE). While the objective function of information measures can be applied directly such as for TCMI (Eq. 3.34), a machine-learning model must be generated for every feature subset in the search prior to evaluating a metric on this model. A metric such as the RMSE in the TB3-algorithm would optimize the subset of features to yield the lowest prediction errors for an underlying statistical model, while the Pearson's coefficient of determination R^2 would yield the highest correlation between the set of features and the property of interest. The RMSE therefore focuses on the model's predictions, while the R^2 to identify the features that might be related to the property of interest. The use of Pearson's coefficient of determination R^2 for feature identification is thus particularly suited to relate the features to the property of interest.

The feature-selection and bounding criterion can be determined with an unbiased estimate of the expected model performance [51] via cross-validation [52, 201] or any other resampling scheme (Section 2.5). K-fold cross-validation, for example, partitions the data set of N samples into K approximately equal-sized ($\approx N/K$), non-overlapping subsets $\vec{X}^{(k)}$ of which the k-th partition is used for validating the model performance (as an independent test set from the joint distribution of $\vec{X}$ and Y) and the remaining $(k-1)$-partitions ($\vec{X}^{(\bar{k})} := \bigcup_{i\neq k}\vec{X}^{(k)}$) are used for creating the machine-learning model $\hat{f}_{\bar{k}}$ (cf., Section 2.5). The set of values of the objective function $\mathcal{G}$,

$$\mathcal{R}_{\mathrm{CV}}(\hat{f},\mathcal{G}) = \left\{\mathcal{G}(Y^{(k)},\hat{f}_{\bar{k}}(\vec{X}^{(k)})) \mid k = 1,\dots,K\right\} \tag{4.10}$$

is then averaged over all folds ($k = 1,\dots,K$),

$$\langle\mathcal{G}(Y,\hat{f}(\vec{X}))\rangle = \frac{1}{|\mathcal{R}_{\mathrm{CV}}|}\sum_{R\in\mathcal{R}_{\mathrm{CV}}}\mathcal{G}(Y_{(R)},\hat{f}(\vec{X}_{(R)})) \equiv \frac{1}{K}\sum_{k=1}^{K}\mathcal{G}(Y^{(k)},\hat{f}_{\bar{k}}(\vec{X}^{(k)}))\ . \tag{4.11}$$

The sample standard deviation is then determined from

$$\mathrm{sdev}_{\mathcal{G}}(Y,\vec{X}) = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left(\mathcal{G}(Y^{(k)},\hat{f}_{\bar{k}}(\vec{X}^{(k)})) - \langle\mathcal{G}(Y,\hat{f}(\vec{X}))\rangle\right)^2} \tag{4.12}$$

for each of the tested feature subsets in the search. Due to the probabilistic nature of the suboptimality test (Eq. 4.5), the TB3-algorithm not only returns the optimal non-redundant subset of relevant features $\vec{X}^*$, but also allows to construct a confidence-based aggregation of sub-optimal feature sets into a redundant subset of features $\vec{X}^\circ$ similar to TCMI (Section 3.5),

$$\vec{X}^\circ := \bigcup_{\vec{X}'\subseteq\vec{X}}\left\{\vec{X}' \mid \langle\mathcal{G}(Y,\hat{f}(\vec{X}'))\rangle \geq \mathcal{G}^{\alpha}_{-}(Y,\hat{f}(\vec{X}^*))\right\}\ , \tag{4.13}$$

i.e., into a set of feature subsets whose feature-selection criteria are statistically equivalent at a confidence level of α. The union of all these feature subsets therefore describe the set of which all statistically equivalent feature subsets can be obtained. Equation 4.13 implies that there may actually be no single optimal feature subset in the search (if $\vec{X}^* \neq \vec{X}^\circ$), but rather a set of (non-)optimal feature subsets that can be used equivalently to construct a statistical model. These subsets of features can have completely different features and be completely distinct from each other. A single statistical model can then be created from the union of all these feature subsets, or multiple machine-learning models from the set of statistically equivalent feature subsets, depending on whether a single model or an ensemble of models is needed for an application.

4.1.3 Evaluation studies

To demonstrate the efficiency of the feature identification with the tolerance-based branch-and-bound (TB3) algorithm, the same examples from Section 3.5 are reviewed and analyzed, all based on a common ground truth of known or empirically identified feature subsets. Machine learning models for these evaluation studies are created using the gradient-boosting decision tree (GBDT) algorithm [249–252] (cf., Appendix A.2) as the feature-selection criterion[5]. The analysis is performed by creating a machine-learning model for every feature subset in the search using the Pearson's coefficient of determination R^2 [101] to determine the prediction performance of the generated GBDT model. Because GBDT algorithms [249–252] are known to be extremely sensitive in modeling multivariate relationships [357], the performance of the GBDT model is monitored and the model-creation process terminated as soon as there is no improvement in the objective function of a held-out data set during model construction (Appendix A.2). Results are summarized in Tab. 4.1.

Overall, the TB3-algorithm identifies all relevant and redundant features of the data sets at a confidence level of $\alpha = 0.95$, thereby increasing the prediction performance of the generated machine-learning model which has a greater impact on smaller data sets than on larger data sets. Except for cases with a comparably small amount of data samples or machine-learning models with high prediction errors, the algorithm never identifies features unrelated to the property of interest either in the optimal minimal non-redundant feature subset $\vec{X}^*$ or in the redundant feature subset $\vec{X}^\circ$. Slight deviations in model performances between $\vec{X}^*$ and $\vec{X}^\circ$ were all within the error bounds or the convergence threshold $\epsilon = 0.01$ (i.e., the allowed variance in the Pearson's coefficient of determination R^2 [101] of the cross-validation), showing that adding more features to the feature subsets neither increase the prediction performance of the cross-validated machine-learning models nor the features provide any additional information for the estimation of the properties of interest.

Concerning the identification of optimal feature subsets, it is interesting to note that all possible feature combinations of $X_1, \ldots X_4$ and $X_{11}, \ldots, X_{14}$ of the Friedman regression data set [347] are identified including X_5 and that an optimal feature subset consists of no more than five features (cf., Section 3.5.3). Likewise, in the high-performance concrete data set [350, 351] four to five features are found – cement X_1, water X_2, blast-furnace slag X_6, and curing time X_8 – in different variations and in combination with the other features (cf., Section 3.5.3). Moreover, in both case studies, all of the topmost feature subsets exhibit similar model performances ($R^2 \approx 0.91$) with differences less than the convergence threshold ϵ. Though, the redundant feature subsets become as large as the total number of features of the data set at a confidence level of $\alpha = 0.95$, the optimal minimal non-redundant feature subset $\vec{X}^*$ and the redundant feature set $\vec{X}^\circ$ for 500 data samples are consistent with the findings in Section 3.5. For example, a feature identification with the TB3-algorithm shows that all features are relevant in the high-performance concrete data set with the

[5] A more thorough evaluation of different machine-learning algorithms for feature identification applied to materials-science applications is discussed in the next 5.

Dataset: features (Section 3.5)		**Score (Pearson's R [101])**			
↲ # data samples	relevant features ($\vec{X}^*$, $\vec{X}^\circ$)	$R^2(Y,\vec{X}^*)$	$R^2(Y,\vec{X}^\circ)$	$R^2(Y,\vec{X})$	ΔR^2
Bivariate normal distribution data set: $X_1,...,X_{11}$					[11 features]
Efficiency: Evaluation of 75 feature subsets out of 2,047 feature combinations (4%).					
50	$\mathbf{X_1}, \mathbf{X_2}, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}$	0.06	-0.41	-0.42	+0.48
100	$\mathbf{X_1}, \mathbf{X_2}$	0.90	0.90	0.76	+0.14
200	$\mathbf{X_1}, \mathbf{X_2}$	0.96	0.96	0.92	+0.04
500	$\mathbf{X_1}, \mathbf{X_2}$	0.99	0.99	0.98	+0.01
Friedman regression data set [347]: $X_1,...,X_{14}$					[14 features]
Efficiency: Evaluation of 625 feature subsets out of 16,383 feature combinations (4%).					
50	$\mathbf{X_1}, X_2, X_3, \mathbf{X_4}, X_5, X_6, ..., X_{10}, X_{11}, \mathbf{X_{12}}, \mathbf{X_{13}}, X_{14}$	0.38	0.17	0.17	+0.11
100	$X_1, \mathbf{X_2}, X_3, \mathbf{X_4}, X_5, X_9, \mathbf{X_{11}}, \mathbf{X_{12}}, \mathbf{X_{13}}, X_{14}$	0.67	0.67	0.59	+0.08
200	$\mathbf{X_1}, X_2, \mathbf{X_3}, \mathbf{X_4}, \mathbf{X_5}, X_8, X_9, X_{10}, X_{11}, \mathbf{X_{12}}, X_{13}, X_{14}$	0.79	0.74	0.73	+0.06
500	$\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \mathbf{X_4}, \mathbf{X_5}, X_{11}, X_{12}, X_{13}, X_{14}$	0.92	0.91	0.89	+0.03
High-performance concrete data set [350, 351]: $X_1,...,X_8$					[8 features]
Efficiency: Evaluation of 193 feature subsets out of 255 feature combinations (76%).					
50	$X_1, X_2, \mathbf{X_3}, X_4, \mathbf{X_5}, \mathbf{X_6}, \mathbf{X_8}$	0.66	0.62	0.62	+0.04
100	$\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, X_4, X_5, X_6, \mathbf{X_8}$	0.83	0.81	0.81	+0.02
200	$\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, X_4, X_5, X_6, \mathbf{X_8}$	0.93	0.92	0.92	+0.01
500	$\mathbf{X_1}, X_2, \mathbf{X_3}, \mathbf{X_4}, X_5, \mathbf{X_6}, \mathbf{X_8}$	0.91	0.91	0.91	+0.00

Tab. 4.1. Feature identification with the TB3-algorithm and a gradient-boosting machine-learning algorithm [249–252] (cf., Appendix A.2) at a confidence level of $\alpha = 0.95$ and a convergence threshold of $\epsilon = 0.01$. Shown are the identified redundant feature subsets $\vec{X}^\circ$ for each of the data sets from Section 3.5. The circled features are known to be not relevant and therefore were misidentified by the algorithm (cf., Section 3.5). Features typeset in bold denote the optimal feature subsets $\vec{X}^*$. The table also shows the average prediction performance of the model obtained with repeated 10-fold cross-validation (5 rounds) utilizing either the identified optimal non-redundant set of features $\vec{X}^*$, the redundant set of features $\vec{X}^\circ$, or the full set of features $\vec{X}$ for different number of data samples. The difference in model performance in the last column refers to the comparison with a model from the optimal feature subset $\vec{X}^*$ and from all features $\vec{X}$ of the data set, $\Delta R^2(Y,\vec{X}^* \leftrightarrow \vec{X}) \equiv R^2(Y,\vec{X}^*) - R^2(Y,\vec{X})$. The reported search efficiency refers to a feature identification with 500 data samples.

exception of fly ash (X_7); however, a non-redundant feature set can already be obtained from four out of eight features as found out by the TB3-algorithm and FS-SISSO method.

The efficiency of the TB3-algorithm for 500 data samples (Tab. 4.1) exemplifies that only a small fraction of all possible feature subset combinations need to be evaluated to identify a set of features that is best suited for a machine-learning model. Even in the extreme case in the high-performance concrete data set, where it is expected that all features are relevant, the TB3-algorithm does not have to evaluate all feature subsets due to effective pruning in the search (Alg. 4.1). This even applies to a smaller number of data samples, although the efficiency decreases due to spurious relationships while reducing the number of data samples. For example, with only 50 data samples the ratio of evaluated feature subsets[6] in the bivariate normal distribution is 0.08 (instead of 0.04). For the Friedman regression data set it is 0.11 (instead of 0.04) and in the high-performance concrete data set it is 0.81 (instead of 0.76). Given that almost all features in the high-performance concrete data set are relevant, pruning has only little effect on the exponential time complexity in the search for the optimal feature subset. Therefore, the efficiency of the TB3-algorithm depends strongly on the extent to which the features are related to each other and to the property of interest. In general, the fewer related features are in the data set, the more efficient the TB3-algorithm becomes.

4.1.4 Summary

Based on the feature-subset search within a probabilistic tolerance, the proposed TB3-algorithm combines the efficiency of the feature-subset search from data-mining tools with the capability of statistical modeling of feature-property relationships from machine learning. As the best subset of features is chosen only from the mean objective function $\langle \mathcal{G} \rangle$, the TB3-algorithm returns an optimal subset with a minimum number of features and the lowest variance in the errors of model predictions. Here, the tolerance parameter (Eq. 4.2) is not only used to efficiently prune the search space and discard feature subsets, whose prediction performance already degrades due to multi-collinear features [99, 357]). It also generates error bounds for the relevance of feature subsets and for a redundant set of features in the search (Eq. 4.13). Thus, the advantage of introducing a tolerance parameter in the search is to enable a robust identification of relevant features, that are optimal or close to optimal for the specified machine-learning algorithm, while providing a more complete picture of the statistical relationships present in the data.

The point is that with a feature-identification method like TB3, simpler models with fewer features can be created with the same prediction performance as models constructed on the full set of features. Because models should not be more complex than necessary (according to Occam's razor [370]), simpler models should be preferred. These models can in turn be used to better analyze and visualize the statistical trend in the data and to derive a more physics-based model. Identifying the relevant features prior to statistical modeling would therefore result in statistical models that are more likely to capture the underlying statistical trend in the data than complex models and that are more likely to be applicable to new data (e.g., new materials). Clearly, the additional computational effort for data sets with very few features does not justify the use of the TB3-algorithm (as in the experiments above using the examples from Section 3.5). However, the more features a data set has, the more efficient the TB3-algorithm is (cf., Section 5.2).

The experiments from Section 3.5 showcase that the TB3-algorithm is effective in reducing the dimensionality (Section 3.1) of the data set, while remaining close to the optimal and in some cases globally optimal solution (Tab. 4.1). In particular, the efficiency of the TB3-algorithm suggests that TB3 can in principle be used even when other feature-identification methods are no longer applicable (cf., Section 5.2.3 as one example of the next Chapter). However, experiments also show that the efficiency of the algorithm strongly depends on the number of data samples and the actual relationships between the features and the property of interest. Overall, the risk of identifying unrelated features is higher for smaller data sets than for larger data sets; therefore, a reliable identification

[6] The lower the ratio of evaluated feature subsets, the more efficient the feature-subset search becomes.

of relevant features at a specified confidence level α requires not only the construction of highly predictive machine-learning models from a reduced feature set but also a deeper understanding of feature-property relationships in the data.

4.2 Feature-dependence maps of materials data

Statistical relationships between the features and the property of interest are often analyzed with the Pearson's R [101] or Spearman's ρ [269] coefficient of determination by computing the pairwise-linear correlation of features. A generalization to non-linear and multivariate dependences are feature-dependence maps [100]. Similar to pairwise-linear correlation maps, feature-dependence maps visualize feature interactions as heat maps and require an accuracy measure that quantifies the dependence between a feature subset $\vec{X}'$ and the dependent feature X. The main difference to pairwise-linear correlation maps is how feature interactions are computed. For feature-dependence maps, the accuracy measure acts as a dependence score, which has a value of 1 if a feature X is completely dependent on the other features $\vec{X}'$ and 0 if there is no such dependence. Here, a feature X is said to be dependent on other features $\vec{X}' \subseteq \vec{X} \setminus X$, if a statistical model can be created with $\vec{X}'$ to estimate X. As such, feature interactions are not necessarily symmetric when interchanging $\vec{X}' \leftrightarrow X$ and therefore are sensitive to arbitrary (injective, surjective, and bijective) relationships of non-linear and multivariate dependences in the data.

4.2.1 Examples

Feature-dependence maps can be generated by creating a statistical model $\hat{f} : \vec{X} \setminus X \longrightarrow X$ for every feature $X \in \vec{X}$ in the data set[7]. Hence, feature-dependence maps can be combined with the developed feature-identification framework (Section 4.1.2) to quickly identify dependent features of the data set based on the feature-selection criterion of the feature-subset search. Figure 4.2 shows the feature-dependence maps for the three evaluation studies from Section 3.5, which were computed with the gradient-boosting decision tree (GBDT, Appendix A.2) algorithm as the feature-selection criterion for the TB3-algorithm (cf., Section 4.1.1).

Feature-dependence maps can also be created with any other feature-identification method, including methods designed for specific machine-learning algorithms [100] (Fig. 4.3). For example, the feature-dependence maps of the three feature-identification methods from Section 3.5 – recursive feature elimination [263] with random forest [266] (RFECV), gradient-boosting decision trees [252, 267, 268] with permutation feature importance [266, 338] (FS-GBDT), and the sure-independence screening and sparsifying operator (FS-SISSO) [197, 371] show very similar, though not always correct, dependences between the features of the data set (Fig. 4.3). In the Friedman regression data set, FS-GBDT is close to the ground truth: features $X_1, \ldots, X_4$ are strongly linearly correlated with $X_{11}, \ldots, X_{14}$ and the target Y is a function of $Y = f(X_1, \ldots, X_5)$ or of the respective correlated features $X_{11}, \ldots, X_{14}$. FS-SISSO correctly identifies some dependences, but still finds spurious relationships, possibly due to the relatively small number of data samples, while RFECV suggests

[7] The creation of a statistical model, $\hat{f} : \vec{X} \longrightarrow Y$, for the property of interest Y might also be of interest and is therefore included in these maps.

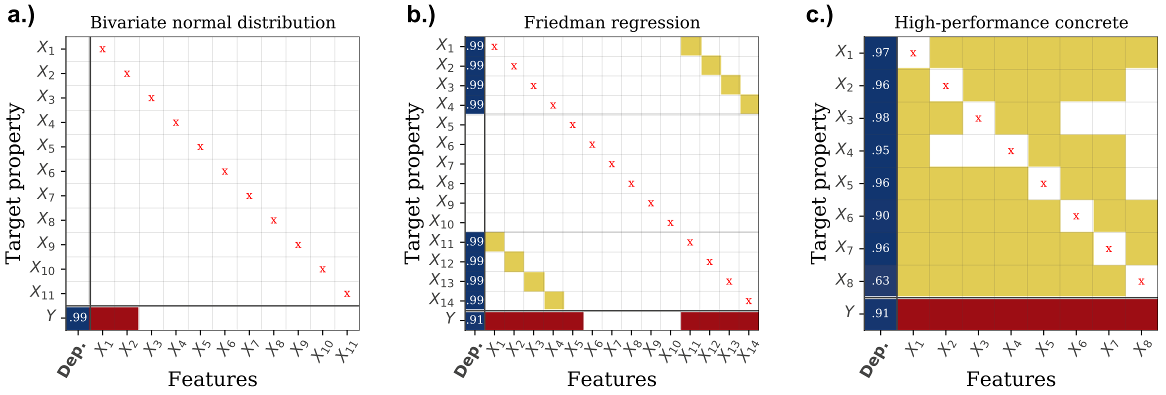


Fig. 4.2. Feature-dependence maps of the three evaluation studies from Section 3.5 using 500 data samples: bivariate normal distribution (a.), Friedman regression (b.), and high-performance concrete (c.). The feature-dependence maps were created with the TB3-algorithm (Section 4.1) using 500 data samples and a gradient-boosting machine-learning algorithm [249–252] (Appendix A.2) at a confidence level of $\alpha = 0.95$ and a convergence threshold of $\epsilon = 0.01$. Shown in the first column (Dep.) is the score of the dependency; this is the Pearson's R coefficient of determination [101] of the 10-fold cross-validated machine-learning model using the identified features. The dependence of the target properties are also shown in the last row.

that features are more dependent on each other than they actually are. This result is to be expected: feature identification based on variable-importance measures of the machine-learning algorithms are sensitive to multi-collinear features [372].

In particular, RFECV and FS-SISSO have the tendency to remove relevant features in the presence of collinear related features [373, 374] (as can be seen in the case of the property of interest Y). As long as at least only one of the multi-collinear features is identified, the redundant and unrelated features have no direct impact on the predictive performance of the statistical models. However, it could be that not all relevant features are identified with either method, so that the statistical models created would lead to significantly larger prediction errors (similar to using information-theoretic methods prior to statistical model). Generating feature-dependence maps and building statistical models therefore require feature-identification methods, such as the TB3-algorithm, that are reliable in relating the features to the property of interest and, in particular, are robust in the presence of multi-collinear features.

4.2.2 Summary

Feature-dependence maps can be seen as a diagnostic tool to identify multivariate inter-correlated features in the data set. Combined with a model-agnostic feature-identification method such as the TB3-algorithm, the identification of related features with different machine-learning algorithms can be compared. For example, by replacing the three feature-identification methods from Section 3.5 with the TB3-algorithm, the same feature-dependence maps can be obtained (Fig. 4.2 as compared to Fig. 4.3). In particular, the TB3-algorithm can be used to identify features that are useful for estimating other properties of interest or are not related to any other feature in a data set and therefore may play an important role in building accurate statistical models for the property of interest. Thus, rather than expending time trying to intuit the relevant features with a feature-identification method specific

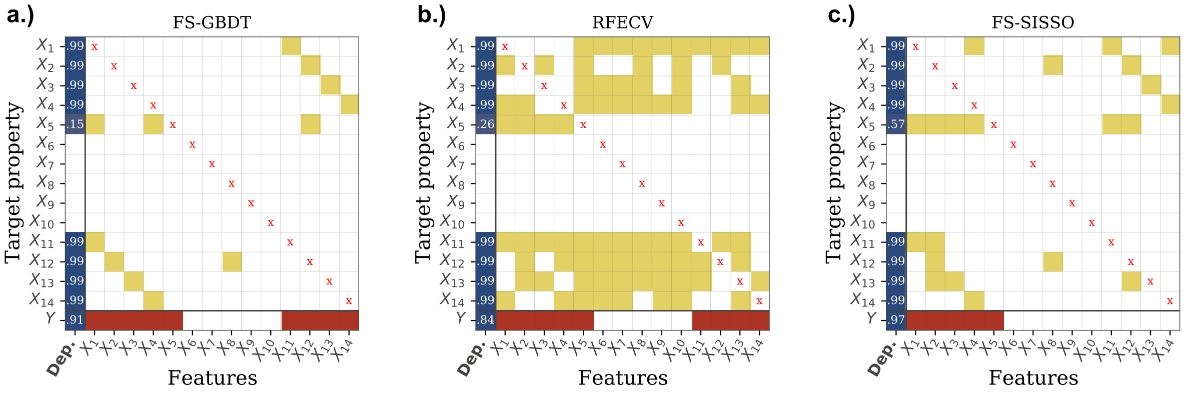


Fig. 4.3. Feature-dependence maps from the Friedman regression data set [347] using 500 data samples obtained with a.) gradient-boosting decision trees [249–252] and permutation feature importance [266, 338] (FS-GBDT), b.) recursive feature elimination [263] and random forests [266] (RFECV), and c.) with the sure-independence screening and sparsifying operator [197, 371] (FS-SISSI) as described in Section 3.5.2. The hyperparameters of all machine-learning models were optimized prior to feature identification (cf., Section 3.5) using 10-fold cross-validation [52, 201] (Section 2.5). The reported metric is the Pearson's R coefficient of determination [101] of the 10-fold cross-validated machine-learning models. Features $X_1, \ldots, X_4$ must be strongly linearly correlated with $X_{11}, \ldots, X_{14}$ respectively and the target Y must be a function of $Y = f(X_1, \ldots, X_5)$ or of their correlated features $X_{11}, \ldots, X_{14}$.

to a machine-learning algorithm, the TB3-algorithm can be utilized as a standardized procedure to automatically create feature-dependence maps and to identify redundant feature sets for generating statistical models and for ensuing data-analysis tasks [249, 375–377], with the model's prediction performance as the only criterion for feature identification (cf., Section 5.2 of the next Chapter).

4.3 Uncertainty estimation of machine-learning models

The prediction performance of a machine-learning model is estimated based on data samples that have not been used for model construction. The performance measure often comes in form of a single evaluation metric such as the root-mean-squared error, the mean absolute error, or the Pearson's coefficient of determination [101] estimated with one of the available resampling techniques (cf., Section 2.5). This produces an average error statistics of the whole model. However, as machine-learning models are inherently statistical and only approximate the property of interest to a limited extent (cf., Section 2.4), the errors between a model's prediction and the actual value of the property of interest can be quite large (cf., Section 5.2). Hence, it can be expected that the model's prediction may vary significantly from the actual value of the property of interest. Predictions of machine-learning models cannot be trusted in particular: when there are not enough data samples available, the exploration and screening of new materials drives the materials search into regions for which there are not enough data samples (cf., [208]), or when the models are applied to new data, e.g., to new materials (cf., [208, 363]). An uncertainty estimate expressed in terms of a prediction interval is therefore desirable and even necessary in many materials-science applications to identify regions of the materials space that are promising to explore [102–104, 363–365] (cf., Section 5.2). Sutton, Boley *et al.* [208]

proposed a method for identifying regions of a materials space in terms of a symbolic description, for which a machine-learning model is applicable. These descriptions can be used to qualitatively identify which materials are likely to have a low prediction error based on the underlying machine-learning model. However, a new material within these regions could still have a machine-learning prediction that is significantly different from the actual value of the property of interest. In addition, the domain of inapplicability, i.e., the region where the machine-learning model is likely to fail, may be more promising in the context of materials discovery and design to search for new materials that exhibit a rare or otherwise uncommon materials behavior. The following rationale is therefore to define an uncertainty estimate for individual machine-learning predictions in order to identify materials which cannot be accurately predicted and as such are promising to explore for the analysis, screening, and prediction of novel materials.

An uncertainty estimate is an interval of the error made in each individual prediction. The uncertainty in the model's predictions can be estimated through the generation of a distribution of predictions with either machine-learning algorithms (logistic regression [51, 52], Gaussian process regression [108, 109, 365], deep neural networks [12, 105–107, 200], random forests [378], etc.), or statistical methods [103, 364, 379–383]. Most of these methods increase the computational requirements of the statistical modeling (e.g., random forest, resampling techniques, and statistical methods), while some of them, namely Gaussian process regression [108, 109] and deep-neural networks (using dropout as a Bayesian approximation [107, 384]) are expensive in the estimation of the prediction error of the property of interest. Furthermore, some methods may need large amount of data (e.g., deep-neural networks), while others make assumptions about the data or the errors (e.g., statistical regression or Gaussian process regression). The problem: For practical data sets, knowledge about the distribution that generated the data is usually not available and therefore needs to be assumed. In the best case, the assumptions are correct, so that an error model provides reliable prediction intervals. However, all these models assume that data samples are i.i.d. generated from the underlying distribution[8]. In practical materials-science applications though, new materials are usually synthesized from variations of promising candidate materials. As such materials data often contain a disproportionate number of samples from specific materials classes or compositions and consequently are rarely i.i.d. [8, 207, 385].

Resampling strategies estimate uncertainties in terms of an ensemble of machine-learning models constructed on distinct partitions of the training data to provide both a mean and a variance for further evaluations of the materials space (Fig. 4.4). They can be applied to any machine-learning algorithm and provide error bounds for uncertainty estimations, where the prediction interval estimates the range in which the property of a new material falls at a given confidence level α (Fig. 4.4). A confidence level of $\alpha = 0.95$, for example, means that the actual values of the property of interest are expected to be within the estimated respective prediction intervals of the constructed machine-learning model in 95% of all the predictions. However, ensemble models may perform poorly on certain materials or in underrepresented materials spaces, i.e., regions of the materials space with only very number of data samples [112, 208]. Moreover, ensemble-based model predictions can be misleading especially if the

[8] In fact, the i.i.d. assumption is a standard assumption in machine learning [51, 52].

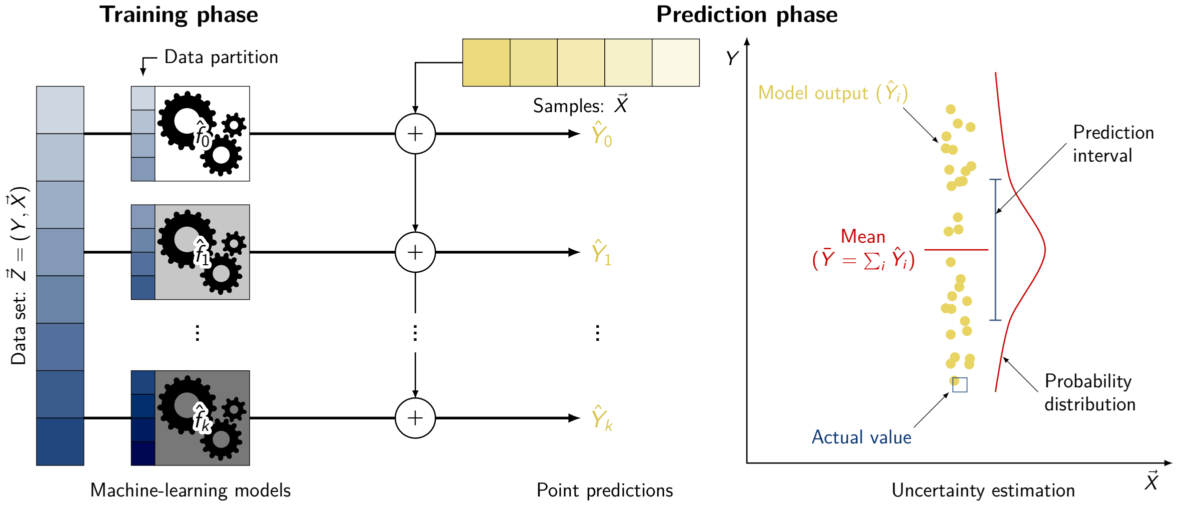


Fig. 4.4. Illustration of an ensemble model to estimate the uncertainty of machine-learning predictions. Shown is the construction and prediction phase of the k machine-learning models $\hat{f}_i$ on different partitions of the data and the prediction on new data $\vec{X}'$. Each machine-learning model of the ensemble provides a point prediction $\hat{Y}_i$ in the distribution of the predicted model outputs. From the distribution of the model outputs, the mean value ($\bar{Y}$), the variance of the prediction, and the prediction interval at a given confidence level can be determined. Depending on the confidence level, the actual value may or may not be included in the prediction interval.

machine-learning predictions are highly correlated due to unfortunate partitioning or heterogeneous data.

The framework described here therefore employs a relatively new method, called conformal prediction [60–64]. Conformal prediction is a model-independent method for estimating the reliability of machine-learning predictions and for estimating prediction intervals for new materials. In contrast to existing methods for uncertainty estimation (e.g., Bayesian inference or logistic regression), conformal prediction requires no additional assumption on the machine-learning model to estimate the error bounds of the model's predictions. It further lessens the i.i.d. requirements for data sampling under the so-called exchangeability condition [60, 62, 63] by assuming that samples need only to be drawn from the same distribution. In addition, it requires minimal computational costs for the estimation of the prediction intervals and provides statistical guarantees (i.e., within a probabilistic tolerance) by relating the size of the prediction intervals to the performance of the machine-learning model, independent of the machine-learning algorithm [62–64].

4.3.1 Conformal prediction

Conformal prediction addresses the weakness of many traditional machine-learning algorithms to provide reliable uncertainty estimates for machine-learning predictions on new data [60–64, 366]. Reliable uncertainty estimates means that, at a confidence level of α, the actual values are within the error bounds of the machine-learning's predictions with a probability P at least $\alpha\%$[9] [386]. Having

[9] $\alpha\% := \alpha \cdot 100\%$

observed a sequence of data[10], where each sample $z_i \in \vec{Z}$ is a pair $(y_i, \vec{x}_i)$, $i = 1, \ldots, n$ of features $\vec{x}_i \in \vec{X}$ and a property of interest $y_i \in Y$,

$$\vec{Z} := (Y, \vec{X}) = \{z_1, \ldots, z_n\} = \{(y_1, \vec{x}_1), \ldots, (y_n, \vec{x}_n)\} \ , \tag{4.14}$$

a conformal predictor is thus an estimator built on top of a traditional machine-learning algorithm, $\hat{f} : \vec{X} \longrightarrow Y$, that outputs a prediction region Γ (a prediction interval in the case of regression or a set of predicted labels in the case of classification) for a new data sample $\vec{x}_{n+1}$ at a given confidence level α. The probability P that the actual value of the property of interest Y is within the prediction region Γ and confidence level α is then given by,

$$P(y_{n+1} \in \Gamma^{1-\alpha}) \geq \alpha \ . \tag{4.15}$$

The α parameter is a tolerance error: the smaller it is, the greater the probability that the actual value of the property of interest is contained within the estimated prediction region Γ.

Exchangeability

A conformal predictor is based on the exchangeability condition which provides statistical guarantees that the probability of error, i.e., $y_{n+1} \notin \Gamma^{1-\alpha}$, does not exceed its significance level $\epsilon = 1 - \alpha$ for any confidence level α and any prediction region Γ. The exchangeability condition [62, 63] states that variables $z_1, \ldots, z_n$ are exchangeable, if for any permutation π of the set $\{1, \ldots, n\}$, the variables $w_i = z_{\pi(i)}$ have the same joint probability distribution as z_i [62–64],

$$\begin{gathered} P(z_1, \ldots, z_n) = P(z_{\pi(1)}, \ldots, z_{\pi(n)}) = P(w_1, \ldots, w_n) \ , \\ \forall \text{ permutations of } \pi = \{1, \ldots, n\} \ . \end{gathered} \tag{4.16}$$

In other words, the samples z_i can be drawn in any order and need not be drawn independently as long as they are drawn from the same distribution [62–64, 387].

By design, a conformal predictor with a confidence level of $\alpha = 0.95$ (significance level $\epsilon = 0.05$) should contain the actual values of the property of interest in 95% of all cases (cf., Eq. 4.15). A conformal predictor that satisfies Equations 4.15 and 4.16 is said to be valid [60–64, 366]. The validity of a conformal predictor can be tested statistically by varying the confidence level between 0 and 1 to ensure that the uncertainty estimates are consistent across multiple confidence levels. This process is called calibration and the resulting curve should therefore correspond to a linear line from 0 to 1. Deviations from this line leads to reported error rates that are higher than expected. However, with diverse data sets of different materials and structures, and a larger number of samples in a data sets, conformal predictors are expected to be always valid [62] (cf., Section 5.2).

Definition

Conformal prediction defines a nonconformity measure A to quantify how different a new sample $\vec{x}_{n+1}$ is with respect to all the other samples in a data set $\vec{Z}$. It then determines the size of the prediction region Γ based on a statistical hypothesis test of the nonconformity measure. The nonconformity

[10] In the literature on conformal prediction, the training set is represented as a multiset $\vec{Z} = \Lbag z_1, \ldots, z_n \Rbag$ – a so-called bag – without any internal order of the elements.

measure is a real-valued function, $A : \vec{Z}\times\vec{Z}' \longrightarrow \mathbb{R}$, created on the data set $\vec{Z}$ and evaluated on a second data set $\vec{Z}'$. The measure assigns a numerical score $\vec{\alpha} = (\alpha_1, \ldots, \alpha_n)$, the so-called nonconformity score, to each sample z_i of $\vec{Z}$,

$$\alpha_i = A(\{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\}, z_i) = \Delta(y_i, D_{\{z_1,\ldots,z_{i-1},z_{i+1},\ldots,z_n\}}(\vec{x}_i)) \ , \tag{4.17}$$

where $\Delta : Y \times \hat{Y} \longrightarrow \mathbb{R}$ refers to a discrepancy measure and $D : \vec{Z} \times X \longrightarrow Y$ to the conformal predictor of the data set. Y is the actual value of the property of interest and $\hat{Y}$ is the estimated value for the property of interest. The discrepancy measure is constructed based on all samples of $\vec{Z}$ except z_i and is applied to $\vec{x}_i$. A nonconformity measure can be any evaluation metric from machine learning. In principle, it can be the average of all samples of $\vec{Z}$ or, in regression, the absolute error, e.g.,

$$\alpha_i = \Delta(y_i, \hat{f}(\vec{x}_i)) = |y_i - \hat{f}(\vec{x}_i)| \ , \quad i = 1, \ldots, n \ . \tag{4.18}$$

The underlying intuition is that new data samples that are less similar to the data samples of a data set should lead to less confident estimates. As the nonconformity measure can be scaled, the numerical value of α_i does not, by itself, quantifies how different a sample z_i is relative to the other samples z_k of a data set $\vec{Z}$. A suitable method for this comparison is to compute the proportion of samples greater than the nonconformity score of z_j for each of sample $z_i = (y_i, \vec{x}_i) \in \vec{Z}$ in the data set $\vec{Z}$,

$$p_j = \frac{|\{i = 1, \ldots, n : \alpha_i \geq \alpha_j\}|}{n} \ . \tag{4.19}$$

Equation 4.19 is called the p-value for z_j and ranges from $1/n$ to 1. The p-value indicates whether z_j is conforming or not: If the p-value is small (close to its lower bound $1/n$), z_j is nonconforming, i.e., z_j is different from all other samples and has a large nonconformity score α_j. If ithe p-value is large (close to its upper bound 1), z_j is similar to the other data samples of $\vec{Z}$ and has a small nonconformity score α_j. Hence, a sample $z_j = (y_j, \vec{x}_j)$ that conforms with $z_i \in \vec{Z}$ at a confidence level α has a p-value greater than $1 - \alpha$. The level of confidence determines the amount of conformity (as measured by the p-value). Generally speaking, lower nonconformity scores α_j are equivalent to high conformity. The higher the conformity, the more likely the machine-learning model can accurately predict the property of interest of that sample.

In case of a new data sample $\vec{x}_k$ whose actual value y_k is not known, a statistical hypothesis test needs to be performed to estimate the value of the property of interest Y of that sample. To this end, each possible value $y \in Y$ is tested as a possible candidate of the sample $z_k = (y, \vec{x}_k)$,

$$p_k^y = \frac{|\{i = 1, \ldots, n : \alpha_i \geq \alpha_k^y\}|}{n} \ , \qquad \alpha_k^y = \Delta(y, \hat{f}(\vec{x}_k)) \ , \qquad \forall y \in Y \ . \tag{4.20}$$

Given that any other value $y \in Y$ than y_k would result in larger nonconformity scores (Eq. 4.17) and lower p-values (Eq. 4.20), only the values that conform to the data set $\vec{Z}$ and therefore are close to y_k have a p-value of $p_k^y > 1 - \alpha$ at a confidence level α. Conversely, this means that samples whose property of interest cannot be predicted well by the conformal predictor have a low p-value. Based on these considerations, it is possible to introduce a (heuristic) measure in Section 4.4 to identify materials that are either not sampled from the same distribution or are promising to synthesize in a real materials-science application.

The corresponding prediction region of the conformal predictor D is generated by the p-values from Equation 4.20. It is the set of all values y, that fulfill the criterion $p_k^y > 1 - \alpha$,

$$\Gamma^{1-\alpha}(\{z_1, \dots, z_{k-1}, z_{k+1}, \dots, z_n\}, \vec{x}_k) = \{y \in Y : p_k^y > 1 - \alpha\} \ . \tag{4.21}$$

Under the assumption that the samples are exchangeable (Eq. 4.16), the probability that y_j is in the region $\Gamma^{1-\alpha}$ at confidence level α is given by [62]

$$P(y_j \in \Gamma^{1-\alpha}(\{z_1, \dots, z_{j-1}, z_{j+1}, \dots, z_n\}, \vec{x}_j)) \geq \alpha \ , \tag{4.22}$$

where the corresponding prediction region is given by

$$y_j \in \left[\min \Gamma^{1-\alpha}(\vec{x}_j), \max \Gamma^{1-\alpha}(\vec{x}_j)\right] \ . \tag{4.23}$$

Conformal prediction regions are approximately constant across all data samples $z_i \in \vec{Z}$ [388]. To generate tighter prediction regions for samples that are more reliable than others, nonconformity measures (Eq. 4.17) can be normalized $\alpha_i \longrightarrow \tilde{\alpha}_i$, for instance with the nearest-neighbor method [62–64]. Normalization can be subject to the “curse of dimensionality” [389, 390]. It is therefore critical to reduce the number of features in a machine-learning model with dimensionality-reduction techniques (Section 3.1) or the TB3-algorithm (Section 4.1) prior to estimating the prediction uncertainty of new samples.

Inductive conformal prediction

Equations 4.20 and 4.21 estimate the prediction interval of the conformal predictor D for a new sample $\vec{x}_{n+1}$ on all previous samples in the data set $\vec{Z}$. This is known as transductive conformal prediction [60]. Transductive conformal prediction requires to update the nonconformity scores α_i and to reconstruct the conformal predictor each time a new sample $z_{n+1} = (y_{n+1}, \vec{x}_{n+1})$ is added to the data set [62, 63]. A more efficient approach, that constructs the conformal predictor only once is inductive conformal prediction [62, 391, 392].

Inductive conformal prediction requires to construct a conformal predictor on a subset of the data, on a so-called proper training set $\vec{Z}_t = (Y_t, \vec{X}_t) = \{z_1, \dots, z_m\}$, and to compute the nonconformity scores,

$$\alpha_j = \Delta(y_j, \hat{f}(\vec{x}_j)) \ , \quad y_j \in \vec{Z}_c \ , \tag{4.24}$$

of the inductive conformal predictor on a separate data set, on a so-called calibration set $\vec{Z}_c = (Y_c, \vec{X}_c) = \{z_{m+1}, \dots, z_n\}$ (Alg. 4.2). The constructed inductive conformal predictor is then applied on a test set $\vec{X} = \{x_{n+1}, \dots, x_{n+l}\}$ of which each value of $y_j \in Y_c$ is tested as a possible hypothesis for the actual value y_{n+l} of the sample $\vec{x}_{n+l} \in \vec{X}$ and checked for (non-)conformity via the p-value as in Equation 4.20,

$$p_{n+l}^y = \frac{|\{j = m+1, \dots, n : \alpha_j \geq \alpha_{n+l}^y\}|}{n - m + 1} \qquad \forall y \in \vec{Y}_c \ . \tag{4.25}$$

```
Data: Data set $\vec{Z} = (Y, \vec{X}) = \{z_1, \ldots, z_n\}$, $z_i = (y_i, \vec{x}_j)$
Input: Test sample $\vec{x}_{n+1}$, machine-learning algorithm $\hat{f}$, confidence level $\alpha$
Result: Prediction region $\Gamma^{1-\alpha}(\vec{x}_{n+1})$
function conformal_prediction($\vec{x}_{n+1}$, $\vec{Z}$, $\hat{f}$, $\alpha$):
    // Setup conformal prediction
    Split data $\vec{Z}$ into training $i = \{1, \ldots, m\}$ and calibration set $j = \{m+1, \ldots, n\}$;
    Construct confidence predictor $\hat{f}$ on training set $\vec{Z}_t = \{z_1, \ldots, z_m\}$;
    Compute nonconformity scores $\alpha_j = \Delta(y_j, \hat{f}(\vec{x}_j))$ on calibration set $\vec{Z}_c = \{z_{m+1}, \ldots, z_n\}$;
    // Compute prediction range $\Gamma^{1-\alpha}(\vec{x}_{n+1}) = \{y_j \in Y : p^{y_j}_{n+1} > 1 - \alpha\}$
    for $y_j \in \vec{Y}_c$ do
        $\alpha^{y_j}_{n+1} = \Delta(y_j, \hat{f}(\vec{x}_{n+1}))$;
        $p^{y_j}_{n+1} = \frac{|\{k = m+1, \ldots, n : \alpha_k \geq \alpha^{y_j}_{n+1}\}|}{n - m + 1}$;
        if $p^{y_j}_{n+1} > 1 - \alpha$ then
            Add $y_i$ to the prediction set $\Gamma^{1-\alpha}(\vec{x}_{n+1})$;
    return $\Gamma^{1-\alpha}(\vec{x}_{n+1})$;
```

Algorithm 4.2: A pseudo-code listing of the conformal prediction algorithm presented in this work.

Considering that the prediction region of the inductive conformal predictor at confidence level α is formed from a statistical hypothesis test, where the most unlikely values y_j are rejected at confidence level α, the prediction region is

$$\Gamma^{1-\alpha}(\{z_1, \ldots, z_n\}, \vec{x}_{n+l}) = \{y \in \vec{Y}_c : p^{y}_{n+l} > 1 - \alpha\} \; , \tag{4.26}$$

and the validity of the prediction region ensures that

$$P(y_{n+1} \in \Gamma^{1-\alpha}(\{z_1, \ldots, z_n\}, \vec{x}_{n+1})) \geq \alpha \; , \tag{4.27}$$

from which a prediction interval can be derived as in Equation 4.23.

4.3.2 The algorithm

The introduction of a calibration set $\vec{Z}_c$ adds some randomness to inductive conformal prediction. Therefore, in order to stabilize the prediction regions, the uncertainty estimation needs be repeated and the data randomly split at each iteration. There are two variants, namely aggregated [393, 394] and cross-conformal prediction [395], which use either bootstrapping or cross-validation to reduce the variance in the estimation of the nonconformity scores and the p-values (Eqs. 4.24 and 4.25) of the prediction intervals. However, both variants have their limitations: aggregated conformal prediction uses about 63% of the data samples [201], while cross-conformal prediction lacks theoretical

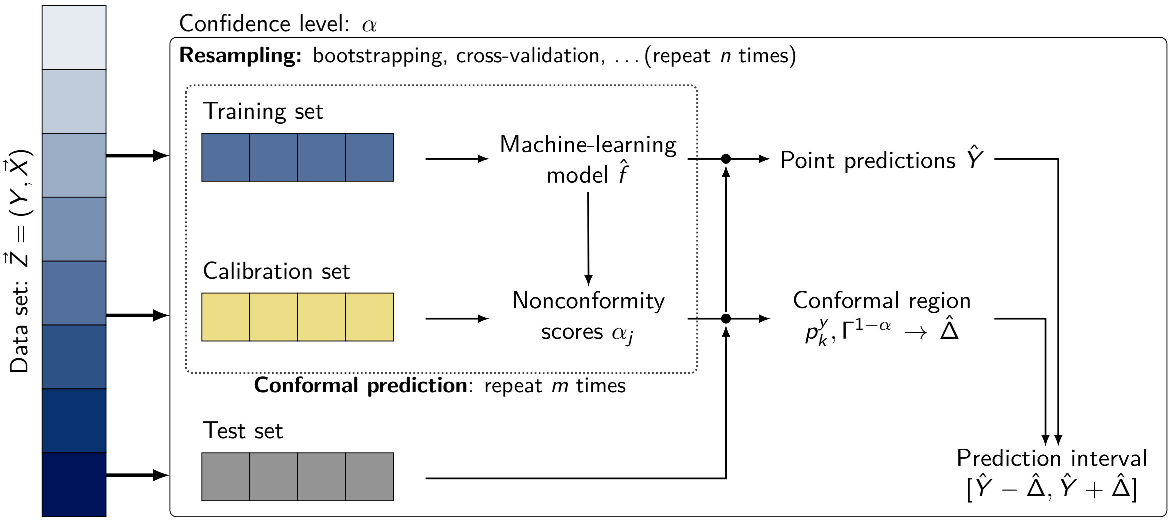

Confidence level: α
Resampling: bootstrapping, cross-validation, . . . (repeat n times)
Data set: $\vec{Z} = (Y, \vec{X})$
Training set
Machine-learning model $\hat{f}$
Point predictions $\hat{Y}$
Calibration set
Nonconformity scores α_j
Conformal region $p_k^y, \Gamma^{1-\alpha} \rightarrow \hat{\Delta}$
Conformal prediction: repeat m times
Test set
Prediction interval $[\hat{Y} - \hat{\Delta}, \hat{Y} + \hat{\Delta}]$

Fig. 4.5. The proposed setup for uncertainty estimation with conformal prediction for new data (out-of-sample predictions) and for samples of the training data (in-sample predictions). For out-of-sample predictions, n is set to 1 and the test set equals the samples of the new data, while the data set is split m times into a training set and a calibration set to create the machine-learning model, to estimate the prediction region via the nonconformity scores, and to determine the prediction interval. For in-sample predictions, the same procedure is repeated n times with different samples of the data set and the prediction interval determined for samples being part of the test set.

guarantees for the prediction regions, i.e., Equation 4.26 cannot be guaranteed [395]. Therefore, an extension of inductive conformal prediction is developed to guarantee reliable uncertainty estimates while stabilizing the prediction intervals by repeatedly constructing the machine-learning model and applying conformal prediction at each iteration, as shown in Figure 4.5.

The setup depicted in Figure 4.5 can be used to estimate the prediction intervals for the samples of the data set (in-sample predictions) or new samples (out-of-sample predictions). In-sample prediction involves randomly partitioning the data set m times into a training, calibration, and a test set, and repeating the procedure n times with different samples in the test set at each iteration. Out-of-sample prediction only involves splitting the data m times into a training set and a calibration set, which are then used to estimate the prediction regions of all new samples of a separate test set.

The setup is similar to in-sample conformal inference [388], which considers only balanced splits of the data set into two halves ($n = 2$) at the expense of less accurate predictive models. In contrast, the proposed setup in this thesis (Fig. 4.5) imposes no restrictions on the splits and stabilizes the prediction regions by randomly splitting the training data multiple times into a proper training set $\vec{Z}_t$ and a calibration set $\vec{Z}_c$ of i.i.d. data samples, while ensuring valid prediction intervals (cf., Eq. 4.26 and [62, 388]). Here, the size of the calibration set is set between 15-30% [396] of the training data, but always greater than or equal to the floor of the inverse of the significance level minus one, $|\vec{Z}_c| \geq \lfloor \frac{1}{1-\alpha} - 1 \rfloor$, to provide sufficient support for the chosen confidence level and to avoid discretization errors in estimating the p-values of the nonconformity scores (Eq. 4.25) [397]. The proper training set is then used to generate the conformal predictor (machine-learning model), while the calibration set is used to calculate the nonconformity scores (cf., Eqs. 4.17 and 4.24). Finally, using both the conformal predictor from the training and the nonconformity scores from the calibration set, the conformal-prediction regions of the remaining data samples in the test set are determined. This procedure is repeated n times to stabilize the prediction regions of the conformal prediction given different splits of the data set. This approach has the advantage that, first, any resampling technique can be used (Section 2.5) and, second, reliable estimates can be obtained for the prediction regions of all samples within the training data and on new samples.

The most time-intensive part is the robust estimation of the prediction regions: the setup depicted in Fig. 4.5 has a time complexity of $\mathcal{O}(m)$ for out-of-sample predictions and $\mathcal{O}(m \cdot n)$ for in-sample predictions. However, to avoid excessively long execution times due to the creation of different conformal predictors, machine-learning models can also be created prior to conformal prediction. The prerequisite for this is that the machine-learning model is created on different partitions of the training data and the model is evaluated for data that have not been used for model construction (as is typically the case, when the prediction performance of these models is evaluated). The predictions are then used for calibrating the model and estimating the prediction intervals of the samples in the test set. Since this approximation requires cross-validating the machine-learning model only once, uncertainty estimates can be determined at minimal additional cost as compared to performance evaluations of the model. Furthermore, the approximation has been shown to provide tighter prediction intervals in the evaluation studies of Section 3.5 than creating the machine-learning model multiple times, while still guaranteeing reliable uncertainty estimates (Tab. 4.2 and Sec. 5.2). Lastly, the approximation enables the application of conformal prediction as a *post hoc* analysis tool to existing machine-learning models, which makes it attractive to provide prediction intervals without having to rebuild the statistical models again.

4.3.3 Examples

To identify regions of the materials space that cannot be adequately predicted by the specified machine-learning models, the data sets in Chapter 3 are analyzed to showcase the uncertainty estimation on simple examples. For all three data sets, the data are randomly split into a training set, consisting 90% of the data and a test set encompassing the remaining 10% of the data, and the procedure is repeated $n = 50$ times. Furthermore, the training set is randomly split $m = 100$ times into a proper training set $\vec{Z}_t$ and a calibration set $\vec{Z}_c$, where the calibration set size is set to 20% of the size of the data used for conformal prediction, or the floor of $\left(\frac{1}{1-\alpha} - 1\right) \times 100\%$ in case of small data sets, whichever is larger. The proper training set is then used to create the underlying machine-learning model, while the calibration set is used to estimate the prediction intervals of the remaining data samples in the test set. As the underlying machine-learning model, the gradient-boosting decision trees algorithm [249–252] (cf., Appendix A.2) is used and as nonconformity score the maximum absolute error $|y - \hat{y}|$ between the actual y and the predicted value $\hat{y}$. The prediction region is then obtained by cross-validating the model and normalizing the prediction intervals using the k-nearest neighbors algorithm [62–64] and the Python package `nonconformomist` [398] for three different levels of confidence: $\alpha = [0.5, 0.8, 0.95]$. In addition, the prediction bands of all three confidence levels are determined as convex hulls of the respective prediction intervals and samples are highlighted whose actual values are outside the prediction interval at a confidence level of $\alpha = 0.95$. Results are shown in (Fig. 4.6).

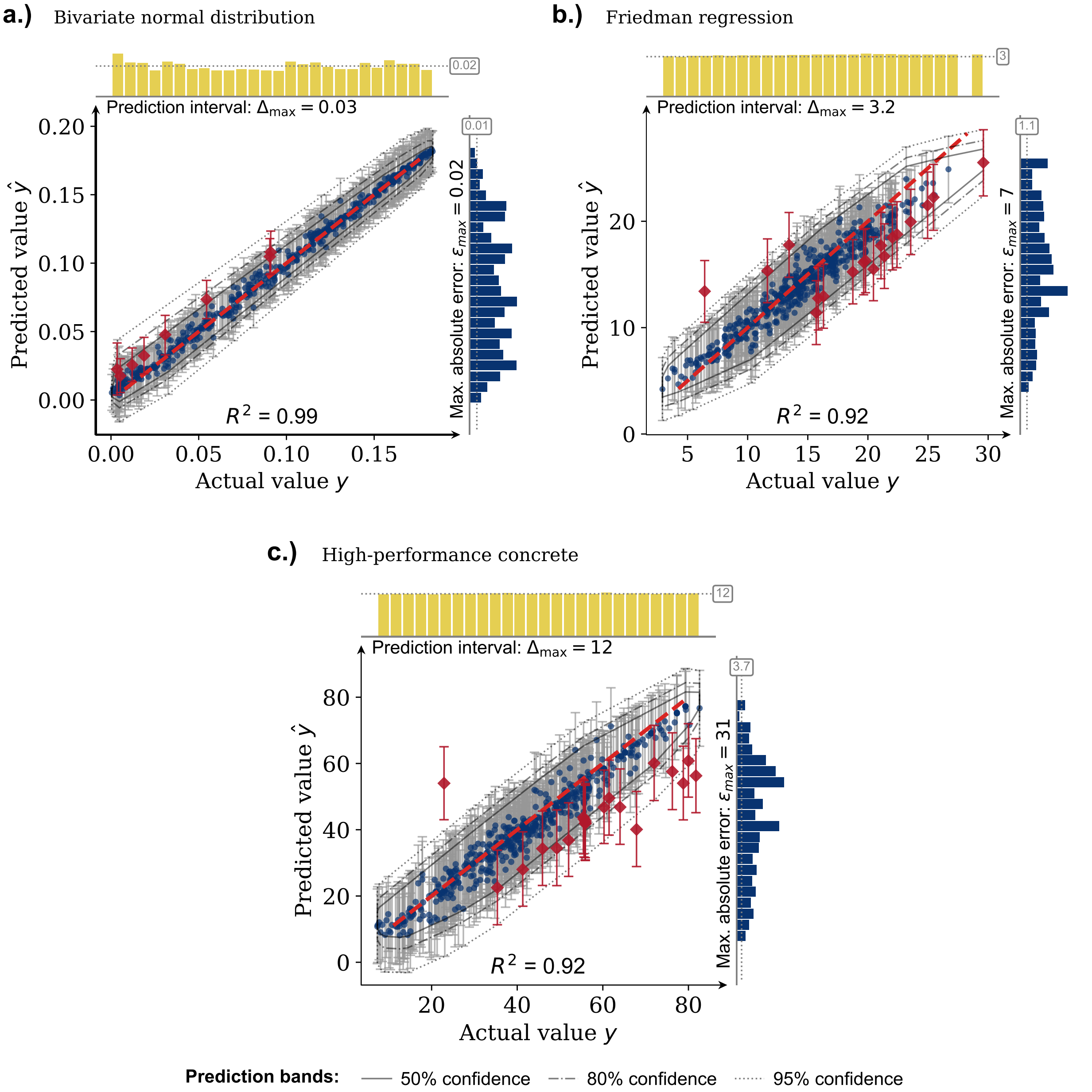


Fig. 4.6. Parity plots of actual values y versus in-sample predictions $\hat{y}$ including prediction bands at confidence levels $\alpha = [0.5, 0.8, 0.95]$ (solid, dashed, dotted) and uncertainty estimations obtained by conformal prediction with a gradient-boosting machine-learning algorithm [249–252] (cf., Appendix A.2) for the three data sets – a.) bivariate Gaussian distribution, b.) Friedman regression data set, and c.) high-performance concrete – as discussed in Chapter 3. Highlighted are all samples whose actual values are outside the prediction interval at a confidence level of $\alpha = 0.95$. The plots above and right of the parity plots show the distribution of the the size of prediction intervals (diagram above the parity plot, Δ) and the maximum absolute error between the predicted $\hat{y}$ and actual value of y (diagram right of the parity plot, $\varepsilon = |y - \hat{y}|$). The numbers in the boxes display the mean values ($\bar{\Delta}$, $\bar{\varepsilon}$), while the maximum errors are given in the texts below or to the left of the diagrams. For comparisons of model performance and prediction intervals, the Pearson's coefficient of determination (R^2) between the actual values y and the predictions $\hat{y}$ is also reported.

In the Friedman and high-performance concrete data sets (Fig. 4.6b. and c.), the prediction intervals are roughly constant. Constant prediction intervals are an indicator that each sample in the data set can be predicted equally well by the machine-learning model. If a prediction interval in a region is larger than in others, it means that the region is underrepresented or the samples cannot be adequately represented by the machine-learning model relative to the other samples. For example, the prediction intervals of the bivariate Gaussian distribution (Fig. 4.6a.) tend to be slightly larger for small values than for larger values. This is to be expected and is due to the addition of noise to the simulated data set, which affects smaller values comparatively more than larger values. The prediction intervals also provide an estimate of the tendency of the errors. In the three examples above, the machine-learning model of the bivariate Gaussian data set slightly overestimates smaller values, while in the Friedman regression and the high-performance concrete data set, larger values are slightly underestimated by the machine-learning models (Fig. 4.6b. and c.).

In all three data sets, the size of the prediction interval $[-\Delta, +\Delta]$ is approximately equal to the maximum absolute error and therefore includes most of the actual values. This is a consistency condition. In fact, conformal prediction includes the actual values within a probabilistic tolerance of at least α% of the training data, emphasizing the validity of the prediction intervals as a whole (Tab. 4.2).

Depending on the confidence level, it can be observed that high confidence levels lead to large prediction intervals. This phenomenon, termed as confidence-efficiency trade-off [62, 64], has the consequence that prediction intervals are fairly large and high confidence levels results in higher uncertainty in the model's predictions [386]. In the context of uncertainty estimation, the confidence-efficiency trade-off provides an important diagnostic tool to assess the reliability of the machine-learning models at various confidence levels. For the case studies as discussed in Chapter 3, the prediction bands of different confidence levels show that even with a relatively low confidence $\alpha = 0.5$, most of the actual values of the samples are covered by the prediction bands of the underlying machine-learning model. From this it can be concluded that the gradient-boosting machine-learning models are capable of efficiently modeling the underlying relationships of the data, thereby providing confidence that the actual values of the predictions are contained with probability at least α on average in the estimated prediction intervals from conformal prediction. However, the sizes of the prediction intervals demonstrate that, despite good model performances, the prediction capabilities are limited. While in the bivariate Gaussian distribution the size of the prediction interval is about 1% of the range of the property of interest, the size of the prediction intervals in the Friedman and high-performance concrete data set is about 10% of the range, indicating a high uncertainty at a confidence level of $\alpha = 0.95$. For the last two data sets it might be worthwhile either to increase the amount of data (through experiments or simulations) or to try out other machine-learning models, depending on the required accuracy of the application.

4.3.4 Summary

To estimate the uncertainty of machine-learning predictions, conformal prediction was employed to compute in-sample prediction intervals of a test set based on splitting the data and using the remaining samples for generating the machine-learning models and the prediction region. The basic idea of using

Data set		**Validity (%) / Prediction interval (Δ)**		
	Confidence level (α):	0.50	0.80	0.95
Bivariate normal distribution data set		66.8% / 0.009	90.6% / 0.016	98.4% / 0.028
Friedman regression data set [347]		57.0% / 1.0	84.6% / 1.9	96.4% / 3.2
High-performance concrete data set [350, 351]		56.4% / 2.4	84.2% / 5.7	96.0% / 12.0

Tab. 4.2. Validity of conformal prediction of the three data sets discussed in Chapter 3 and their respective prediction interval at three different confidence levels $\alpha = [0.5, 0.8, 0.95]$. Validity denotes the probability that a model's estimate is within the prediction interval (Eq. 4.22), while the prediction interval denotes the uncertainty of a single (point) estimate.

conformal prediction is to define an uncertainty estimate for individual machine-learning predictions and to identify materials which cannot be predicted well by the machine-learning model and as such are promising to explore for the analysis, screening, and prediction of novel materials. The validity of a conformal predictor can be tested statistically by varying the confidence level between 0 and 1 to ensure that the uncertainty estimates are consistent across multiple confidence levels. Higher error rates indicate that conformal prediction is either applied to too few samples or that some samples are not drawn from the same distribution and thus exhibit an entirely different materials behavior.

Unique to conformal prediction is that it can be combined with any machine-learning algorithm, requires minimal additional computational costs as compared to the performance assessments of the model, and that no parameterizations of the model are needed to obtain reliable uncertainty estimates at a given confidence level [62, 63, 399]. They are two caveats of conformal prediction. First, as conformal prediction is based on the exchangeability condition (Eq. 4.16), all samples in a data set must come from the same distribution. Second, there is a trade-off between the size of the prediction intervals and the confidence level α: the lower the confidence level is, the greater the probability that the actual value of the property of interest is contained within the estimated prediction interval, but the larger the prediction intervals. As the uncertainty estimate is dependent on the confidence level and the prediction performance of the machine-learning model, this means that confidence levels can be adjusted to filter out materials whose target property is below or above some user-defined thresholds [386] or to analyze which samples cannot be predicted by a machine-learning model (cf., Sections 4.4 and 5.2).

4.4 Identification of anomalous materials

The fact that conformal prediction [60–64] produces valid prediction regions at specified confidence levels (cf., Section 4.3.1) can be exploited to determine how different a material is relative to a set of other materials [400, 401]. This difference can be expressed by the non-conformity score and the p-value of conformal prediction (Eqs. 4.18 and 4.19),

$$p^y_{n+1} = \frac{|\{k = 1, \ldots, n : \alpha_k \geq \alpha^y_{n+1}\}|}{n} < \epsilon \ , \quad \alpha^y_{n+1} = \Delta(y_{n+1}, \hat{f}(\vec{x}_{n+1})) \ . \tag{4.28}$$

Equation 4.28 states that the p-value of any material with features $\vec{x}_{n+1}$ and target property y_{n+1} below the significance level of $\epsilon = 1 - \alpha$ does not conform well to other materials $z_i = (y_i, \vec{x}_i) \in \vec{Z}$ in the data set. Such a material is termed anomalous. An anomalous material is therefore a material with a large nonconformity score and a large prediction error of the machine-learning model. Anomalous materials should not be confused with outliers in statistics. Although both describe data samples that differ significantly from other samples in the data set, outliers can be attributed to variability in the measurement or experimental error whereas anomalous materials not.

Equation 4.28 can be seen as a statistical hypothesis test: with a probability of at least α% materials are classified as anomalous because their actual value is not within the estimated prediction interval of the machine-learning model and with a probability of at most $\epsilon\% = 100\% - \alpha\%$ materials are erroneously identified as anomalous because the exchangeability assumption of the training data is violated. This happens, for example, if the training data are not representative for the modeling of the property of interest, there are only very few other materials with similar features to the anomalous material, or the actual values of the property of interest are generated by a different mechanism (or distribution) than the training data (cf., Section 4.3.1).

Equation 4.28 requires the determination of the actual value of property of interest of each new sample in a data set. Therefore, an anomalous material can only be determined from the available data and cannot be used to screen the materials space for materials whose property of interest cannot be accurately estimated by the underlying machine-learning model. Further, Equation 4.28 considers only the error of the machine-learning prediction $\hat{f}(\vec{x}_{n+1})$ relative to the actual value y_{n+1} and not the distance of the sample $z_{n+1} = (y_{n+1}, \vec{x}_{n+1})$ relative to the other samples $z_i \in \vec{Z}$ in a data set. In particular, this means that an anomalous material is characterized only by the property of interest and not by its properties. Therefore, a heuristic is proposed in the following to classify materials as anomalous, when the actual sample is also anomalous in its features $\vec{x}_{n+1}$ and when the actual value of the property of interest has not yet been determined. This heuristic, called credibility (Alg. 4.3), can help to efficiently identify materials whose property of interest cannot be accurately estimated, or when more samples are needed to reliably predict the property of interest.

4.4.1 The algorithm

The identification of anomalous materials is closely related to outlier [402] and conformal-anomaly detection [400, 401]. Both methods determine whether or not a sample was generated from the same distribution of the training data. While outlier detection utilizes order statistics and determinants [403], machine-learning algorithms such as random forest [404], clustering methods [405, 406], or density-based estimators [407–409], and makes assumptions about the data, conformal-anomaly detection calculates the p-value as in Equation 4.28 by taking the difference between the machine-learning prediction $\hat{f}(\vec{x}_{n+1})$ and the actual value of y_{n+1} of a sample $z_{n+1} = (y_{n+1}, \vec{x}_{n+1})$ relative to the set of other samples in the training data $z_i \in \vec{Z}$. Because the p-value indicates when a material is non-conforming relative to the training data [400, 401], conformal-anomaly detection is thus equivalent to identifying the materials whose actual values of the property of interest cannot be predicted by the machine-learning model.

To identify an anomalous material z_{n+1}, outlier and conformal-anomaly detection require the knowledge of the property of interest y_{n+1}. As the property of interest y_{n+1} is often not known in advance, the proposed heuristic in this thesis estimates the probability that the actual value of a material's property of interest is within the prediction interval as obtained by conformal prediction. The idea is to compute the p-value based on the machine-learning's prediction $\hat{f}(\vec{x}_{n+1})$ relative to the k-nearest neighbor samples with similar feature values by using the median of their properties of interest as an initial estimate (Alg. 4.3).

Without having to determine the actual value y_{n+1} of a new sample $z_{n+1} = (y_{n+1}, \vec{x}_{n+1})$, the heuristic, first, estimates the prediction region of the property of interest $\Gamma^{1-\alpha}(z_{n+1})$ (line 12 of Algorithm 4.3) and, second, checks whether the range $r(\vec{x}_k)$ of the values of the property of interest of nearest-neighbor samples z_k (line 13) overlaps with the prediction region of z_{n+1}, i.e., $r(\vec{x}_k) \cap \Gamma^{1-\alpha}(\vec{x}_{n+1}) \neq \varnothing$. Then, whenever the p-value of z_k is less than $1-\alpha$ or $r(\vec{x}_k)$ does not overlap with $\Gamma^{1-\alpha}(\vec{x}_{n+1})$, z_k is anomalous relative to the data set in terms of its feature values and its property of interest. If indeed $P(r(\vec{x}_k) \in \Gamma^{1-\alpha}(\vec{x}_{n+1})) > \alpha$, then the p-value (line 16–19) is determined according to Equation 4.28 and again tested on nonconformity with Equation 4.28 and the machine-learning prediction $\hat{y}_{n+1}$ (line 23).

The outcome of this heuristic can be turned into a score, hereafter referred to as credibility, by repeating the procedure with random partitions of the data and averaging over all hypothesis tests,

$$\Pi^{1-\alpha}(\vec{x}_{n+1}) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}(p^{y}_{n+1} > 1-\alpha \wedge \Gamma^{1-\alpha}(\vec{x}_{n+1}) \cap r(\vec{x}_k) \neq \varnothing) \,, \tag{4.29}$$

where $\mathbb{1}(A)$ is the indicator function that is one if A is true and zero otherwise. Equation 4.29 indicates how well a material's property of interest can be predicted based on the available set of data and the generated machine-learning model: a high credibility indicates a high probability of finding the actual value within the prediction interval, whereas a low credibility indicates an anomalous material whose property of interest cannot be accurately estimated by the machine-learning model.

4.4.2 Examples

Applying the credibility measure (Eq. 4.29) to all three data sets from Chapter 3, no anomalous data samples (or materials) can be found at a confidence level greater than $\alpha > 0.9$. Although anomalous data samples or materials can be found at (much) lower confidence levels, these data samples or materials cannot be verified as there are either no anomalous samples in the data set (bivariate normal distribution, Friedman regression data set) or it is not known which materials in the high-performance concrete data set are subject to different materials behavior. Therefore, to showcase the heuristic on a simple example, the gradient-boosting decision trees (GBDT) algorithm [249–252] (cf, Appendix A.2) is applied on a linear relationship $f(x) = x + \varepsilon$ with noise ε between a feature x and the property of interest $y \equiv f(x)$. For this purpose, 200 samples were generated $\vec{Z} = \{(x_1, y_1), \ldots, (x_{200}, y_{200})\}$, the data randomly split $m = 50$ times into a proper $\vec{Z}$ and a calibration set $\vec{Z}_c$, and the procedure repeated for every sample $z_k \in \vec{Z}$ in the data set to compute the prediction interval $\Gamma^{1-\alpha}$ and the credibility $\Pi^{1-\alpha}$ as in Alg. 4.3. To identify anomalous materials, a total of ten samples were randomly selected, excluded from the model construction, and y perturbed in the range of $[-y, y]$ to varying degrees. The corresponding results of the machine-learning predictions in relation to the actual values of the data set at a confidence level $\alpha\% = 90\%$ are shown in Figure 4.7.

```
Data: Data set $\vec{Z} = (Y, \vec{X}) = \{z_1, \ldots, z_n\}$, $z_i = (y_i, \vec{x}_j)$
Input: Machine-learning algorithm $\hat{f}$, nearest-neighbor estimator $\hat{g}$, confidence level $\alpha$
       + parameters of $\hat{f}$ as well as of $\hat{g}$
Result: true if material $z_k$ is anomalous, false otherwise.
function identify_anomalous_materials($z_k$, $\hat{f}$, $\hat{g}$, $\alpha$):
    credibility = 0;
    anomalous = false;
    for $i = 1, \ldots, N$ do
        Split data $\vec{Z}^{-k} = \{z_j | j \neq k\}$ randomly into training $\vec{Z}_t$ and calibration set $\vec{Z}_c$;
        // Conformal prediction
        Construct confidence predictor $\hat{f}$ on training set $\vec{Z}_t$;
        Construct nearest-neighbor estimator $\hat{g}$ on calibration set $\vec{Z}_c$;
        Compute nonconformity scores $\alpha_j = \Delta(y_j, \hat{f}(\vec{x}_j))$ on calibration set $\vec{Z}_c$;
        // Estimate prediction interval and nearest-neighbor samples
        $\vec{Z}_{\mathrm{NN}} \longleftarrow$ k-nearest neighbors obtained from $\hat{g}(\vec{x}_k)$;
        $r_\Gamma = \left[\min \Gamma^{1-\alpha}(\vec{x}_k), \max \Gamma^{1-\alpha}(\vec{x}_k)\right]$;
        $r_k = \left[\min_y \vec{Z}_{\mathrm{NN}}, \max_y \vec{Z}_{\mathrm{NN}}\right]$;
        // conformal-anomaly detection
        $p_k = 0$;
        if $r_k \cap r_\Gamma \neq \varnothing$ then
            $\hat{y}_k = \operatorname{median} \Gamma^{1-\alpha}(\vec{x}_k)$;
            $\alpha_k = \Delta(\hat{y}_k, \hat{f}(\vec{x}_k))$;
            $p_k = |\{j = 1, \ldots, |\vec{Z}_c| : \alpha_j \geq \alpha_k\}| / |\vec{Z}_c|$;
        // Check hypothesis of an anomalous material
        if $p_k > 1 - \alpha$ then
            credibility $\longleftarrow$ credibility + 1;
    if credibility$/N < 1 - \alpha$ then
        anomalous = true;
    return anomalous;
```

Algorithm 4.3: A pseudo-code listing of the presented algorithm for identifying anomalous materials in materials-science applications.

Overall, the GBDT algorithm accurately models the linear trend of the functional relationship $f(x)$ (as expected) with a Pearson's coefficient of determination of $R^2 = 0.96$ and a root-mean-squared error of RMSE $= 0.01$. As depicted in Fig. 4.7, the samples with the perturbed values are randomly distributed in the target-property domain with eight samples within and two samples outside the prediction band of the training data. The credibility $\Pi^{1-\alpha}$ supports the results of the visual inspection: all samples outside the prediction bands have a low credibility, while the samples

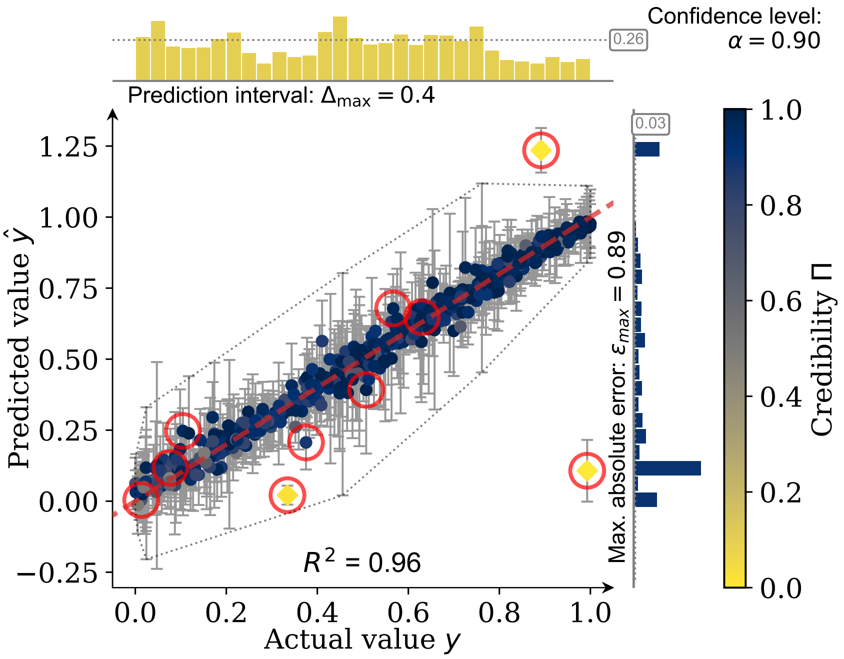


Fig. 4.7. Identification of anomalous materials illustrated by the example of a linear dependence $f(x) = x + \varepsilon$ between the feature x and the target property of interest $y \equiv f(x)$. Shown is the prediction performance of a gradient-boosting decision trees algorithm [249–252] (Appendix A.2) on 200 samples with added noise ε, prediction intervals (error bars), the prediction band (dotted line), and the credibility Π of each sample at confidence level of $\alpha = 0.90$. The plots above and right of the parity plots show the distribution of the size of prediction intervals (diagram above the parity plot, Δ) and the maximum absolute error between the predicted $\hat{y}$ and actual value of y (diagram right of the parity plot, $\varepsilon = |y - \hat{y}|$). The numbers in the boxes display the mean values ($\bar{\Delta}$, $\bar{\varepsilon}$), while the maximum errors are given in the texts below or to the left of the diagrams. To identify anomalous materials, a total of ten samples were randomly selected and y perturbed in the range of $[-y, y]$ to varying degrees (circles). As visualized, samples outside the prediction band, not enclosed in the prediction intervals of the other samples, or that cannot be estimated from samples with similar features are shown to have low credibility scores. Of the ten samples with perturbed values (circles), this is true for three samples that cannot be estimated with the machine-learning model at the specified confidence level of $\alpha = 0.90$ (cf., Section 4.3.1).

within the prediction bands (with the exception of one sample) and the training data have a high credibility. The two samples outside and the one sample inside the prediction band, have been verified to be anomalous ($\Pi < 1 - \alpha$): both samples were strongly under- and overestimated, respectively, by the machine-learning model and have actual values and features that are anomalous relative to the training data.

Although the calculation of credibility on this simplified example allows only very few conclusions to be drawn for material-science applications, it is to be noted that none of the anomalous materials have large prediction intervals as compared to the rest of the samples. This is because the machine-learning model was not constructed on any of the anomalous materials, thereby significantly underestimating the size of the prediction interval. Therefore, the size of the prediction interval cannot and should not be used as an indicator for materials whose property of interest cannot be accurately estimated by the underlying statistical model. The example shows: to identify anomalous materials the features and the property of interest of a new material should be set in relation to the other materials in the data set. Therefore, in large and complex material data sets, the concept of credibility has the potential to help explore the materials space and to identify materials whose property of interest cannot be predicted by the machine-learning model (cf., Section 5.2).

4.4.3 Summary

Credibility (Eq. 4.29) identifies materials whose property of interest cannot be well predicted by the machine-learning model, either because the materials are not representative for the modeling of the property of interest or there are not enough data to reliably predict the property of interest. These materials are termed anomalous and are characterized by a low credibility that can be calculated solely on the basis of a training data set and the features of a new sample (Alg. 4.3). Anomalous materials result in biased machine-learning models that underestimate or overestimate the property of interest of a material (Fig. 4.6). The aim is to understand why machine-learning models may not accurately model the property of interest of these materials in order to identify possible limitations of the machine-learning models.

In the context of materials, anomalous materials may have a rare or otherwise uncommon target property or behave differently than similar materials or structures. In a simple, yet illustrative case study, all anomalous materials were successfully identified. Given that a single case study has limited validity in a field with a diverse set of materials data, the proposed heuristics should be seen as a diagnostic tool to understand and to identify potential problems with the generated machine-learning models, especially when the actual values of new samples are not known. Nonetheless, the introduction of a heuristic, that can identify materials whose properties of interest are difficult to estimate, opens up new opportunities to isolate specifically which materials require further investigations or a more thorough analysis of the constructed machine-learning models.

4.5 Discussion

The identification of relevant feature subsets is a new concept for automatizing feature selection and model construction. It combines filter, wrapper, and embedded methods in a common framework for materials science (Section 4.1). The framework comprises not only a feature-selection workflow to reduce and optimize the feature representation of materials, but also the creation of statistical models that perform well on new data. As the identification of features related to the property of interest requires highly predictive models [358], special focus was put on the applicability of the framework to be independent of any specific feature-selection criterion (cf., Section 4.2).

The strength of the proposed feature-identification framework with the TB3-algorithm is to effectively reduce the number of features in a data set (Section 3.1), to identify the optimal minimal non-redundant feature subset, and, without additional computational cost, to provide an aggregated set of redundant features (Eq. 4.13). The main disadvantage is that the computing time of the TB3-algorithm may increase exponentially with more feature interactions. Tests have shown that the TB3-algorithm can in principle be applied to thousands of features (cf., Section 5.2.3). However, an efficient and flexible objective function is needed to quickly sift through the still large space of possible combinations of relevant features subsets. Whereas filter methods may potentially miss relevant features in the construction of predictive models, wrapper and embedded methods tightly couple the identification of feature subsets with the prediction performance of the machine-learning model. Experiments demonstrate that the found (redundant) feature subsets using wrapper methods do not degrade the prediction performance of the constructed machine-learning models. Furthermore, the identified feature subsets are not only optimal for a machine-learning algorithm (Tab. 4.1), but are also close or identical to the ground truth of the evaluated data sets (Chapter 3). Unrelated features have been identified in none of the examples by the TB3-algorithm in combination with highly predictive machine-learning models. In the case of small data sets and machine-learning models with only moderate predictive performance, however, unrelated features were identified as relevant leading to spurious relationships in the characterization of the property of interest. Therefore, the chance of identifying unrelated features using machine learning is higher for smaller data sets than for larger data sets, and is higher the lower the model's prediction performance.

Critical factors in identifying the features related to the property of interest are multi-collinear features and strong relationships between the features and the property of interest. A deeper understanding of these feature relationships was achieved with the TB3-algorithm and feature-dependence maps of materials data (Section 4.2). Feature-dependence maps [100] visualize feature interactions as heat maps and can be created with any feature-identification method, including various variable-importance measures designed for specific machine-learning algorithms. Examples indicate that a single-best model as provided by a single machine-learning algorithm, constructed from the features of the variable-importance measure or from a minimal non-redundant feature subset, may not adequately represent the relationships in the data. This has the consequence that relationships between features may remain undetected or may even be erroneously identified (Section 4.2). The framework therefore utilizes the TB3-algorithm as a standardized procedure for automatically generating feature-dependence maps from the identified redundant feature subsets. The generation of feature-dependence maps with the TB3-algorithm has the advantage that the algorithm is not only model-agnostic but also takes into account the multiplicity of feature subsets with the same predictive performance (Eq. 4.13). The results are feature-dependence maps within a probabilistic tolerance, which can be used as a diagnostic tool to identify related features in the data set or to further analyze the statistical relationships in the data with *post hoc* analysis techniques [249, 375–377].

The reliability of feature-dependence maps is closely related to the prediction performance of machine-learning models. The prediction performance of a machine-learning model is commonly quantified by a single evaluation metric on data that have not been used for model construction. As has been shown, a single metric, while useful for estimating the goodness-of-fit of the model, neither provides information about the quality of model outputs [363] nor an error estimation in regions with only few training data. Consequently, an uncertainty estimate is desirable and even necessary in many materials-science applications (Section 4.3). Although it is possible to estimate the uncertainty of the model outputs by making assumptions about the model, the distribution of the data or the errors, or through the generation of a distribution of predictions – for example with resampling techniques – it cannot be guaranteed that the predictions are reliable, e.g., that in 95% of the cases the actual value of the property of interest is within the uncertainty estimation at a confidence level of $\alpha = 0.95$ [363, 366]. Conformal prediction [62–64] is a general methodology that is applicable to any machine-learning algorithm and provides guaranteed uncertainty estimates within a probabilistic tolerance. The most time-intensive part is the robust estimation of the prediction intervals. Therefore, an approximation has been proposed that estimates the prediction intervals based on the machine-learning predictions of the data that have not been used for model construction. The uncertainty estimates can be determined at minimal additional cost as compared to the regeneration of the statistical models on different partitions of the data set, while providing tighter prediction intervals and guaranteeing reliable uncertainty estimates at a specified confidence level. In general, the size of the prediction interval depends on the confidence level: higher confidence levels lead to larger prediction intervals. However, the confidence level can be adapted for specific needs in order to obtain precise uncertainty estimates for the prediction of new materials [386].

Uncertainty estimates from conformal prediction hold on average: In the worst case the actual values are still outside the prediction intervals, but only for no more than $\epsilon\% = 100\% - \alpha\%$ of the samples. A heuristic was proposed that identifies those materials whose properties of interest are outside the prediction interval or which are difficult to predict by the machine-learning model. This heuristic can be determined solely on the basis of the training data, the features of the material (Alg. 4.3), and without knowledge of the actual value of the property of interest for that material. In a proof-of-concept study with known ground truth, the introduced heuristic credibility measure successfully identified all materials exhibiting a rare or otherwise unusual property of interest relative to the available data used for modeling the feature-property relationship. As results on such a simple example do not allow to generalize the concept of identifying anomalous materials to materials data with unknown ground truth, the proposed heuristic should be considered primarily as a diagnostic tool to understand and identify potential difficulties with the generated machine-learning models when applied to new data.

Chapter 5

Computational materials-science applications

The identification of relevant feature subsets from a large set of candidate features is central to the creation of accurate machine-learning models for a given class of materials (cf., Chapter 4). The success of such a feature identification depends on clean and curated materials data, knowledge of hypothetical relevant features of the underlying physical processes (cf., Section 3.1), and the machine-learning algorithm to identify the features that are related to the properties of interest (cf., Section 4.1).

So far, recent studies have identified relevant features either heuristically by means of constructing feature subsets iteratively [55, 161], selecting them on the basis of predefined criteria [171, 172, 177, 179], or introducing them without extensively and systematically analyzing their relevance [162, 163, 165–168, 335]. In contrast, the developed feature-identification framework (Chapters 3 and 4) has the potential to automatize the model creation and to enable a data-driven identification and characterization of related features to the property of interest. Therefore, three materials-science applications with increasing number of data samples and features are analyzed to discuss and demonstrate the challenges of identifying relevant features in materials science and to investigate the applicability of the developed feature-identification framework for the quantitative prediction of the crystal structure of octet-binary compound semiconductors [55, 161], the prediction of structural properties of perovskites [410, 411], and the prediction of elastic properties of inorganic crystalline compounds [166, 412].

Applying the developed framework on available materials data has several advantages. First, all three data sets are characterized by clean and curated materials data. Second, prior knowledge of relevant features of the target properties enables a direct comparison of the developed framework with established machine-learning approaches. And third, all three applications are challenging in terms of feature identification, which includes the analysis of a small data set, the determination of feature dependences, and the systematic generation of predictive machine-learning models.

5.1 Conceptual framework

The automatic identification of relevant features and the systematic generation of machine-learning models comprises several steps, in the following, formalized as a model-agnostic protocol for feature identification and model construction (Fig. 5.1). One of the key components in the data-driven identification and characterization of relevant feature subsets is the tolerance-based branch-and-bound algorithm (TB3, Section 4.1). The TB3-algorithm (Alg. 4.1) identifies a ranked list of relevant feature subsets to optimize the materials representation for a given machine-learning algorithm. As such, identified feature subsets are model dependent. First, because of determining the relevant feature subsets by using a machine-learning model as the feature-selection criterion in the feature-subset search (Section 3.1.2 and Section 3.5). And second, because of inherent assumptions in the machine-learning algorithm that influence the exploration of relevant feature subsets.

With regard to a model-independent understanding of the materials-science problem, the question arises whether relevant feature subsets can be identified with a different machine-learning algorithm for feature identification and model creation other than the use of filter methods[1]. The possibility to identify relevant feature subsets independent of the machine-learning algorithm is hereafter referred to as transferability. For simplicity, transferability is determined on the basis of the Jaccard similarity [413] or equivalently the Tanimoto[2] coefficient [414],

$$J(\vec{X},\vec{X}') = \frac{|\vec{X}\cap\vec{X}'|}{|\vec{X}\cup\vec{X}'|} = \frac{|\vec{X}\cap\vec{X}'|}{|\vec{X}|+|\vec{X}'|-|\vec{X}\cap\vec{X}'|} \,, \qquad 0 \leq J(\vec{X},\vec{X}') \leq 1 \,, \tag{5.1}$$

between the identified non-redundant sub-optimal feature subsets $\vec{X},\vec{X}'$ of two different feature-identification methods. The Jaccard similarity (coefficient) ranges from a model-independent feature identification ($J(\vec{X},\vec{X}') \approx 1$) to a model-dependent identification ($J(\vec{X},\vec{X}') \approx 0$), e.g., if feature subsets are not representative for the actual statistical relationship or implicit assumptions about the machine-learning algorithm prohibit the identification of relevant feature subsets.

To this end, feature-property relationships are analyzed in a three-stage approach by first identifying the relevant features, second by generating machine-learning models based on these features, and finally by evaluating the model's prediction performance with a machine-learning model employing fixed (hyper-)parameter settings to enable model inter-comparisons (Fig. 5.1). In particular, feature-property relationships are examined by utilizing diagnostic tools such as feature-dependence maps on the data set (Section 4.2). A comparison of different feature-identification methods is performed to identify common and most frequently identified features. Even though these methods may, in principle, operate differently (e.g., information-based feature selection versus feature-identification methods using machine-learning algorithms such as SISSO or GBDT[3]), common identified features are a strong indication of the relevance of features used to evaluate the relevance of each feature (sub-

[1] This assumes that machine-learning algorithms are able to model arbitrary relationships in the data.

[2] The Jaccard and Tanimoto coefficients are identical in that they are defined by the ratio of the intersection and union. It was first proposed by Jaccard [413] and was later formulated independently again by Tanimoto [414].

[3] Symbolic regression models such as SISSO depend on the number and combination of operators (cf., Appendix A.1). Ensemble of decision trees such as GBDT rely on the number and depth of the generated trees (cf., Appendix A.2). As such, identified features are implicitly dependent on the parameters (and hyper-parameters) of the machine-learning model.

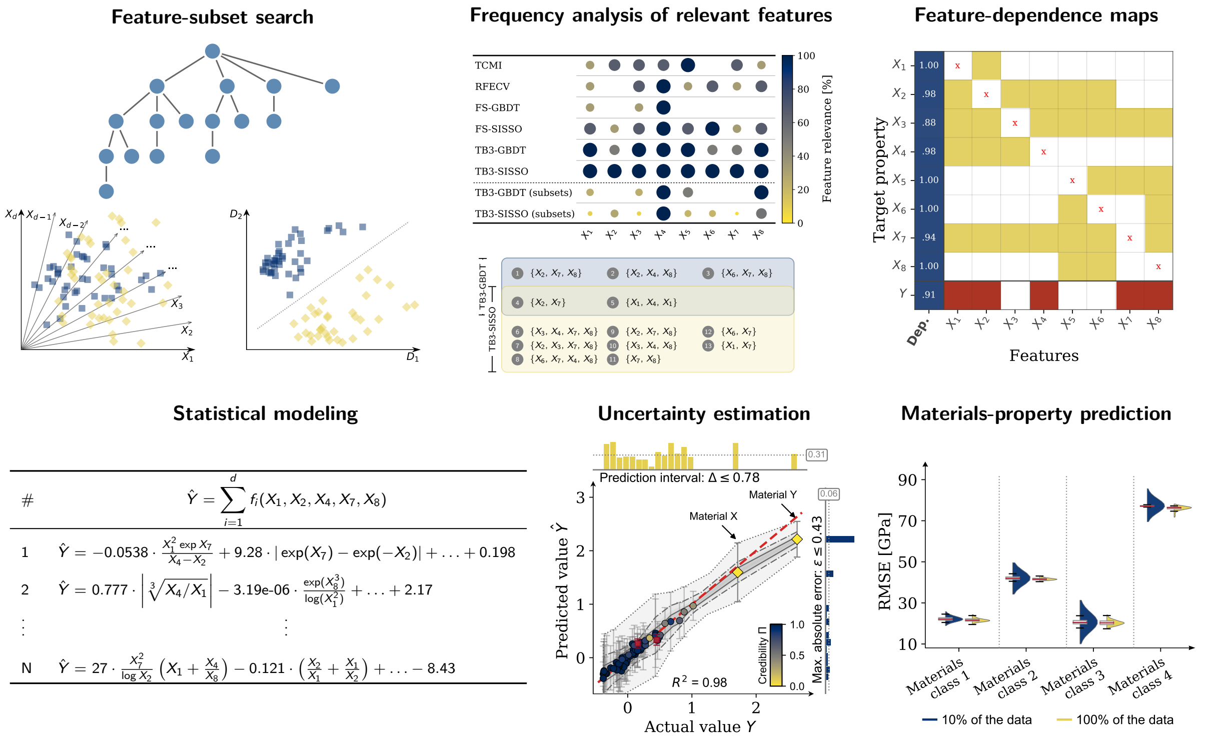


Fig. 5.1. Conceptual framework of the proposed feature identification and model construction ranging from the feature-subset search of relevant features, an analysis of the most frequently identified features, the identification of feature dependencies, the construction of statistical models, the estimating of the prediction uncertainty, materials-property predictions, and the application of the generated machine-learning models to new data.

set) and the transferability of each machine-learning model. Further, ensemble-based and conformal prediction (Section 4.3.1) are employed to estimate the prediction errors of the machine-learning models and to identify materials, whose properties of interest are difficult to estimate (Section 4.4).

5.1.1 Feature identification

The three-stage approach has two major advantages. First, feature-property relationships can be analyzed independently from feature identification and, second, a more flexible machine-learning model can be used for feature identification than for model construction. For example, a piecewise-constant machine-learning algorithm such as the GBDT algorithm [249–252, 267, 268] (cf., Appendix A.2) generally makes fewer assumptions on the statistical relationships than symbolic-regression algorithms (cf., Appendix A.1); unlike symbolic regression algorithms, they do not imply an algebraic functional relationship between the features and the property of interest and hence can identify a larger number of statistical relationships in the data. Moreover, piecewise-constant machine-learning algorithms are generally faster than symbolic-regression algorithms in the modeling of statistical relationships, hence saving expensive computational resources when searching for the relevant features in the data set. This is especially useful when the feature-identification is more computationally demanding than building a machine-learning model based on all the features of the data set. The performance of the GBDT and SISSO algorithm for feature identification is therefore investigated in detail on the octet-binary compound semiconductors [55], perovskites [411], and inorganic crystalline compounds data set [166, 412], using each algorithm (referred to as TB3-GBDT and TB3-SISSO, respectively) to search for the (sub-)optimal and redundant sets of relevant features prior to analyzing feature-property relationships, transferability, and relevant feature subsets of the data set.

For both feature-identification methods, TB3-GBDT and TB3-SISSO, the Pearson's determination coefficient R^2 [101] of the generated machine-learning model is maximized to identify the relevant feature subsets of the data set (thereby minimizing the root-mean-squared error of the generated machine-learning models). Mean and standard deviation are estimated by means of 10-fold cross-validation on the remaining fold not being used for feature identification.

For TB3-SISSO, a symbolic-regression model is built with at most three terms (*desc_dim* ≤ 3[4]) by applying the operator set $\Upsilon = \left\{\cdot + \cdot, \cdot - \cdot, \cdot \times \cdot, \cdot / \cdot, |\cdot - \cdot|, \cdot^{-1}, \cdot^{2}, \cdot^{3}, \cdot^{1/3}, \exp(-\cdot), \exp \cdot, \ln \cdot, \sqrt{\cdot}\right\}$ recursively up to three times (*rung* ≤ 3[4]), while keeping only those terms with a maximum of *maxcomplexity* $= 5$[4] features and *subs_sis* $= 300$[4] feature combinations in each iteration. (cf., Appendix A).

For TB3-GBDT, the LightGBM [252, 267, 268] algorithm is used – a modification of the GBDT algorithm to improve the efficiency and scalability of the machine-learning model. GBDT models are constructed iteratively with 90% of the data (*subsample* $= 0.9$[5], *subsample_freq* $= 1$[5]) to subsequently reduce the overall prediction error of the GBDT model, performing a maximum of 2000 iterations (*n_estimators*[5]) with a shrinkage factor (*learning_rate*[5]) of $\gamma = 0.1$. In each iteration, decision trees are built based on a splitting criterion to find the feature $X \in \vec{X} = \{X_1, \ldots, X_d\}$ that best separates the samples of the data (cf., Appendix A). Decision trees are tree-like models consisting of nodes (features) and leaves (predictions). A minimum of 1% of the data in each leaf node (*min_child_samples*[5]) was used in the applications for predicting the property of interest. In addition, the squared residuals and the model complexity (*eval_metric* = [l1, l2_root][5]) of the LightGBM machine-learning models were monitored to terminate the model optimization, once either the ℓ_1- or ℓ_2-norm [52, 249] has not improved in the last 200 iterations (*early_stopping_rounds*[5]).

5.1.2 Prediction performance of identified feature subsets

The evaluation of optimal and redundant feature subsets is based on the prediction performance of the generated machine-learning model, i.e., on the accuracy to estimate the property of interest on new data. In principle, any machine-learning algorithm is applicable, but among the machine-learning algorithms available to date, flexible and universally applicable algorithms are recommended for creating predictive machine-learning models. Since the creation of a physics-based model and the modeling of feature-property relationships in the data are substantially facilitated by an accurate and deterministic machine-learning model, the SISSO algorithm [176, 197] is used as the reference machine-learning algorithm for estimating the prediction performance of identified feature subsets with the TB3-algorithm (TB3-GBDT and TB3-SISSO, Section 4.1), with total cumulative mutual in-

[4] The names refer to the settings defined in the Fortran code of the reference [197].

[5] All names in the parentheses refer to the parameters in the documentation of the LightGBM package (https://lightgbm.readthedocs.io/) [252, 267, 268].

formation (TCMI, Section 3.3.3), and with the three feature-identification methods discussed in Section 3.5: recursive feature elimination [263] using random forest (RFECV) [266], gradient-boosting decision trees using permutation feature importance (FS-GBDT) [252, 266, 338], and feature selection through hyper-parameter optimization ($\textit{rung} = [1, 2, 3]$[4], $\textit{maxcomplexity} = [3, 5, 10]$[4]) of a 10-fold cross-validated SISSO model (cf., Section 2.5, FS-SISSO).

In the following, the prediction performance for each of the identified feature subsets is estimated by partitioning the data set into 10 folds (10-fold cross validation), using 9 folds to generate a SISSO model and the remaining fold to estimate the mean accuracy (Pearson's coefficient of determination R^2 [101]) and the root-mean-squared error (RMSE) using the generated model. In practice, the dimension of the SISSO model and the recursive application of the operator set are hyper-optimized (Section 2.5). However, in order to compare and benchmark different feature-identification methods and subsets in the following, these hyper-parameters are fixed. For the model construction the operator set $\Upsilon = \left\{\cdot + \cdot, \cdot - \cdot, \cdot \times \cdot, \cdot / \cdot, |\cdot - \cdot|, \cdot^{-1}, \cdot^{2}, \cdot^{3}, \cdot^{1/3}, \exp(-\cdot), \exp\cdot, \ln\cdot, \sqrt{\cdot}\right\}$ is applied, where the number of terms in the symbolic expression is set to *desc_dim* $= 3$[4], the number of operator combinations is set to $\textit{maxcomplexity} = 10$[4], the number of feature combinations is limited to $\textit{rung} = 3$[4], and *subs_sis* $= 100$[6] combinations were evaluated at each SIS step [197].

In addition to the estimation of the prediction performance, a frequency analysis of the identified feature subsets is performed. The same identified features subsets are also used to estimate the minimum number of features (the smallest feature-subset cardinality) required to construct a machine-learning model without degrading its prediction performance. More specifically, the minimum number of features is heuristically determined with the help of feature-dependence maps (Section 4.2) and identified feature subsets with the smallest feature-subset cardinality, whose predictive performance are similar or better than that of a machine-learning model constructed with the full set of features.

5.1.3 Feature-dependence maps

Feature-dependence maps facilitate the identification of multivariate non-linear related features and the most promising property of interest for a given machine-learning algorithm (Section 4.2). In particular, they can be used to analyze the interrelationships between features (independent of a particular property of interest) and to visualize statistical relationships of potentially related features in the data (cf., Section 5.2). The generation of feature-dependence maps can be time-consuming, especially when applied to large data sets with hundreds or even thousands of features. However, not all dependences between the features must necessarily be determined. For instance, if compositional or general-purpose features (Sec. 2.5) are used, the relationships between most of the features are known or can be computed once with TB3-algorithm.

In the framework, feature-dependence maps are generated with the TB3-algorithm using either GBDT or SISSO as the feature-selection criterion (Section 4.1). Essentially, a feature-subset search is performed with a convergence threshold of $\epsilon = 0.01$ (cf., Section 4.2) for each feature of the data set at a confidence level of $\alpha = 0.95$ (Section 4.1). Within this probabilistic tolerance level, the set

[6] The number of expressions used for generating the models mostly determines the computational requirements. The large the number, the higher the computational payload and the longer it needs to build the models.

of identified features are then visualized as blocks in the feature-dependence maps, which represent the interactions between features and between features and the property of interest.

5.1.4 Model construction

Due to the limited availability of materials data and the implicit assumptions in the statistical modeling of relationships, a feature identification possibly identifies many feature-subset combinations in the search (cf., Chapter 4). It is therefore expected that a search for the optimal feature subset leads to a multiplicity of competing machine-learning models with distinct relevant feature subsets, but similar prediction performances (Section 4.1). Because the TB3-algorithm generates a ranked list of minimally non-redundant feature subsets without additional computational cost, competing machine-learning models of identified minimally non-redundant feature subsets can be tested statistically for similar prediction performance.

The most commonly used method to compare the prediction performance of two machine-learning models is the 5x2-cv paired t-test [415]. The paired 5x2-cv t-test splits the data set into two sets and repeats the splitting five times, using both sets alternately to construct the model and evaluate the prediction performance (5-times repeated 2-fold cross-validation, Section 2.5). The variance of the differences is then compared by performing a hypothesis tests (using the Student's t-test [369]) and whenever the feature subsets of the tested machine-learning model have a statistical similar prediction performance as the feature subsets of the best-performing machine-learning model at a confidence level of $\alpha \leq 0.95$, both models and feature subsets are considered statistically equivalent.

5.1.5 Uncertainty estimation

The multiplicity of competing machine-learning models can be considered as an ensemble of machine-learning models with different feature subsets, but similar prediction performances. These machine-learning models in turn can be constructed on the same partitions of the data set and evaluated on data that have not been used for model construction – to provide both a mean and a variance for each (point) prediction and predictions interval with statistical guarantees (Section 4.3). The developed framework for feature identification and model construction facilitates the combination of competing machine-learning models into an ensemble of machine-learning models so as to improve the overall performance of the machine-learning algorithm and to account for the various statistical relationships in the data [416].

The ensemble mean and prediction intervals can either be used to direct the search for new materials [58, 103, 160, 367, 368] or to estimate the prediction error of the ensemble machine-learning models [360, 363, 378]. In the framework, the ensemble average of the machine-learning predictions and the prediction intervals are determined as the αth-percentile ($\alpha = [0.5, 0.8, 0.95]$) of the different model predictions on the same set of data (Section 4.3). Specifically, the materials data that are not being used for the model construction are used to estimate the in-sample prediction intervals [60–64] (Section 4.3.2) by randomly splitting the materials data $m = 100$ times into a proper training set and a calibration set, while repeating the procedure $n = 50$ times for each material in the data set (cf., Fig. 4.5). As will be shown explicitly, the prediction errors based on the ensemble mean and prediction intervals do not provide any statistical guarantees about the uncertainty of the model's predictions (Section 4.3). Hence, prediction intervals are computed using conformal prediction [62–64] at three different confidence levels $\alpha = [0.5, 0.8, 0.95]$ with the setup as described in Section 4.3.1.

5.1.6 Anomalous materials

While conformal prediction provides statistical guarantees about the prediction intervals, it does not provide any information about the prediction error of machine-learning predictions on new data (referred to as reliability). Therefore, the developed feature-identification and model-construction framework uses the credibility measure defined in Section 4.4 to identify materials, whose properties of interest are difficult to estimate (Eq. 4.29). Based on the materials data, credibility is computed based on the Algorithm 4.3 in the same step as the prediction intervals, e.g., by randomly splitting the materials data $m = 100$ times into a proper training set and a calibration set, while repeating the procedure $n = 50$ times for each material in the data set at confidence levels of $\alpha = [0.5, 0.8, 0.95]$.

The lower the credibility, the more difficult it becomes to estimate the materials' properties of interest. Below a specified threshold ($\Pi \leq 1 - \alpha$) these materials are classified as anomalous, meaning they are either not representative for the modeling of the target properties or they cannot be described by the specified set of materials data at a confidence level of α. The aim is to understand why machine-learning models may not accurately model the property of interest of these materials in order to identify possible limitations of the machine-learning models (cf., Section 4.3).

5.2 Datasets

The developed framework for feature identification and model construction aims to automate the model creation in materials-science applications and to enable a data-driven identification and characterization of related features to the property of interest using machine learning. In the following, three materials-science applications are reviewed to study the applicability of the developed framework for feature identification and model construction. Common to all materials-science applications is the availability of clean and curated materials data and the representation of the materials problem by properties of the constituent elements excluding atomic interactions (Section 2.5). However, it is emphasized that the developed framework is neither limited to the choice of materials representation nor to the investigated materials-science applications.

5.2.1 Octet-binary compound semiconductors

Octet-binary compound semiconductors are AB-type materials formed by groups of I/VII, II/VI, III/V, and IV/IV elements of two atomic species. They are known to crystallize in a wide range of crystal structures. Since 1970, a plethora of methods have been developed to qualitatively predict the stability of octet-binary compound semiconductors and the preferred crystal structure based on their energy differences [55, 161, 197, 418–426]. To date, octet-binary compound semiconductors are among the most thoroughly characterized materials classes in materials science and are therefore ideally suited to benchmark and test the developed feature-identification and model-construction framework. In the

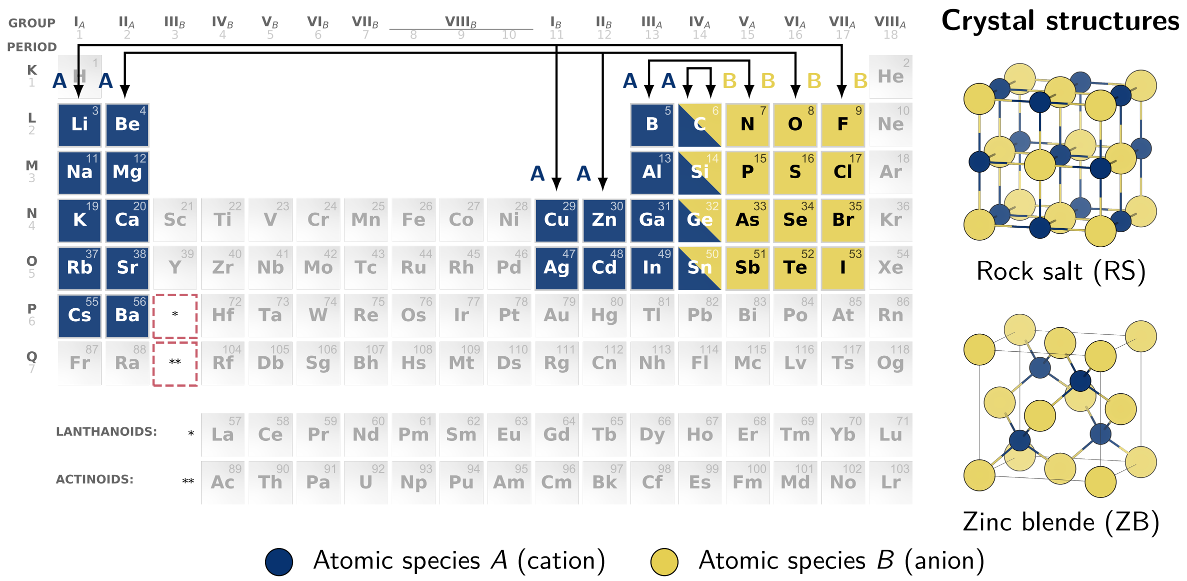


Fig. 5.2. A periodic table of elements of the 82 octet-binary compound semiconductors data set [55, 161] and the two analyzed crystal structures: rock salt and zinc blende. Highlighted are elements that occur in the crystal structures as cations (A) or anions (B). Pairs with a full valence shell are connected by arrows.

following, the quantitative estimation of the energy differences between fourfold (e.g., zinc blende) and sixfold (e.g., rock salt) crystal structures (cf., Fig. 5.2) are investigated in greater detail.

The data set

Predicting the crystal structure and stability of octet-binary compound semiconductors on the basis of their chemical composition has been a challenge for materials science for more than half a century [417, 427]. One of the reasons for this is that the energy differences from related crystal structures of octet-binary compound semiconductors are less than 0.001% of the total energy of a single atom and therefore require very accurate predictions [55, 426].

Recently, Ghiringhelli, Vybiral *et al.* [55, 161] investigated a total of 82 octet-binary compound semiconductors to quantitatively predict the energy difference $\Delta E = E_{\mathrm{RS}} - E_{\mathrm{ZB}}$ of rock-salt and zinc-blende crystal structures using machine learning. They calculated the total energy of rock-salt and zinc-blende crystal structures within the framework of the density-functional theory [25, 26] and the local-density approximation exchange-correlation functional [428], but stressed that the approach is neither limited to the type of exchange-correlation functional nor to the crystal structure and machine-learning algorithm (cf., Section 2.2).

In their work, Ghiringhelli, Vybiral *et al.* focused on creating predictive models that were faster to compute and easier to obtain than the target property ΔE itself [55]. To this end, they iteratively constructed a candidate list of features from seven electro-chemical atomic properties of each atomic species A/B such as the atomic ionization potential IP, electron affinity EA[7], the energies of the

[7] The ionization potential (IP) and electron affinity (EA) were calculated from the free, isolated, spinless and spherically symmetric atom, via the half occupied Kohn-Sham orbital of the half positive (negative) charged atom.

highest-occupied and lowest-unoccupied Kohn-Sham levels, H and L, and the maximum of the radial probability densities of the valence s-, p-, and d-orbitals, r_s, r_p, and r_d, respectively.

To uniquely determine the materials representation, the features of the constituent elements A/B in the features were ordered such that the element A had the smallest Mulliken electronegativity $\mathrm{EN(A)} < \mathrm{EN(B)}$, where $\mathrm{EN} = -(\mathrm{IP} + \mathrm{EA})/2$. Because the ordering of the elements resulted in machine-learning models that were not symmetric with respect to the exchange of atomic species A and B, this thesis additionally investigates the influence of the Mulliken electronegativity as an additional feature in estimating the energy difference ΔE of octet-binary compound semiconductors.

The main difficulty of using this data set for the prediction of the energy difference between rock-salt and zinc-blende crystal structures is the small number 82 materials relative to the moderate number of 16 features. In addition, the data set contains only very few materials with high energy differences, such as boron nitride ($\Delta E_{\mathrm{BN}} = 1.71\,\mathrm{eV}$) and diamond ($\Delta E_{\mathrm{CC}} = 2.64\,\mathrm{eV}$) as compared to the 95% of the data which have energy differences of less than $\Delta E \leq 0.80\,\mathrm{eV}$. As a consequence, the energy difference ΔE has the form of a highly skewed non-Gaussian distribution, leading to a potential bias in identifying relevant features and constructing machine-learning models. Due to the small number of materials structures, the analysis is further complicated by issues of the "curse of dimensionality" [389, 390] (Section 2.6) such as spurious relationships between features (Fig. 5.5), high sensitivity to multi-collinear features (Tab. 5.1), and large variations in the predictive performance of machine-learning models (Tabs. 5.2 and 5.3).

Feature identification

Tests were performed on three randomly chosen subsets of 20, 41, and 82 octet-binary compound semiconductors to identify the relevant features for the energy difference ΔE between rock-salt and zinc-blende crystal structures. Feature-identification methods from Chapter 3 (TCMI, RFECV, FS-GBDT, and FS-SISSO) and the developed feature-identification framework from Chapter 4 (TB3-GBDT and TB3-SISSO) were used to identify the related features to the predict the energy difference between rock-salt and zinc-blende crystal structures. Results are reported in Table 5.1 and machine-learning models were finally built with the SISSO algorithm using fixed hyper-parameter settings [197] (cf., Sections 5.1.2 and A.1).

Overall, there is a clear trend of increasing prediction performance with increasing number of data samples and larger feature-subset cardinalities. With the exception of FS-GBDT (which is affected by issues with the permutation feature importance [266, 338] and spurious relationships between features), all investigated feature-identification methods (TCMI, RFECV, FS-SISSO, TB3-GBDT, TB3-SISSO) achieve similar prediction performance on the complete data set with a Pearson's coefficient of determination of about $R^2 \approx 0.95$ and a root-mean-squared error of about $\mathrm{RMSE} \approx 80\,\mathrm{meV}$. Remarkably, feature-identification methods not only show a high variability in the relevance of individual features, but identified features also vary considerably across different subsets of data samples. Both of these results indicate that statistical relationships between the energy difference ΔE and the identified feature (subsets) are highly non-linear and therefore are computationally difficult to identify. Although octet-binary compound semiconductors can in principle be uniquely determined by two atomic properties[8], the investigated feature-identification methods identify almost all features of both

[8] By definition, octet-binary compound semiconductors are uniquely determined by the atomic charges Z(A) and Z(B) [418, 419], neither of which is included in the data set.

#	Dependence Measure	Features #	Relevant features	Performance R^2	RMSE [meV]
20 samples*	TCMI	8	H(A), L(A), r$_d$(A), EA(B), EN(B), L(B), r$_p$(B), r$_s$(B)	1.00±0.00	452±380
	RFECV	2	r$_p$(A), r$_s$(A)	1.00±0.00	174±88
	FS-GBDT	2	r$_d$(A), r$_p$(A)	1.00±0.00	168±88
	FS-SISSO	7	EN(A), r$_p$(A), r$_s$(A), EA(B), EN(B), H(B), IP(B)	1.00±0.00	148±84
	TB3-GBDT		–	–/–	–/–
	TB3-SISSO		–	–/–	–/–
41 samples	TCMI	7	EA(A), EN(A), L(A), r$_p$(A), EA(B), IP(B), r$_s$(B)	0.84±0.15	86±34
	RFECV	7	EA(A), L(A), r$_d$(A), r$_p$(A), r$_s$(A), H(B), r$_p$(B)	0.76±0.25	134±88
	FS-GBDT	3	EA(A), r$_d$(A), r$_p$(A)	0.81±0.25	111±41
	FS-SISSO	9	EA(A), IP(A), L(A), r$_p$(A), r$_s$(A), H(B), L(B), r$_p$(B), r$_s$(B)	0.80±0.18	111±61
	TB3-GBDT	16	EA(A), **EN(A)**, **H(A)**, **IP(A)**, L(A), r$_d$(A), **r$_p$(A)**, **r$_s$(A)**, EA(B), **EN(B)**, H(B), IP(B), L(B), r$_d$(B), r$_p$(B), r$_s$(B)	0.85±0.16/ **0.88±0.11**	123±106/ **84±49**
	TB3-SISSO	16	EA(A), EN(A), H(A), IP(A), **L(A)**, r$_d$(A), **r$_p$(A)**, r$_s$(A), **EA(B)**, EN(B), H(B), IP(B), L(B), r$_d$(B), **r$_p$(B)**, r$_s$(B)	0.85±0.16/ **0.91±0.13**	123±106/ **87±67**
82 samples	TCMI	8	EN(A), H(A), IP(A), r$_p$(A), EA(B), EN(B), L(B), r$_d$(B)	0.94±0.05	97±44
	RFECV	12	EN(A), IP(A), L(A), r$_d$(A), r$_p$(A), r$_s$(A), EA(B), EN(B), H(B), L(B), r$_d$(B), r$_p$(B)	0.95±0.06	63±22
	FS-GBDT	6	EN(A), L(A), r$_p$(A), r$_s$(A), EN(B), r$_d$(B)	0.86±0.16	424±627
	FS-SISSO	12	EA(A), EN(A), H(A), IP(A), L(A), r$_p$(A), r$_s$(A), EA(B), EN(B), H(B), IP(B), r$_p$(B)	0.95±0.05	82±73
	TB3-GBDT	12	EA(A), EN(A), IP(A), L(A), r$_d$(A), r$_p$(A), **r$_s$(A)**, **EA(B)**, EN(B), IP(B), **r$_p$(B)**, r$_s$(B)	0.94±0.05/ **0.92±0.10**	72±32/ **97±41**
	TB3-SISSO	16	EA(A), **EN(A)**, **H(A)**, **IP(A)**, L(A), r$_d$(A), **r$_p$(A)**, **r$_s$(A)**, EA(B), EN(B), H(B), **IP(B)**, L(B), r$_d$(B), r$_p$(B), r$_s$(B)	0.94±0.06/ **0.91±0.13**	92±69/ **92±62**
Reference [197]		12	EA(A), EN(A), H(A), IP(A), L(A), r$_p$(A), r$_s$(A), EA(B), EN(B), H(B), IP(B), r$_p$(B)	0.95±0.05	81±72
Energy difference		Stats: $\Delta E = [-0.38\,, 2.64]$ eV	mean = 0.11 eV		std = 0.44 eV

Tab. 5.1. Prediction performance of identified redundant feature subsets (normal face) and optimal non-redundant feature subsets (bold face) of different feature-identification methods for estimating the energy difference ΔE between rock-salt and zinc-blende crystal structures: total cumulative mutual information (TCMI, Section 3.3.3), recursive feature elimination [263] using random forest (RFECV) [266], gradient-boosting decision trees using permutation feature importance (FS-GBDT) [252, 266, 338], SISSO using 10-fold cross-validation and hyper-parameter optimization (FS-SISSO) [197] (cf., Section 2.5), and the tolerance-based branch-and-bound algorithm (TB3) with GBDT and SISSO as feature-selection criterion (Section 4.1). Prediction performances were estimated using SISSO by means of 10-fold cross-validation (cf., Section 5.1). Reported are the mean $\bar{x}$ and standard deviation Δx as $\bar{x} \pm \Delta x$ of the Pearson's coefficient of determination $x = R^2$ [101] and the root-mean-squared error (x = RMSE). The RMSE is in units of millielectronvolts (meV). The reported prediction performance from Ref. [197] is also shown. * In the case of 20 samples, the R^2 performance statistic could not be reliably determined due to the small size of test samples (cf., [429]).

atomic species as relevant. This discrepancy is further reflected in the feature-subset cardinalities of the redundant and minimal non-redundant feature subsets of the TB3-algorithm using the GBDT or the SISSO algorithm as the feature-selection criterion (Section 4.1): while the redundant feature subsets include almost all features of the data set, the minimal non-redundant feature subsets require 3–4 features in order to predict the energy difference between rock-salt and zinc-blende crystal structures as accurately as a machine-learning model created with all features (Tab. 5.1). In contrast, a feature identification with the GBDT algorithm (FS-GBDT) identifies about 2–6, RFECV about 2–12, TCMI about 2–7 and FS-SISSO about 7–12 features as relevant. Whereas FS-GBDT and RFECV tend to identify features of atomic species *A* only, they were outperformed by TCMI and FS-SISSO. In terms of accuracy and feature-subset cardinality (cf., Section 3.5), TCMI has lower prediction errors as FS-SISSO, while TB3-GBDT and TB3-SISSO consistently yield the best prediction performance and the smallest feature-subset cardinality.

It is to be noted that the identified feature subsets of the TB3-GBDT algorithm do not necessarily lead to the same feature subsets as the identified feature subsets of the TB3-SISSO algorithm, i.e., the transferability of identified (sub-optimal) feature subsets is limited when different machine-learning algorithms are used for the search and model construction. A comparison (Fig. 5.3) shows that in all 9 non-redundant feature subsets identified by TB3-GBDT and 49 non-redundant feature subsets identified by TB3-SISSO, only three of the identified feature subsets are found by both methods: $\{\mathrm{EA(A)}, \mathrm{r}_p\mathrm{(A)}, \mathrm{IP(B)}\}$, $\{\mathrm{H(A)}, \mathrm{r}_s\mathrm{(A)}, \mathrm{EN(B)}\}$, and $\{\mathrm{r}_s\mathrm{(A)}, \mathrm{EA(B)}, \mathrm{r}_p\mathrm{(B)}\}$. This corresponds to a Jaccard similarity coefficient of 0.05, which can be explained by the fact that the limited availability of materials data in the data set leads to differences in the identification of feature subsets of multicollinear features (Fig. 5.3, no. 4 and 14).

A frequency analysis (Fig. 5.4) shows that (although different machine-learning algorithms and hence different feature-selection criteria were used[9]) the atomic radii $\mathrm{r}_s\mathrm{(A)}$, $\mathrm{r}_p\mathrm{(A)}$ are the most frequent selected features for estimating the energy difference between rock-salt and zinc-blende crystal structures, followed by the radius $\mathrm{r}_p\mathrm{(B)}$, the Mulliken electronegativity EN(B) and related features (IP, EA) of both atomic species. These results are in agreement with prior investigations, where the atomic radii $\{\mathrm{r}_s\mathrm{(A)}, \mathrm{r}_p\mathrm{(A)}, \mathrm{r}_s\mathrm{(B)}, \mathrm{r}_p\mathrm{(B)}\}$ [418, 419][10] and the feature subset $\{\mathrm{EA(B)}, \mathrm{IP(B)}, \mathrm{r}_s\mathrm{(A)}, \mathrm{r}_p\mathrm{(A)}, \mathrm{r}_d\mathrm{(A)}, \mathrm{r}_s\mathrm{(B)}, \mathrm{r}_p\mathrm{(B)}\}$ [55, 161] have been identified as relevant. In fact, the atomic radii are

[9] As a reminder, wrapper methods identify the relevant features based on the selected machine-learning algorithm and model. The identification of features as such is thus significantly influenced by the model parameters used. It therefore requires the determination of an appropriate set of parameters for each model in order to compare different feature-identification methods (cf., Sec. 5.1).

[10] John & Bloch [418] uses the difference between the total effective core radii of atoms A and B (r_σ) against the corresponding s-p hybridization of both sites (r_π) to separate the crystal structures of binary compounds,

$$r_\sigma = \left|\left(\mathrm{r}_p\mathrm{(A)} + \mathrm{r}_s\mathrm{(A)}\right) - \left(\mathrm{r}_s\mathrm{(B)} + \mathrm{r}_p\mathrm{(B)}\right)\right|, \quad r_\pi = \left(\mathrm{r}_p\mathrm{(A)} - \mathrm{r}_s\mathrm{(A)}\right) + \left(\mathrm{r}_p\mathrm{(B)} - \mathrm{r}_s\mathrm{(B)}\right).$$

Zunger [419] uses the difference between the total effective core radii of atoms A and B and the sum of the orbital non-locality of the s- and p-electrons on each site,

$$r_\sigma = \left|\left(\mathrm{r}_p\mathrm{(A)} + \mathrm{r}_s\mathrm{(A)}\right) - \left(\mathrm{r}_s\mathrm{(B)} + \mathrm{r}_p\mathrm{(B)}\right)\right|, \quad r_\pi = \left|\mathrm{r}_p\mathrm{(A)} - \mathrm{r}_s\mathrm{(A)}\right| + \left|\mathrm{r}_p\mathrm{(B)} - \mathrm{r}_s\mathrm{(B)}\right|.$$

Both, definitions use different values for r_s and r_p: John & Bloch [418] derived the atomic positions from the "Pauli-force" model potential introduced by Simons [430], whereas Zunger [419] derived the atomic radii from the screened density-functional atomic pseudo-potentials. In comparison, Ghiringhelli, Vybiral *et al.* [55] computed the atomic radii using density-functional theory [25, 26] within the local-density approximation [428]. As for some elements (carbon, nitrogen, TB3-GBDT

1 {r_s(A), EA(B), r_s(B)}
2 {EN(A), r_s(A), r_p(B)}
3 {EN(A), r_s(A), r_s(B)}
4 {EN(A), r_s(A), IP(B)}
5 {r_d(A), r_s(A), r_p(B)}
6 {L(A), r_p(A), r_p(B)}

TB3-SISSO

7 {IP(A), r_p(A), IP(B)}
8 {r_s(A), EA(B), r_p(B)}
9 {r_s(A), EN(B)}
10 {EA(A), r_p(A), IP(B)}
11 {r_p(A), EN(B)}

12 {EN(A), H(A), IP(A), r_p(A), r_s(A), IP(B)}
13 {EN(A), r_d(A), r_p(A), EA(B), r_p(B), r_s(B)}
14 {H(A), r_p(A), r_s(A), IP(B), r_p(B)}
15 {EN(A), r_p(A), EN(B), r_d(B), r_p(B)}
16 {H(A), r_p(A), r_s(A), H(B), r_p(B)}
17 {IP(A), r_d(A), r_p(A), EA(B), r_s(B)}
18 {H(A), IP(A), r_p(A), r_s(A), IP(B)}
19 {EN(A), r_d(A), r_p(A), EN(B), r_p(B)}
20 {EN(A), IP(A), r_p(A), EA(B), r_s(B)}
21 {IP(A), r_p(A), r_s(A), IP(B), r_p(B)}
22 {H(A), r_p(A), H(B), IP(B), r_p(B)}
23 {IP(A), r_p(A), IP(B), L(B), r_p(B)}
24 {H(A), r_p(A), IP(B), r_s(B)}
25 {H(A), IP(A), r_p(A), r_p(B)}
26 {EA(A), EN(A), r_p(A), r_p(B)}
27 {H(A), r_p(A), EN(B), r_p(B)}
28 {r_d(A), r_p(A), r_d(B), r_p(B)}
29 {EN(A), IP(A), r_p(A), r_p(B)}
30 {r_d(A), r_p(A), r_p(B), r_s(B)}
31 {H(A), r_p(A), EN(B), r_s(B)}
32 {r_d(A), r_p(A), IP(B), r_s(B)}
33 {H(A), IP(A), r_p(A), IP(B)}
34 {r_d(A), r_s(A), H(B), r_s(B)}
35 {H(A), IP(A), r_p(A), r_s(B)}
36 {IP(A), r_p(A), EA(B), r_s(B)}
37 {r_p(A), EN(B), IP(B), r_p(B)}
38 {r_d(A), r_p(A), IP(B), r_p(B)}
39 {r_d(A), r_p(A), IP(B), L(B)}
40 {L(A), r_p(A), r_s(A), r_p(B)}
41 {H(A), r_p(A), EA(B), IP(B)}
42 {EN(A), r_p(A), EA(B), r_s(B)}
43 {EN(A), r_d(A), r_p(A), r_s(B)}
44 {EN(A), H(A), r_p(A), r_p(B)}
45 {r_p(A), r_s(A), L(B), r_p(B)}
46 {r_d(A), r_p(A), EN(B), r_p(B)}
47 {r_p(A), r_s(A), r_p(B), r_s(B)}
48 {EN(A), r_s(A), H(B), r_s(B)}
49 {r_d(A), r_p(A), H(B), r_p(B)}
50 {H(A), r_p(A), IP(B), r_p(B)}
51 {EN(A), r_d(A), r_p(A), r_p(B)}
52 {H(A), r_p(A), r_s(A), r_s(B)}
53 {r_d(A), r_p(A), r_s(A), r_s(B)}
54 {H(A), r_p(A), H(B), r_p(B)}
55 {r_d(A), r_s(A), EA(B), r_s(B)}
56 {H(A), r_s(A), EA(B), r_s(B)}
57 {r_p(A), EA(B), r_p(B)}
58 {r_p(A), r_d(B), r_p(B)}
59 {r_p(A), H(B), r_s(B)}
60 {EA(A), r_p(A), r_p(B)}
61 {IP(A), r_p(A), r_p(B)}
62 {r_p(A), EN(B), r_s(B)}
63 {r_p(A), r_s(A), r_p(B)}
64 {H(A), r_s(A), r_p(B)}
65 {r_p(A), r_d(B), r_s(B)}
66 {IP(A), r_s(A), H(B)}
67 {r_p(A), H(B), r_p(B)}
68 {EA(A), r_p(A), r_d(B)}
69 {L(A), r_p(A), EN(B)}
70 {r_s(A), IP(B), r_s(B)}
71 {EN(A), r_s(A), EN(B)}
72 {L(A), r_p(A), IP(B)}
73 {r_d(A), r_p(A), IP(B)}
74 {r_p(A), EA(B), EN(B)}
75 {r_s(A), H(B), r_s(B)}
76 {r_p(A), EA(B), H(B)}
77 {H(A), r_p(A), EN(B)}
78 {r_s(A), H(B), r_p(B)}
79 {r_s(A), EN(B), r_s(B)}
80 {EA(A), r_s(A), r_s(B)}
81 {EA(A), r_s(A), H(B)}
82 {L(A), r_s(A), H(B)}
83 {r_s(A), EA(B), r_s(B)}
84 {r_d(A), r_s(A), IP(B)}
85 {L(A), r_s(A), r_d(B)}
86 {EN(A), r_p(A), IP(B)}
87 {r_p(A), r_s(B)}
88 {r_p(A), r_p(B)}
89 {r_p(A), r_d(B)}
90 {r_p(A), IP(B)}
91 {r_s(A), r_p(B)}
92 {r_p(A), EA(B)}

■ TB3-GBDT (1–6) □ TB3-GBDT ∩ TB3-SISSO (7–11) □ TB3-SISSO (11–92)

Fig. 5.3. Identified minimally non-redundant (sub-optimal) feature subsets of the tolerance-based branch-and-bound algorithm (TB3) with GBDT or SISSO as the feature-selection criterion (Section 4.1). Shown is the intersection and union of the two feature-identification methods corresponding to a Jaccard index of 0.06. It should be noted that due to the different machine-learning algorithms, a direct comparison between TB3-GBDT and TB3-SISSO may not be appropriate. The numbering is used for referencing purposes.

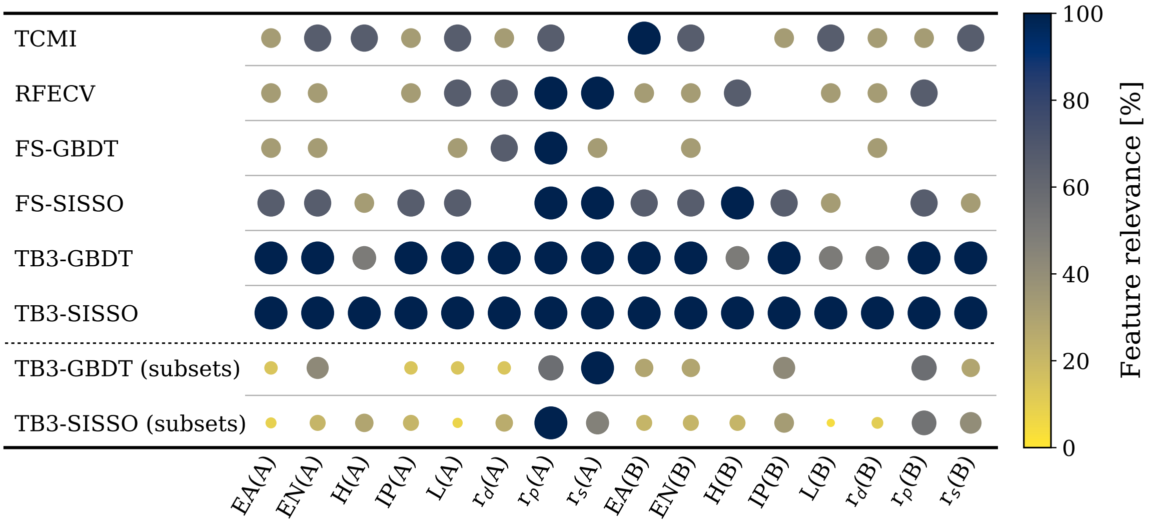


Fig. 5.4. Heat-map of most frequent selected or occurring features in identified (redundant) feature subsets of the investigated feature-identification methods (cf., Section 5.1). The size and color reflect the frequency of the features (= relevance). A frequency analysis of identified minimal non-redundant feature subsets of the tolerance-based branch-and-bound algorithm (TB3) with GBDT and SISSO is also performed and shown for reference.

closely related to the pseudo-potential of an atom and thus determine the total energy of a material, while the electronegativity and electron affinity are a measure of the attractive interaction between a valence electron and the atomic nucleus and thus determine the type of bonding (e.g., covalent or ionic bonding).

Results further confirm that feature subsets are not symmetric with respect to the exchange of atomic species *A* and *B*, e.g., that features of both atomic species do not appear symmetrically in the identified feature subsets (cf., [55]). In particular, there is a general tendency to assign greater relevance to the features of atomic species *A* than to those of atomic species *B*. For instance, L(A) is more frequently selected than L(B). This asymmetry is independent of whether or not the Mulliken electronegativity has been identified in the minimal non-redundant feature subsets and can be explained by the fact that the atomic species *A* constrains the electro-chemical properties of the atomic species *B*. A symmetrization of the features was explicitly constructed and tested, but resulted in the same identification of the atomic radii (r_s, r_p) and the Mulliken electronegativity (EN) as the most relevant features of the data set. Thus, even though the Mulliken electronegativity is not part of many of the identified minimal non-redundant feature subsets, it still plays an important role in the statistical description of the energy difference of rock-salt and zinc-blende crystal structures of octet-binary compound semiconductors.

oxygen, and fluorine) the atomic radius r_s is larger than r_p, Ghiringhelli, Vybiral *et al.* used the absolute differences and sums of the atomic radii,

$$r_\sigma = \left|r_p(\mathrm{A}) + r_s(\mathrm{A})\right| - \left|r_s(\mathrm{B}) + r_p(\mathrm{B})\right|, \quad r_\pi = \left|r_p(\mathrm{A}) - r_s(\mathrm{A})\right| + \left|r_p(\mathrm{B}) - r_s(\mathrm{B})\right|.$$

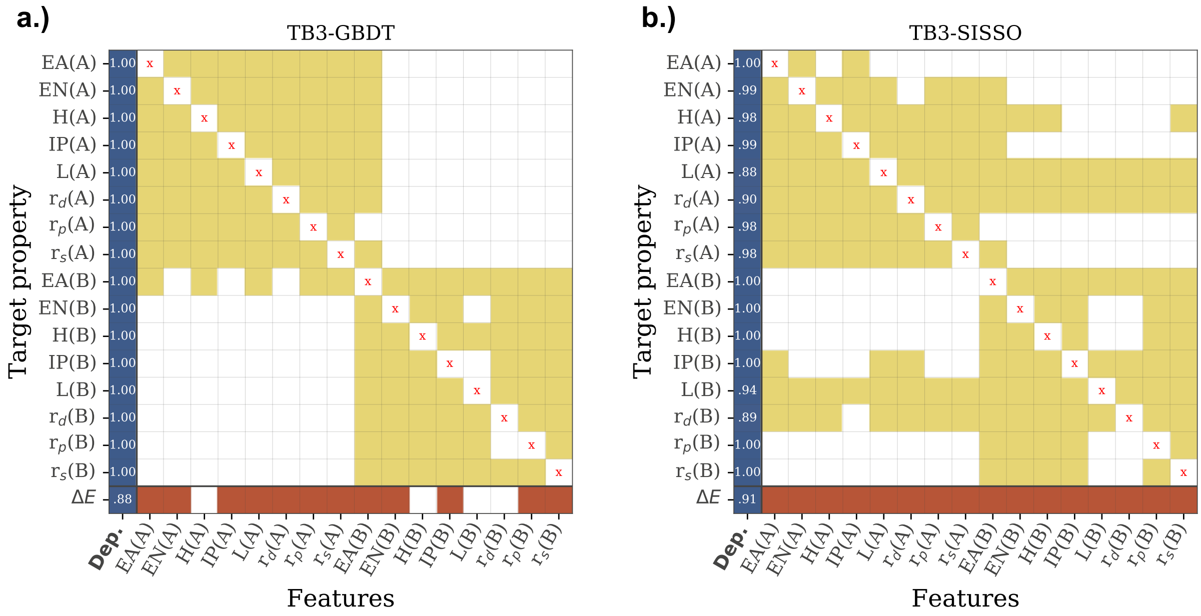


Fig. 5.5. Feature-dependence maps of the octet-binary compound semiconductor data set [55, 161] created with the tolerance-based branch-and-bound algorithm (TB3) and two different machine-learning algorithms: (a) gradient-boosting decision trees (GBDT) [249–252, 267, 268] and (b) the sure-independence screening and sparsifying operator (SISSO) [197]. Feature-dependence maps were created at a confidence level of $\alpha = 0.95$ and a convergence threshold of $\epsilon = 0.01$ (cf., Section 4.2). The score(=strength) of the dependence is shown in the first column (Dep.). The score is the Pearson's R coefficient of determination [101] of the 10-fold cross-validated machine-learning models based on the identified features (colored boxes). The dependence of the energy difference between rock-salt and zinc-blende crystal structures (ΔE) is given in the last row.

Feature dependences

In comparison to the minimally non-redundant feature subsets found by the tolerance-based branch-and-bound algorithm (TB3, Tab. 5.1), redundant feature subsets consist of almost all features of the data set. This indicates that there are many interactions between the features and the energy difference of rock-salt and zinc-blende crystal structures. Indeed, the feature-dependence maps generated with TB3-GBDT and TB3-SISSO reveal a block-like structure of feature interactions between atomic features for each of the atomic species A or B (Fig. 5.5). A more detailed analysis of the identified non-redundant feature subsets identifies not only physical relationships such as $\mathrm{H} - \mathrm{L} \propto \mathrm{IP} - EA$ [161] or $\mathrm{EN} = -(\mathrm{IP}+\mathrm{EA})/2$, but also completely dependent features such as r_p with r_s or the electron affinity EA(B) to features of atomic species A or vice versa (Fig. 5.5). These findings are not surprising as the atomic properties of each atomic species A/B implicitly depend on the atomic charge Z (and therefore are necessarily related) and the inherent ordering of the atomic species in ascending order of the Mulliken electronegativity constrains the electro-chemical properties of constituent elements A/B. Feature-dependence maps further reveal a relation between the electron affinity (EA) of atomic species B to estimate the atomic properties of atomic species A. The set of atomic properties is therefore not independent (Fig. 5.5).

Differences in the feature-dependence maps of the GBDT or the SISSO algorithm are due to small fluctuations in the cross-validated prediction performances of identified non-redundant sub-optimal feature subsets. For example, in the TB3-GBDT search, the highest occupied molecular orbital H of the atomic species B can be almost completely estimated by the Mulliken electronegativity EN(B) or the ionization potential alone IP(B) ($R^2 = 0.99$). However, the prediction performance slightly improves ($\Delta R^2 = +0.01$), once a second feature from the atomic species B is added. Similarly, L and r$_d$ are dependent of both atomic species in the TB3-SISSO search: the prediction performance of L and r$_d$ are slightly better ($\Delta R^2 = +0.02$) when features of the respective other atomic species are present in the minimally non-redundant feature subsets.Although the electron affinity (EA) can in principle be expressed by any other combination of elemental properties of the same atomic species, the TB3-algorithm only identifies the electron affinity as a function of the Mulliken electronegativity (EN) and the ionization potential (IP) alone. This is due to the fact, that the TB3-algorithm searches for the optimal minimally non-redundant feature subsets. Thus, feature subsets with larger feature-subset cardinalities, but same prediction errors, cannot be identified and hence are not included in the feature-dependence maps.

A comparison of the TB3-GBDT and TB3-SISSO feature-dependence maps otherwise shows a qualitatively good agreement in the feature dependences: both feature-dependence maps have a block-like structure of feature interactions and the energy difference between rock-salt and zinc-blende crystal structures is a non-linear function of the atomic species, $\Delta E = \Delta E(Z_A, Z_B)$ (Fig. 5.5). As the features also implicitly dependent on the atomic charges Z, it is to be expected that ΔE can be accurately estimated by at least two features, one from each atomic species A/B. Indeed, identified minimally non-redundant feature subsets for the prediction of ΔE always consist of at least one feature from each atomic species A/B. Especially TB3-SISSO and TB3-GBDT (Fig. 5.3) further suggest that the electron affinity EA(B) is important in the estimation of ΔE due to strong statistical correlations to both atomic species.

Statistical models

The prediction performance of identified non-redundant sub-optimal feature subsets of TB3-GBDT and TB3-SISSO are summarized in Tabs. 5.2 and 5.3. Machine-learning models can accurately predict the energy differences between rock-salt and zinc-blende crystal structures of octet-binary compound semiconductors with an $R^2 > 0.9$. Yet, they are characterized by large standard deviations (of up to 50% of the actual values) in the cross-validation prediction performance. The high variance suggests that the performance of the model may be limited by the small amount of materials data, or that the relationship cannot be accurately modeled by an analytic relationship with the present features.

GBDT algorithms [249–252, 267, 268] are known to have issues with modeling very few data samples [243, 358–360, 431]. As such, machine-learning models constructed with feature subsets identified by the TB3-GBDT algorithm (or using the GBDT algorithm for modeling) have larger prediction errors (RMSE $\approx$ 100 meV, $R^2 \approx 0.91$) than machine-learning models with the TB3-SISSO algorithm (RMSE $\approx$ 80 meV, $R^2 \approx 0.95$, Tabs. 5.2 and 5.3).

Prediction performances of constructed machine-learning models from both methods are comparable to prior investigations [55, 161, 197] (all with an approximately cross-validated RMSE $\approx$ 80 meV

#	Features	GBDT		SISSO	
	+ Model	R^2	RMSE [meV]	R^2	RMSE [meV]
1	**EN(A), r$_s$(A), IP(B)**	0.85±0.13	138±123	0.91±0.08	106±39
	$\Delta E = -0.256 \cdot \frac{\left(\sqrt[3]{EN(A)} + \sqrt[3]{IP(B)}\right) EN(A)}{r_s(A)^4} - 699 \cdot \left(\ln\left(r_s(A)\right)\right)^2 \exp\left(EN(A)+IP(B)\right) + 0.0892 \cdot \left(EN(A) - IP(B) + \ln\left(EN(A)\right)\left(EN(A)+IP(B)\right)\right) - 0.688$				
2*	**r$_s$(A), EA(B), r$_p$(B)**	0.81±0.26	142±140	0.92±0.10	97±41
	$\Delta E = 3.96 \cdot \frac{\sqrt{r_p(B)}}{\exp\left(r_s(A)\right)\left(r_p(B)^3 + r_s(A)^3\right)} + 0.159 \cdot \frac{r_p(B)+r_s(A)}{r_s(A)^2 \left(\exp\left(EA(B)\right) + \exp\left(-r_s(A)\right)\right)} - 1.03 \cdot \frac{r_s(A)^2 \left(r_p(B) - r_s(A)\right) \ln\left(r_p(B)\right)}{EA(B)^3} - 0.974$				
3	**IP(A), r$_p$(A), EN(B)**	0.85±0.15	145±171	0.86±0.12	232±402
	$\Delta E = 0.0475 \cdot \frac{EN(B)^3}{r_p(A)^2 \sqrt[3]{\exp\left(EN(B)\right)}} + 0.0539 \cdot \frac{r_p(A)}{\ln\left(r_p(A)\right)} \cdot \left\lvert\frac{EN(B)}{IP(A)} - \frac{IP(A)}{EN(B)}\right\rvert - 0.0221 \cdot \left\lvert\frac{EN(B)+IP(A)}{\ln\left(r_p(A)\right)} - \frac{EN(B)IP(A}{EN(B)-IP(A)}\right\rvert - 0.469$				
4	**r$_p$(A), EN(B)**	0.83±0.18	157±177	0.85±0.13	125±052
	$\Delta E = 0.0454 \cdot \frac{EN(B)^3}{r_p(A)^2\sqrt[3]{\exp\left(EN(B)\right)}} - 0.233 \cdot \exp\left(-\frac{1}{EN(B)\ln\left(r_p(A)\right)}\right) + 4.12\text{e-}06 \cdot r_p(A)^3 r_p(A) \exp\left(r_p(A)\right) \cdot EN(B) \ln\left(EN(B)\right) - 0.307$				
5	**EN(A), r$_s$(A), r$_p$(B)**	0.84±0.12	143±117	0.82±0.13	251±262
	$\Delta E = 3.86 \cdot \frac{\sqrt{r_p(B)}}{\exp\left(r_s(A)\right)\left(r_p(B)^3+r_s(A)^3\right)} + 0.45 \cdot \left\lvert\left(\ln\left(r_s(A)\right)\right)^2 - \frac{\ln\left(r_s(A)\right) r_p(B)}{r_s(A)}\right\rvert + 3.85\text{e-}06 \cdot \frac{\left(\exp\left(EN(A)\right)\right)^2}{\ln\left(r_s(A)\right) EN(A) r_s(A)} - 0.385$				
6*	**EA(A), r$_p$(A), IP(B)**	0.86±0.10	137±126	0.83±0.16	149±87
	$\Delta E = -0.243 \cdot \frac{\ln\left(EA(A)-IP(B)\right)\sqrt[3]{IP(B)}}{r_p(A)^2} - 0.646 \cdot \frac{\exp\left(-r_p(A)\right) - \ln\left(r_p(A)\right)}{EA(A)IP(B)} - 6.74\text{e-}05 \cdot \frac{\exp\left(-r_p(A)\right)}{\ln\left(r_p(A)\right)\sqrt{\exp\left(IP(B)\right)}} - 0.389$				

(continues on next page)

(continued from previous page)

7	**EA(A), r$_p$(A), EN(B)**	0.86±0.13	143±168	0.86±0.19	131±106
	$\Delta E = 0.0467 \cdot \frac{EN(B)^3}{r_p(A)^2 \sqrt[3]{\exp\big(EN(B)\big)}} + 0.383 \cdot \frac{\exp\big(-r_p(A)\big) - \ln\big(r_p(A)\big)}{EA(A)EN(B)} - 0.0106 \cdot \left\lvert \frac{EA(A)^2}{\ln\big(r_p(A)\big)} - \frac{EN(B)^2}{\exp\big(r_p(A)\big)} \right\rvert - 0.378$				
8	**IP(A), r$_s$(A), EN(B)**	0.85±0.20	135±165	0.85±0.19	134±74
	$\Delta E = 0.199 \cdot \frac{\Big(\ln\big(EN(B)\big)\Big)^2}{r_s(A)^3 \sqrt{r_s(A)}} - 185 \cdot \exp\big(IP(A)r_s(A)\big) \cdot \left\lvert \frac{EN(B)}{IP(A)} - \frac{IP(A)}{EN(B)} \right\rvert + 0.0377 \cdot \frac{\ln\big(r_s(A)\big)}{EN(B)IP(A)\ \Big(\exp\big(IP(A)\big) + \exp\big(-EN(B)\big)\Big)} - 0.164$				
9*	**H(A), r$_s$(A), EN(B)**	0.85±0.20	134±16	0.85±0.26	116±53
	$\Delta E = 0.0265 \cdot \frac{EN(B)^3}{r_s(A)^3 \sqrt[3]{\exp\big(EN(B)\big)}} + 0.052 \cdot \left\lvert \frac{H(A)^2}{EN(B) - H(A)} - \ln\big(r_s(A)\big)\big(H(A) + EN(B)\big) \right\rvert - 3.55\text{e-}06 \cdot EN(B)^3 \ln\big(r_s(A)\big)\ \Big(\exp\big(-H(A)\big) + \exp\big(r_s(A)\big)\Big) - 0.434$				

Tab. 5.2. Ranked list of statistical equivalent symbolic-regression models for the prediction of the energy difference ΔE between rock-salt and zinc-blende crystal structures in ascending order of the SISSO [197] prediction errors (RMSE - ΔRMSE). Shown is the prediction performance of identified, best-performing, non-redundant, sub-optimal feature subsets of the tolerance-based branch-and-bound algorithm (TB3) using the gradient-boosting decision trees (GBDT) algorithm [249–252, 267, 268] as feature-selection criterion Sections 3.1.2 and 4.1). For comparison, 10-fold cross-validated (cf., Section 5.1) prediction performances of GBDT and the sure-independence screening and sparsifying operator (SISSO) [197] are reported in terms of the mean $\bar{x}$ and standard deviation Δx as $\bar{x} \pm \Delta x$ of the Pearson's coefficient of determination $x = R^2$ [101] and the root-mean-squared error (x = RMSE). The RMSE is in units of millielectronvolts (meV). Feature subsets marked with an asterisk (*) are identified by TB3 using SISSO as the feature-selection criterion (cf., Tab. 5.3).

#	**Features**	**GBDT**		**SISSO**	
	+ Model	R^2	RMSE [meV]	R^2	RMSE [meV]
1	**EA(A), r$_p$(A), r$_p$(B)**	0.83±0.15	144±126	0.95±0.04	71±38
	$\Delta E = 1.08 \cdot \frac{\big(r_p(A) + r_p(B)\big)}{\sqrt[3]{r_p(A)}\big(r_p(A)^3 + r_p(B)^3\big)} + 0.0555 \cdot \left\lvert \sqrt{\exp\big(r_p(A)\big)} - \frac{\exp\left(EA(A)\right)}{\exp\left(-r_p(B)\right)} \right\rvert - 0.0669 \cdot \left\lvert \frac{\exp\big(-EA(A)\big)r_p(B)}{r_p(A)} - \Big\lvert \exp\big(r_p(B)\big) - \exp\big(-EA(A)\big) \Big\rvert \right\rvert - 0.405$				
2	**L(A), r$_p$(A), r$_s$(A), r$_p$(B)**	0.77±0.28	161±145	0.95±0.04	64±23
	$\Delta E = 5.76 \cdot \frac{\exp\big(-r_s(A)\big)}{r_p(B)^2 \exp\big(r_p(A)/r_p(B)\big)} - 0.586 \cdot L(A)r_p(B)\big(r_p(B) - r_s(A)\big) \cdot \left\lvert \frac{L(A)}{r_p(A)} - \frac{L(A)}{r_s(A)} \right\rvert - 0.0288 \cdot \frac{\Big\lvert r_p(B) - \big\lvert r_p(B) - r_s(A) \big\rvert \Big\rvert}{\exp\big(L(A)\big)\sqrt[3]{L(A)}} - 0.218$				

(continues on next page)

(continued from previous page)

3	**IP(A), $\mathbf{r}_p$(A), EN(B), $\mathbf{r}_s$(B)**	0.83±0.21	145±156	0.94±0.04	81±32
	$\Delta E = 1.72 \cdot \frac{\sqrt[3]{r_s(B)}}{r_p(A)^3 + r_s(B)^3} + 0.729 \cdot \frac{IP(A)^3 \lvert r_p(A) - r_s(B) \rvert}{\exp(EN(B)) + \exp(-IP(A))} - 5.43 \cdot \frac{\lvert r_s(B) - (r_p(A) - r_s(B)) \rvert r_p(A)}{\exp(r_p(A)) EN(B)} - 0.0817$				
4	**$\mathbf{r}_p$(A), EA(B), $\mathbf{r}_p$(B)**	0.77±0.29	152±138	0.94±0.04	75±26
	$\Delta E = 2.39 \cdot \frac{\exp(-r_p(A)/r_p(B))}{r_p(B)^2 \sqrt{r_p(A)}} - 0.0452 \cdot (r_p(A)^3)^3 \exp(EA(B) r_p(A)) + 0.533 \cdot \left\lvert \sqrt[3]{\exp(-r_p(A))} - (\exp(EA(B)) + \exp(-r_p(B))) \right\rvert - 0.277$				
5	**IP(A), $\mathbf{r}_p$(A), $\mathbf{r}_p$(B)**	0.84±0.13	139±130	0.94±0.05	81±24
	$\Delta E = 2.21 \cdot \frac{\exp(-r_p(A)/r_p(B))}{r_p(B)^2 \sqrt{r_p(A)}} - 0.00105 \cdot \frac{(IP(A)^2)^3 \exp(IP(A))}{\exp(-r_p(B))} - 3.24\text{e-}05 \cdot (r_p(A) - r_p(B))^3 IP(A)^3 IP(A) r_p(B) + 0.067$				
6	**$\mathbf{r}_d$(A), $\mathbf{r}_p$(A), $\mathbf{r}_s$(B)**	0.80±0.18	153±118	0.93±0.04	93±48
	$\Delta E = 1.48 \cdot \frac{r_d(A) - r_p(A)}{\lvert r_d(A) - r_p(A) \rvert \, (r_p(A)^3 + r_s(B)^3)} - 0.081 \cdot \frac{\lvert r_s(B) - \lvert r_p(A) - r_s(B) \rvert \rvert \, r_d(A)}{(r_d(A) - r_p(A)) r_p(A)} - 0.0826 \cdot \left\lvert \exp(r_s(B)) - \exp(-r_s(B)) - \frac{\lvert r_d(A) - r_s(B) \rvert}{\lvert r_d(A) - r_p(A) \rvert} \right\rvert - 0.129$				
7	**$\mathbf{r}_p$(A), $\mathbf{r}_s$(A), $\mathbf{r}_p$(B)**	0.71±0.48	171±154	0.95±0.06	84±47
	$\Delta E = 1.91 \cdot \frac{\sqrt{r_p(B)} r_s(A)}{r_p(A) \, (r_p(B)^3 + r_s(A)^3)} + 0.172 \cdot \left\lvert \frac{r_s(A)^2}{r_p(A)} - \exp(r_p(B)) \cdot \lvert r_p(A) - r_s(A) \rvert \right\rvert + 0.0495 \cdot \lvert r_p(B) + r_p(B) + r_s(A) - \exp(r_s(A)) \cdot \lvert r_p(A) - r_s(A) \rvert \rvert - 0.559$				
8	**H(A), IP(A), $\mathbf{r}_p$(A), $\mathbf{r}_p$(B)**	0.84±0.14	142±138	0.93±0.06	137±205
	$\Delta E = 1.91 \cdot \frac{\left(\sqrt[3]{r_p(B)}\right)^2}{r_p(A)^3 + r_p(B)^3} - 0.0174 \cdot (H(A) IP(A))^3 r_p(B) \exp(IP(A)) + 0.232 \cdot \exp\left(-\left\lvert \frac{H(A)}{r_p(B)} - \frac{IP(A)}{r_p(A)} \right\rvert\right) - 0.101$				

(continues on next page)

(continued from previous page)

9	**H(A), r$_s$(A), r$_p$(B)**	0.85±0.12	136±122	0.93±0.06	139±149
	$\Delta E = 1.41 \cdot \frac{\sqrt{r_p(B)/r_s(A)}}{r_p(B)^3 + r_s(A)^3} + 1.16 \cdot \left\lvert \sqrt[3]{\exp\big(H(A)\big)} - \big(\exp\big(H(A)\big) + \exp\big(-r_s(A)\big)\big) \right\rvert - 0.00618 \cdot \frac{\big(r_p(B) - r_s(A)\big)^2 r_p(B)^2}{\exp\big(H(A)\big)} - 0.422$				
10	**r$_d$(A), r$_p$(A), r$_d$(B), r$_p$(B)**	0.80±0.19	153±124	0.93±0.06	83±31
	$\Delta E = 3.89 \cdot \frac{\big(r_d(A) - r_p(B)\big)\, r_p(B)}{r_d(A)\,\big(r_p(A)^3 + r_p(B)^3\big)} - 0.324 \cdot \left\lvert \frac{r_d(B) \cdot \lvert r_d(A) - r_p(B) \rvert}{r_d(A)} - \frac{r_d(B) \cdot \lvert r_p(A) - r_p(B) \rvert}{r_p(A)} \right\rvert + 0.0935 \cdot \left\lvert \frac{\ln\big(r_p(A)\big) r_p(A)}{r_d(A)} - \frac{r_p(B)}{r_d(B} + \frac{r_p(B)}{r_p(A)} \right\rvert - 0.245$				
24*	**r$_s$(A), r$_p$(B), EA(B)**	0.81±0.26	142±140	0.92±0.10	97±41
	$\Delta E = 3.96 \cdot \frac{\sqrt{r_p(B)}}{\exp\big(r_s(A)\big)\,\big(r_p(B)^3 + r_s(A)^3\big)} + 0.159 \cdot \frac{r_p(B) + r_s(A)}{r_s(A)^2\,\big(\exp\big(EA(B)\big) + \exp\big(-r_s(A)\big)\big)} - 1.03 \cdot \frac{r_s(A)^2 \big(r_p(B) - r_s(A)\big) \ln\big(r_p(B)\big)}{EA(B)^3} - 0.974$				
43*	**EA(A), r$_p$(A), IP(B)**	0.86±0.10	137±126	0.83±0.16	149±087
	$\Delta E = -0.243 \cdot \frac{\ln\big(EA(A) - IP(B)\big) \sqrt[3]{IP(B)}}{r_p(A)^2} - 0.646 \cdot \frac{\exp\big(-r_p(A)\big) - \ln\big(r_p(A)\big)}{EA(A) IP(B)} - 6.74\text{e-}05 \cdot \frac{\exp\big(-r_p(A)\big)}{\ln\big(r_p(A)\big) \sqrt{\exp\big(IP(B)\big)}} - 0.389$				
50*	**H(A), r$_s$(A), EN(B)**	0.85±0.20	134±16	0.85±0.26	116±53
	$\Delta E = 0.0265 \cdot \frac{EN(B)^3}{r_s(A)^3 \sqrt[3]{\exp\big(EN(B)\big)}} + 0.052 \cdot \left\lvert \frac{H(A)^2}{EN(B) - H(A)} - \ln\big(r_s(A)\big)\big(H(A) + EN(B)\big) \right\rvert - 3.55\text{e-}06 \cdot EN(B)^3 \ln\big(r_s(A)\big)\,\big(\exp\big(-H(A)\big) + \exp\big(r_s(A)\big)\big) - 0.434$				

Tab. 5.3. Ranked list of statistical equivalent symbolic-regression models for the prediction of the energy difference ΔE between rock-salt and zinc-blende crystal structures in ascending order of the SISSO [197] prediction errors (RMSE - ΔRMSE). Shown is the prediction performance of identified, best-performing, non-redundant, sub-optimal feature subsets of the tolerance-based branch-and-bound algorithm (TB3) using the sure-independence screening and sparsifying operator (SISSO) [197] as feature-selection criterion Sections 3.1.2 and 4.1). For comparison, 10-fold cross-validated (cf., Section 5.1) prediction performances of SISSO and the gradient-boosting decision trees (GBDT) algorithm are reported in terms of the mean $\bar{x}$ and standard deviation Δx as $\bar{x} \pm \Delta x$ of the Pearson's coefficient of determination $x = R^2$ [101] and the root-mean-squared error (x = RMSE). The RMSE is in units of millielectronvolts (meV). Feature subsets marked with an asterisk (*) are also identified by TB3 using GBDT as the feature-selection criterion (cf., Tab. 5.2). In terms of prediction performance, TB3-SISSO outperforms TB3-GBDT (cf., Tab. 5.2).

cf., Tab. 5.1)[11], but depend on less than half of the features. According to Occam's razor [370], simpler models are preferable to more complex models. Feature identification can thus help reduce the number of features prior to statistical modeling and create statistical models that are more likely to capture the underlying relationships in the data.

Results further show that feature subsets including the atomic radii r_s and r_p (cf., Fig. 5.4) are among the best-performing feature subsets. This also applies to the feature subset suggested by John & Bloch [418, 419] – though with only three atomic radii $\mathrm{r}_s(\mathrm{A})$, $\mathrm{r}_p(\mathrm{A})$, $\mathrm{r}_p(\mathrm{B})$ (Tab. 5.3, #7) as the subset with all four atomic radii of the s- and p-orbitals turned out not to improve the Pearson's coefficient of determination in the subspace search ($R^2 = 0.93 \pm 0.07$, RMSE = (99 ± 68) meV as compared to $R^2 = 0.95 \pm 0.06$, RMSE = (84 ± 47) meV). Tests have shown that none of the feature subsets with a feature-subset cardinality $4 \leq n \leq 7$ resulted in a lower prediction error as the optimal non-redundant feature subset found by the TB3-algorithm (cf., Tabs. 5.2 and 5.3). As a consequence, the feature subset proposed by Ghiringhelli, Vybiral *et al.* [55, 161] was not found by the TB3-algorithm, although this feature set in principle consists of many of the relevant features as identified by the frequency analysis (cf., Fig. 5.4).

Common to TB3-GBDT and TB3-SISSO is the feature subset $\{\mathrm{r}_s(\mathrm{A}), \mathrm{EA}(\mathrm{B}), \mathrm{r}_p(\mathrm{B})\}$ (Tab. 5.2, #2 and Tab. 5.3, #24 or Tab. 5.3, #4 given the strongly linear correlation between r_s and r_p) in agreement with the most frequently identified relevant features (Fig. 5.4) and the results from the feature-dependence maps (Fig. 5.5). Given that a symbolic-regression model based on features subsets common to the TB3-GBDT algorithm, TB3-SISSO algorithm, and the most frequently identified features performs well, is a strong indicator of a model-independent identification of relevant features. However, it is to be emphasized that all symbolic-regression models of identified non-redundant sub-optimal feature subsets with TB3-GBDT and TB3-SISSO are statistically equivalent and none of the feature subsets is superior in terms of prediction performance at a confidence level of $\alpha = 0.95$. In fact, identified non-redundant sub-optimal feature subsets clearly indicate that there are many competing machine-learning models with distinct feature subsets but similar prediction performances (Chapter 4). As each statistical machine-learning model and feature subset represents a potential physical relationship in the data, there is no optimal feature subset and a single-best machine-learning model (cf., [358]) from which the energy difference between rock salt and zinc blende can be statistically determined. These results emphasize that the energy difference is a non-linear function of features of both atomic species and that there is no unique description for the quantitative prediction of the energy difference of octet-binary compound semiconductors data set (cf., Tab. 5.1 and Fig. 5.5). Yet, the multiplicity of machine-learning models facilitates a combination of competing machine-learning models to improve the overall performance of the machine-learning algorithm. Such an ensemble of machine-learning models is known to be more robust [416] and reliable in terms of estimating the uncertainty in the model predictions [360, 363, 378], while using all statistically equivalent machine-learning models for model predictions (cf., [358]).

[11] The generation of a symbolic-regression model based on the SISSO algorithm took one day to complete on one (only identified relevant features) or on three nodes (with all features of the data set) respectively, each with two Intel Xeon E5-2698 v3 processors (= 32 cores/node) and without hyper-threading.

↲ **Algorithm (TB3-...)**		**Ensemble prediction**			**Conformal prediction**		
	Confidence level (α)	Validity ($\hat{\alpha}$)	Mean (Δ)	Max (Δ)	Validity ($\hat{\alpha}$)	Mean (Δ)	Max (Δ)
a.) Full octet-binary compound semiconductors data set							
GBDT	0.50	*0.46*	0.05	0.54	*0.46*	0.05	0.13
	0.80	*0.72*	0.09	1.29	*0.87*	0.13	0.34
	0.95	*0.82*	0.12	2.19	*0.96*	0.31	0.77
SISSO	0.50	*0.55*	0.05	0.44	*0.52*	0.04	0.10
	0.80	*0.91*	0.09	0.81	*0.88*	0.10	0.25
	0.95	*0.94*	0.14	1.33	*0.99*	0.35	0.85
b.) Same data set with the top-most 5th-percentile of ΔE of octet-binary compound semiconductors not being used for model construction							
GBDT	0.50	*0.20*	0.05	0.63	*0.52*	0.08	0.24
	0.80	*0.46*	0.08	0.79	*0.87*	0.19	0.56
	0.95	*0.65*	0.11	1.06	*0.98*	0.49	1.63
SISSO	0.50	*0.37*	0.05	0.57	*0.52*	0.06	0.26
	0.80	*0.77*	0.10	0.97	*0.85*	0.15	0.64
	0.95	*0.93*	0.16	1.58	*0.96*	0.45	1.98

Tab. 5.4. Validity of ensemble and conformal prediction obtained from TB3-GBDT and TB3-SISSO at three different confidence levels $\alpha = [50\%, 80\%, 95\%]$, where the confidence level of the ensemble prediction was computed as the αth-percentile of the ensemble predictions. The validity ($\hat{\alpha}$) denotes the probability that the actual value x of the underlying process is inside the prediction interval $x \in [\bar{x} - \bar{\Delta}, \bar{x} + \bar{\Delta}]$ (Eq. 4.22) of a (point) prediction $\bar{x}$ with uncertainty $\pm\bar{\Delta}$ ($\bar{\Delta} = \text{Mean}(\Delta) \leq \text{Max}(\Delta)$). Shown are the performance statistics of the full octet-binary compound semiconductors data set (a) and the same data set with the top-most 5th-percentile of ΔE of octet-binary compound semiconductors not being used for model construction (b).

Uncertainty estimation

Uncertainty estimates based on an ensemble of machine-learning models take advantage of the fact that the prediction mean and uncertainty can be computed as the average and the αth-percentile of the different model predictions on the same set of data (Section 4.3). However, ensemble-based uncertainty estimates on small data sets are often biased, under-estimate the actual prediction error, and do not guarantee a reliable estimation of the prediction uncertainty on (new) data (Figs. 5.6 and 5.7 and Tab. 5.4). This can be seen for example in Fig. 5.6a., where the actual values of 44, 23, and 15 materials are outside the (50th, 80th, 95th)-percentile prediction intervals of the ensemble prediction. In contrast, prediction intervals based on conformal prediction have a statistical guarantee that in about α% of the cases the actual values of the target property are within the prediction interval at a confidence level of α% (Tab. 5.4).

Clearly, the validity (i.e., the percentage of actual values outside the prediction interval) of the conformal prediction [62–64] is violated at a confidence level of $\alpha = 0.50$ of the TB3-GBDT ensemble. Though this effect is not observed at higher confidence levels, a violation of the validity (Eqs. 4.22 and 4.27) is a strong indicator of a non-exchangeable data set (Section 4.3.1). Non-exchangeability implies that the materials of the data set are not randomly selected and the accuracy of the machine-learning model is significantly determined by the choice of training data. In fact, the large deviations of the symbolic-regression models (Tabs. 5.2 and 5.3), the proportional ensemble-based prediction errors to the energy difference between rock-salt and zinc-blende crystal structures (Fig. 5.6a. and c.), and the low credibility of zinc blende compounds such as boron nitride (BN) or diamond (CC) (Fig. 5.6b. and d.) corroborate the necessity to include crystal structures with a high energy difference into the training set to accurately model the octet-binary compound semiconductor data set (e.g., $R^2 \approx 0.98$, RMSE $\approx 80\,\mathrm{meV}$). Conversely, the exclusion of octet-binary compound semiconductors[12] with energy differences above the 95th-percentile in the data set results in at least 50% larger prediction errors ($R^2 \approx 0.96$, RMSE $\approx 134\,\mathrm{meV}$) and twice as large prediction intervals at the same validity (Tab. 5.4). Still, due to the very good extrapolation capabilities of the symbolic-regression models, there is a clear trend in the model predictions and a correct identification of boron nitride and diamond as the most stable zinc-blende crystal structures (cf., [161]).

It is to be noted that ensemble-based prediction errors are smaller than conformal-based prediction errors at the same confidence level α, but are as large as conformal-prediction intervals at the same validity $\hat{\alpha}$ (Tab. 5.4). As conformal prediction can be further characterized by a lower maximum prediction error Δ, lower variances in prediction errors, and statistical guarantees [363, 366] of the prediction uncertainties, conformal prediction provides a more robust estimate of the model's prediction errors than ensemble-based prediction.

Anomalous materials

One of the key advantages of conformal prediction is the identification of anomalous materials from the full set of data (Section 4.4). Anomalous materials are characterized by a low credibility, namely regions of the materials space that are underrepresented or cannot be adequately estimated by the specified machine-learning models [386]. For example, at a confidence level of $\alpha = 0.95$ there are two materials that are classified as anomalous (Fig. 5.6): boron nitride (BN) and diamond (CC). Both materials exhibit large prediction errors ($\Delta \geq 0.3\,\mathrm{eV}$ and $\varepsilon \geq 0.4\,\mathrm{eV}$) and are therefore difficult to estimate from the ensemble of machine-learning models: both with models based on all materials data (Fig. 5.6) and with models generated with materials from the lower 95th percentile of the energy difference (Fig. 5.7). Although boron nitride and diamond are not classified as anomalous at all confidence levels (especially when the top 5th-percentile of the materials is excluded), diamond and boron nitride are truly anomalous: Due to their small atomic sizes and fourfold strong, non-polar covalent bonds [432], they are not only the two most-stable zinc-blende crystal structures in the data set, but are also the hardest naturally occurring and abundant materials on Earth [433–435].

Summary

From the analysis of the octet-binary compound semiconductors data set, the following conclusion can be drawn: Due to the small number of materials in the data set, machine-learning predictions are characterized by large prediction errors and strong feature interactions between the features of each atomic species A/B. Although octet-binary compound semiconductors can in principle be

[12] Germanium carbide (GeC), boron arsenide (BAs), boron phosphide (BP), boron nitride (BN), and diamond (CC).

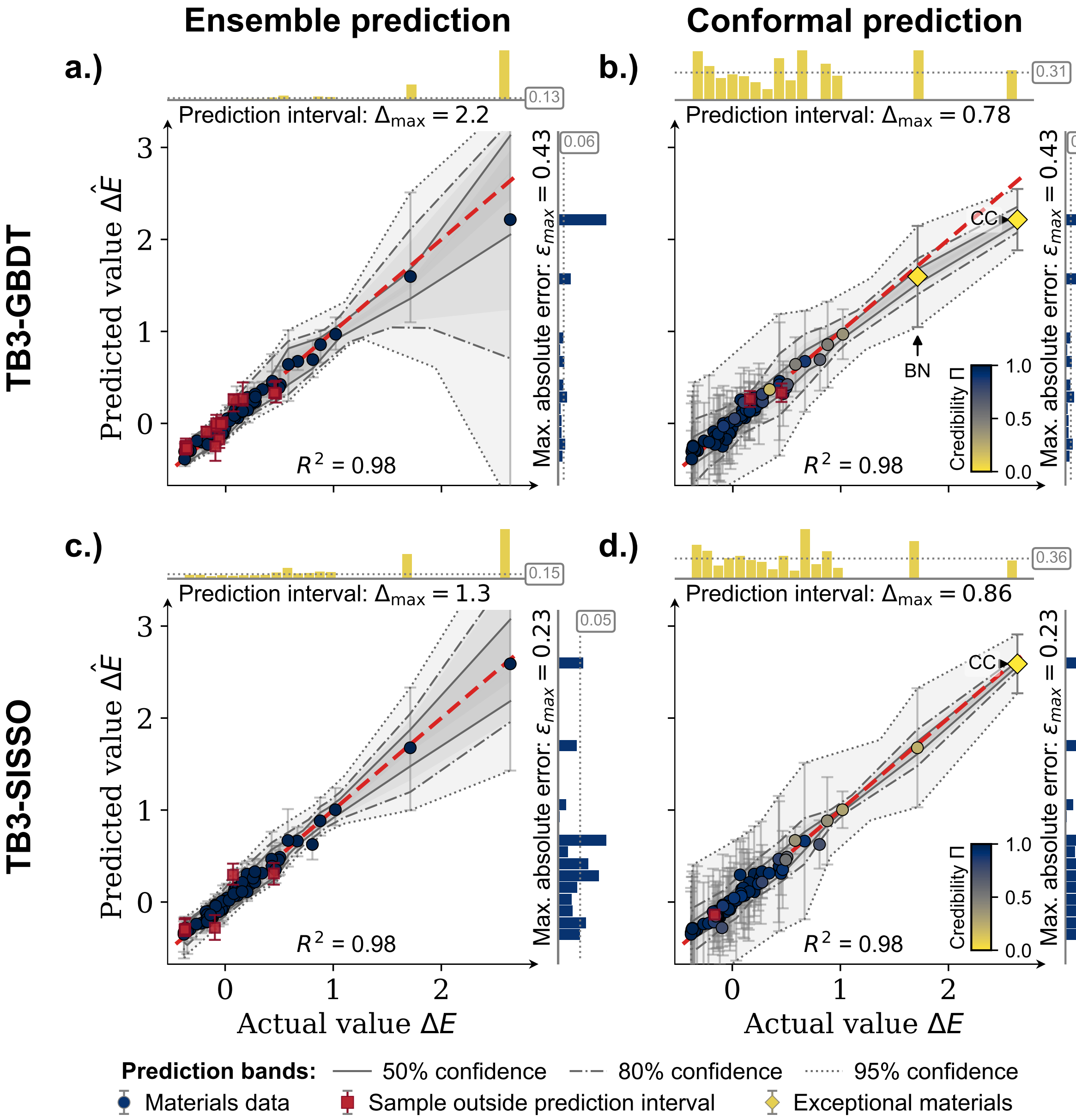


Fig. 5.6. Ensemble prediction performance of machine-learning models from TB3-GBDT and TB3-SISSO to estimate the energy difference ΔE between rock-salt and zinc-blende crystal structures. Shown are the prediction bands (50th, 80th, 95th-percentiles of model predictions), the credibility Π (Eq. 4.29) at a confidence level of $\alpha = 0.95$, the distribution of the prediction errors (diagram above the parity plot, Δ), and the errors of the machine-learning model (diagram right of the parity plot, $\varepsilon = |\Delta E - \hat{\Delta E}|$). The numbers in the boxes report the mean value, while the maximum errors are shown below or left of the diagrams. Units are in electronvolt (eV). Predictions outside the prediction intervals are depicted as squares and anomalous materials as diamond-shape symbols.

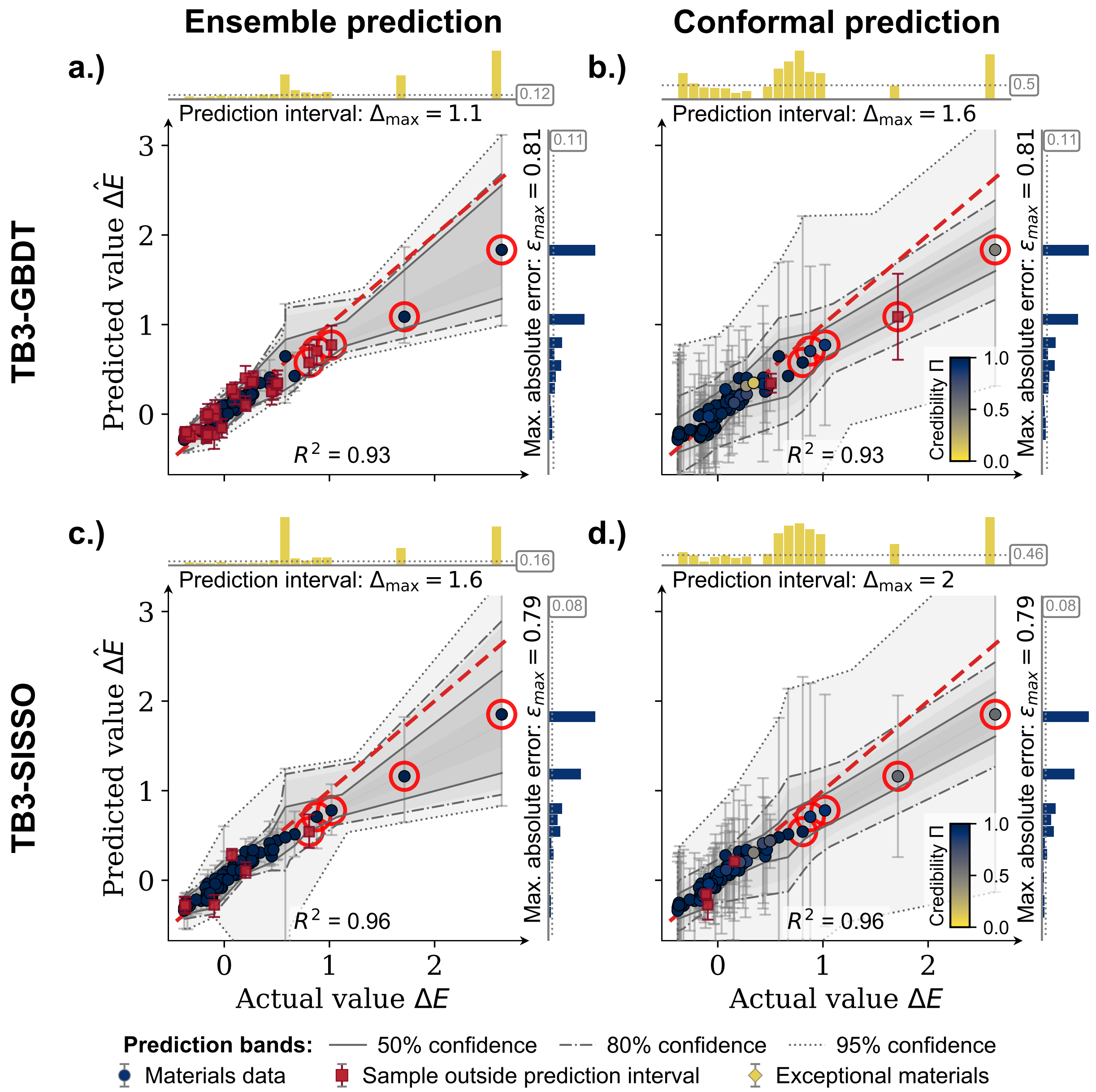


Fig. 5.7. Ensemble prediction performance of machine-learning models from TB3-GBDT and TB3-SISSO to estimate the energy difference ΔE between rock-salt and zinc-blende crystal structures. Shown are the prediction bands (50th, 80th, 95th-percentiles of model predictions), the credibility Π (Eq. 4.29) at a confidence level of $\alpha = 0.95$, the distribution of the prediction errors (diagram above the parity plot, Δ), and the errors of the machine-learning model (diagram right of the parity plot, $\varepsilon = |\Delta E - \hat{\Delta E}|$). The numbers in the boxes report the mean value, while the maximum errors are shown below or left of the diagrams. Units are in electronvolt (eV). Predictions outside the prediction intervals are depicted as squares and anomalous materials as diamond-shape symbols. In contrast to Fig. 5.6, octet-binary compound semiconductors with energy differences above the 95th-percentile in the data set (circles) were not used for model construction.

uniquely determined by two atomic properties, investigated feature-identification methods assign relevance to almost all features of the data set. With the exception of the TB3-algorithm, TCMI, RFECV, FS-GBDT, and FS-SISSO identify more features as dependent than actually required for creating machine-learning models. Thus, they can be viewed as conservative feature-identification methods for quantitatively predicting the crystal structure of octet-binary compound semiconductors (Tab. 5.1 and Fig. 5.5). In contrast, the TB3-algorithm (TB3) has been shown to yield the lowest prediction errors and the smallest feature-subset cardinality, comprising at least one feature from each atomic species A/B. In particular, machine-learning models constructed with the identified feature subset from TB3 have been shown to require fewer features, while achieving the same prediction error as the investigated feature-identification methods TCMI, RFECV, FS-GBDT, and FS-SISSO (cf., [55, 161, 197]).

Despite the difficulties of the data set, the developed framework for feature identification and model construction enables a systematic analysis of the data set, addresses the multiplicity of the material's problem, and provides a tool to estimate the prediction intervals of the machine-learning model's predictions. By comparing the identified non-minimal feature subsets of two different machine-learning algorithms (GBDT and SISSO), it has been argued that the limited availability of materials data led to differences in the identification of feature subsets of multi-collinear features (Fig. 5.3) and that the TB3-algorithm can only identify a subset of all sub-optimal non-redundant feature subsets in the data set (cf., Section 4.1). Common features subsets of both methods include the most frequently identified features from a frequency analysis of all feature-identification methods (Fig. 5.4): the atomic radii (r_s and r_p), the Mulliken electronegativity (EN), and the electron affinity (EA). While there is a clear indication that the atomic radii of the constituent elements statistically describe the energy difference between rock-salt and zinc-blende crystal structures (Tabs. 5.2 and 5.3), there is no optimal feature subset and single-best machine-learning model of the data set. It is rather the opposite. A set of competing machine-learning models with distinct feature subsets but statistically equivalent prediction performances have been constructed from the data set (Chapter 4). In addition, simpler models were found than in previous studies [55, 161, 197] and machine-learning models constructed from the TB3-SISSO algorithm (Tab. 5.3) outperformed those from the TB3-GBDT algorithm[13] (Tab. 5.2). Combined to an ensemble of machine-learning models, reliable uncertainty estimates revealed that boron nitride and diamond are the two materials with the largest prediction errors and the inclusion of crystal structures with a high ΔE into the training set is a necessary prerequisite for an accurate modeling of the octet-binary compound semiconductor data set.

5.2.2 Perovskites

Perovskites[14] are compounds with the general chemical formula ABX_3 (Fig. 5.8, right), where the A-site cation is 12-fold coordinated, the B-site cation is octahedrally 6-fold coordinated (and smaller than A), and the X-site is either a 2-coordinated oxygen ($X = \text{O}$) or halogen ($X = \{\text{F, Cl, Br, I, At, Ts}\}$) anion [410]. Common to all perovskite compounds is their structural and functional flexibility originating

[13] The energy difference between rock-salt and zinc-blende crystal structures of octet-binary compound semiconductors can be better modeled with a symbolic-regression model than with a piecewise-constant model.

[14] Named after the Russian mineralogist L. A. Perovski [436].

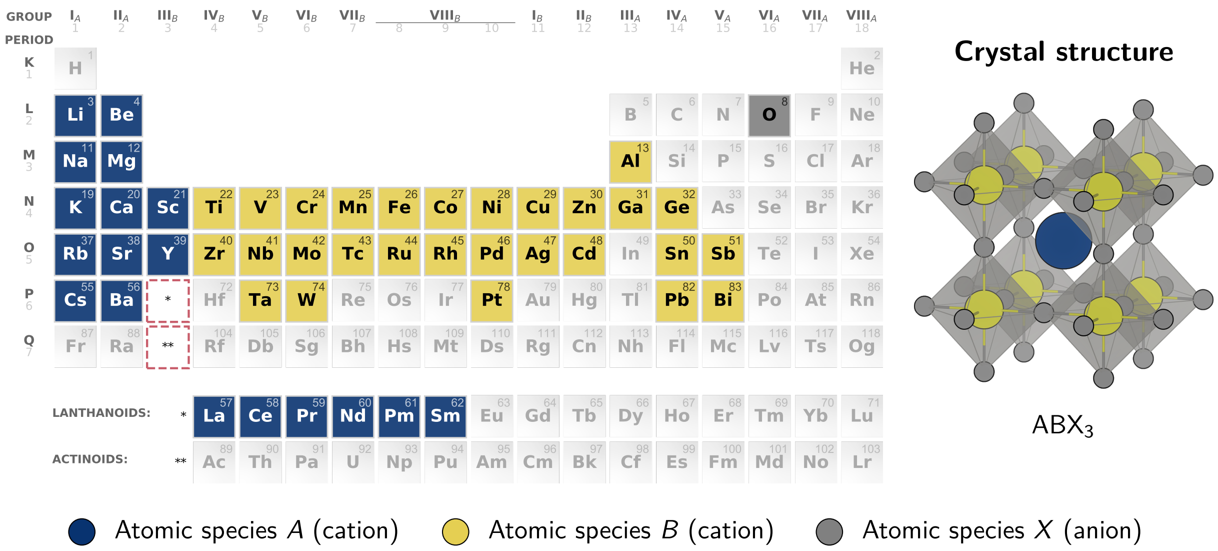

GROUP
PERIOD
I_A II_A III_B IV_B V_B VI_B VII_B VIII_B I_B II_B III_A IV_A V_A VI_A VII_A VIII_A
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
K L M N O P Q
H He
Li Be B C N O F Ne
Na Mg Al Si P S Cl Ar
K Ca Sc Ti V Cr Mn Fe Co Ni Cu Zn Ga Ge As Se Br Kr
Rb Sr Y Zr Nb Mo Tc Ru Rh Pd Ag Cd In Sn Sb Te I Xe
Cs Ba * Hf Ta W Re Os Ir Pt Au Hg Tl Pb Bi Po At Rn
Fr Ra ** Rf Db Sg Bh Hs Mt Ds Rg Cn Nh Fl Mc Lv Ts Og
LANTHANOIDS: * La Ce Pr Nd Pm Sm Eu Gd Tb Dy Ho Er Tm Yb Lu
ACTINOIDS: ** Ac Th Pa U Np Pu Am Cm Bk Cf Es Fm Md No Lr
Crystal structure
ABX3
Atomic species A (cation)
Atomic species B (cation)
Atomic species X (anion)

Fig. 5.8. Crystal structure of the ABX_3 perovskite and the used constituent elements in the 504 perovskite oxides data set [411]. Highlighted are elements in the periodic table that occur in the crystal structures as cations (A/B) and anions (X). The cations of the A-site include alkali metals, alkaline-earth metals, and rare earths, while the cations of the B-site comprise metals that form the center of a $[BX_6]^{4-}$ octahedra surrounded by oxygen anions (X-site).

from the compositional incorporation of a wide range of cations A/B of different sizes and oxidation states ($A^{1+}B^{5+}X_3^{2-}$, $A^{2+}B^{4+}X_3^{2-}$, $A^{3+}B^{3+}X_3^{2-}$), different dopant ratios and vacancies, and, by tilting or distorting the BO_6 octahedron, the symmetry breaking of the perovskite's cubic structure (Pm$\bar{3}$m) into a total of 14 space groups of tetragonal and rhombohedral phases [335, 437]. Owing to the compositional and functional flexibility [438], the structural family of perovskite compounds exhibits a large variety of interesting and intriguing physical and chemical properties including superconductivity [439, 440], ferreoelectricity [441], thermodynamical stability [171, 442, 443], catalytic reactivity [437, 438, 444], photoactivity [164, 445, 446], and electronic and ionic transport properties [447].

The dataset

The most common materials class of perovskite compounds are oxides ABO_3. Perovskite oxides have a high structural stability and are characterized by the possibility to control the oxidation state of the B-cation and to correlate the physical and chemical properties of the constituent elements to the property of interest by substituting the cation at the A-site with cations of different oxidation states or radii [438, 448]. As their structural stability is crucial for establishing design principles and tailoring perovskite oxides for catalysis [437, 438], Foppa, Scheffler *et al.* [411] analyzed a total of 504 cubic perovskite oxides (Fig. 5.8, left) for predicting the bulk modulus using machine learning. In their work, Foppa, Scheffler *et al.* successfully employed the symbolic-regression algorithm SISSO [197]. In particular, they modeled the explicit dependence of the bulk modulus B_0 as a function of the lattice constant a_0, the nuclear charges Z, and the ionic charge (= oxidation state) of the A-site cation c(A)[15] and, by applying SISSO [411] on the same set of features, achieved smaller prediction errors than using a semi-empirical formula [410].

Foppa, Scheffler *et al.* [411] computed the materials properties of the 504 perovskite oxides with the all-electron, full-potential electronic structure code FHI-aims [449] using numeric atom-centered basis functions, highly accurate ("tight") and scaled atomic zero-order regular approximation [450] at the GGA level of the theory (cf., Section 2.2) with the PBEsol exchange-correlation functional [451]. The k-points grids were converged with respect to the total energy (< 1 meV per atom in the unit cell) and a parametrically-constrained geometry relaxation [452] was performed to preserve the ideal cubic perovskite structure (Pm$\overline{3}$m space group), while optimizing the crystal-structure geometry with the Broyden-Fletcher-Goldfarb-Shanno algorithm [453–456]. The bulk modulus B_0 was then determined by fitting the Birch-Murnaghan equation of state [457, 458] to three electronic energies at and near ($\pm 1\%$) the equilibrium lattice constant a_0. They further computed 16 atomic features based on the non-spin polarized density-functional theory using the HSE exchange-correlation functional [459, 460] (cf., Section 2.2).

In total the data set includes 16 features[16]: eight for each atomic species (A/B). Among the atomic features are the nuclear charge (Z), the ionization potential (IP), electron affinity (EA), Mulliken electronegativity (EN), the energies of the highest-occupied and lowest-unoccupied Kohn-Sham levels, H and L, and the radius of maximum electronic density for the valence s- and p-orbitals, r_s and r_p [411]. The property of interests are the bulk modulus B_0 and the equilibrium lattice constant a_0 based on the 16 atomic features.

The difficulty of the data set is to find the relevant (atomic) features for the target properties B_0 and a_0 from a larger class of materials as compared to the octet-binary compound semiconductors data set. Thus, the feature identification and model construction is complicated by higher computational requirements for the construction of symbolic-regression models[17], strongly related features of the data set (cf., Fig. 5.12), and the modeling of the bulk modulus based on the atomic features from the same composition but different oxidation states [410, 411].

Feature identification

To better understand the structural stability and the material's behavior, relevant features are identified with the tolerance-based branch-and-bound algorithm (TB-GBDT, TB3-SISSO) and the four feature-identification methods from Chapter 3 (TCMI, RFECV, FS-GBDT, and FS-SISSO) on three randomly chosen sample subsets of 126, 252, and 504 perovskites oxides. Machine-learning models are finally built with the SISSO algorithm using fixed hyper-parameter settings [197] (cf., Sections 5.1.2 and A.1). Results are summarized in Tables 5.5 and 5.6.

[15] The ionic charge c(A) is the oxidation state of the cation at the site A, that roughly corresponds to the periodic table group of the A element and is 1 for alkali metals, 2 for alkaline earth metals and 3 for rare earths, including lanthanides, in the data set.

[16] The data set was kindly provided by the authors in a private communication.

[17] In contrast to the octet-binary compound semiconductor data set, the generation of a symbolic-regression model based on the SISSO algorithm requires about 24 nodes (instead of 3 nodes), each with two Intel Xeon E5-2698 v3 processors (= 32 cores/node) without hyper-threading, in order to build a machine-learning model within one day.

#	**Dependence**	**Features**		**Performance**	
	measure	#	Relevant features	R^2	RMSE [mÅ]
126 samples	TCMI	6	H(A), L(A), r_p(A), r_s(A), c(A), Z(B)	0.75±0.11	91±15
	RFECV	7	IP(A), Z(A), r_s(A), EA(B), Z(B), r_p(B), r_s(B)	0.83±0.06	80±15
	FS-GBDT	6	Z(A), r_s(A), EA(B), Z(B), r_p(B), r_s(B)	0.86±0.09	64±16
	FS-SISSO	4	r_s(A), EA(B), Z(B), r_s(B)	0.88±0.05	64±13
	TB3-GBDT	14	**EA(A)**, EN(A), H(A), IP(A), L(A), **Z(A)**, r_p(A), r_s(A), c(A), H(B), IP(B), **Z(B)**, **r_p(B)**, r_s(B)	0.87±0.09/ **0.69±0.12**	68±22/ **108±22**
	TB3-SISSO	15	EN(A), H(A), IP(A), **Z(A)**, r_p(A), **r_s(A)**, c(A), **EA(B)**, EN(B), H(B), **IP(B)**, **L(B)**, **Z(B)**, r_p(B), r_s(B)	0.88±0.05/ **0.80±0.15**	64±13/ **83±40**
252 samples	TCMI	6	EA(A), IP(A), L(A), r_s(A), Z(B), r_s(B)	0.81±0.10	75±9
	RFECV	6	Z(A), r_s(A), EA(B), Z(B), r_p(B), r_s(B)	0.83±0.09	72±6
	FS-GBDT	7	Z(A), r_s(A), EA(B), L(B), Z(B), r_p(B), r_s(B)	0.83±0.06	74±7
	FS-SISSO	10	Z(A), c(A), EA(B), EN(B), H(B), IP(B), L(B), Z(B), r_p(B), r_s(B)	0.88±0.05	62±10
	TB3-GBDT	15	EA(A), EN(A), H(A), IP(A), L(A), **Z(A)**, r_p(A), r_s(A), **c(A)**, EA(B), H(B), IP(B), **Z(B)**, r_p(B), r_s(B)	0.87±0.08/ **0.71±0.14**	61±8/ **95±18**
	TB3-SISSO	17	EA(A), EN(A), **H(A)**, IP(A), L(A), Z(A), r_p(A), **r_s(A)**, c(A), **EA(B)**, **EN(B)**, H(B), **IP(B)**, **L(B)**, **Z(B)**, **r_p(B)**, r_s(B)	0.87±0.08/ **0.87±0.06**	61±10/ **64±11**
504 samples	TCMI	6	H(A), Z(A), r_s(A), Z(B), r_p(B), r_s(B)	0.86±0.03	66±5
	RFECV	8	IP(A), Z(A), r_s(A), EA(B), L(B), Z(B), r_p(B), r_s(B)	0.87±0.03	64±5
	FS-GBDT	9	L(A), Z(A), r_s(A), EA(B), H(B), L(B), Z(B), r_p(B), r_s(B)	0.84±0.03	70±4
	FS-SISSO	12	H(A), Z(A), r_s(A), c(A), EA(B), EN(B), H(B), IP(B), L(B), Z(B), r_p(B), r_s(B)	0.88±0.03	61±6
	TB3-GBDT	16	EA(A), EN(A), H(A), IP(A), L(A), **Z(A)**, r_p(A), r_s(A), **c(A)**, EA(B), H(B), IP(B), L(B), **Z(B)**, r_p(B), r_s(B)	0.88±0.03/ **0.72±0.06**	60±8/ **93±7**
	TB3-SISSO	13	H(A), IP(A), Z(A), **r_s(A)**, **c(A)**, **EA(B)**, **EN(B)**, **H(B)**, IP(B), L(B), **Z(B)**, r_p(B), r_s(B)	0.88±0.03/ **0.86±0.02**	60±8/ **67±4**
Lattice constant		Stats: $a_0 = [3.57, 4.40]$ Å		mean = 3.91 Å	std = 0.18 Å

Tab. 5.5. Prediction performance of identified redundant feature subsets (normal face) and optimal non-redundant feature subsets (bold face) of different feature-identification methods for estimating the equilibrium lattice constant a_0: total cumulative mutual information (TCMI, Section 3.3.3), recursive feature elimination [263] using random forest (RFECV) [266], gradient-boosting decision trees using permutation feature importance (FS-GBDT) [252, 266, 338], SISSO using 10-fold cross-validation and hyper-parameter optimization (FS-SISSO) [197] (cf., Section 2.5), and the tolerance-based branch-and-bound algorithm (TB3) with GBDT and SISSO as feature-selection criterion (Section 4.1). Prediction performances were estimated using SISSO by means of 10-fold cross-validation (cf., Section 5.1). Shown are the mean $\bar{x}$ and standard deviation Δx as $\bar{x} \pm \Delta x$ of the Pearson's coefficient of determination $x = R^2$ [101] and the root-mean-squared error (x = RMSE). The RMSE is in the unit of milli angstrom (mÅ).

#	Dependence measure	Features #	Relevant features	Performance R^2	RMSE [GPa]
126 samples	TCMI	8	EN(A), L(A), r_p(A), EA(B), EN(B), IP(B), Z(B), r_p(B)	0.47±0.20	32.7±7.4
	RFECV	7	IP(A), L(A), c(A), EA(B), Z(B), r_p(B), r_s(B)	0.60±0.22	26.7±6.8
	FS-GBDT	3	EA(B), Z(B), r_s(B)	0.55±0.17	28.4±7.2
	FS-SISSO	6	r_s(A), c(A), EA(B), Z(B), r_p(B), r_s(B)	0.57±0.21	31.3±7.4
	TB3-GBDT	14	EN(A), H(A), IP(A), Z(A), r_p(A), **r_s(A)**, **c(A)**, EA(B), **EN(B)**, **IP(B)**, L(B), **Z(B)**, r_p(B), r_s(B)	0.65±0.19/ **0.64±0.24**	26.4±4.6/ **24.9±5.1**
	TB3-SISSO	17	EA(A), EN(A), H(A), **IP(A)**, L(A), Z(A), **r_p(A)**, r_s(A), **c(A)**, **EA(B)**, EN(B), **H(B)**, IP(B), L(B), **Z(B)**, r_p(B), **r_s(B)**	0.63±0.19/ **0.64±0.22**	26.9±4.3/ **28.0±7.7**
252 samples	TCMI	7	H(A), L(A), Z(A), r_p(A), r_s(A), H(B), Z(B)	0.46±0.18	29.7±3.8
	RFECV	6	IP(A), c(A), EA(B), Z(B), r_p(B), r_s(B)	0.68±0.09	23.2±3.8
	FS-GBDT	8	IP(A), r_s(A), c(A), EA(B), IP(B), Z(B), r_p(B), r_s(B)	0.70±0.10	22.3±3.0
	FS-SISSO	9	r_s(A), c(A), EA(B), EN(B), H(B), IP(B), Z(B), r_p(B), r_s(B)	0.69±0.10	22.4±3.1
	TB3-GBDT	16	**EA(A)**, EN(A), H(A), IP(A), Z(A), r_p(A), r_s(A), **c(A)**, EA(B), EN(B), H(B), **IP(B)**, **L(B)**, Z(B), **r_p(B)**, r_s(B)	0.72±0.12/ **0.55±0.08**	21.7±4.6/ **27.5±3.4**
	TB3-SISSO	17	EA(A), EN(A), H(A), IP(A), L(A), Z(A), r_p(A), r_s(A), **c(A)**, **EA(B)**, EN(B), H(B), IP(B), **L(B)**, **Z(B)**, r_p(B), r_s(B)	0.74±0.13/ **0.59±0.10**	21.2±4.3/ **26.0±2.9**
504 samples	TCMI	8	H(A), L(A), r_p(A), r_s(A), EA(B), EN(B), H(B), IP(B)	0.58±0.11	26.6±2.8
	RFECV	6	IP(A), c(A), L(B), Z(B), r_p(B), r_s(B)	0.69±0.03	22.8±2.4
	FS-GBDT	13	H(A), IP(A), L(A), r_p(A), r_s(A), c(A), EA(B), H(B), IP(B), L(B), Z(B), r_p(B), r_s(B)	0.73±0.05	21.4±1.6
	FS-SISSO	10	IP(A), r_s(A), c(A), EA(B), EN(B), H(B), L(B), Z(B), r_p(B), r_s(B)	0.75±0.05	20.3±2.0
	TB3-GBDT	17	EA(A), EN(A), H(A), IP(A), L(A), **Z(A)**, r_p(A), r_s(A), **c(A)**, **EA(B)**, EN(B), H(B), IP(B), L(B), Z(B), **r_p(B)**, r_s(B)	0.75±0.04/ **0.64±0.05**	20.4±2.1/ **24.9±2.2**
	TB3-SISSO	14	**H(A)**, **IP(A)**, L(A), Z(A), r_s(A), c(A), **EA(B)**, **EN(B)**, H(B), **IP(B)**, **L(B)**, **Z(B)**, r_p(B), **r_s(B)**	0.75±0.04/ **0.74±0.05**	20.4±2.1/ **21.1±2.5**
Bulk modulus		Stats: B_0 = [59.9 , 238.1] GPa		mean = 175.1 GPa	std = 41.0 GPa

Tab. 5.6. Prediction performance of identified redundant feature subsets (normal face) and optimal non-redundant feature subsets (bold face) of different feature-identification methods for estimating the bulk modulus B_0: total cumulative mutual information (TCMI, Section 3.3.3), recursive feature elimination [263] using random forest (RFECV) [266], gradient-boosting decision trees using permutation feature importance (FS-GBDT) [252, 266, 338], SISSO using 10-fold cross-validation and hyper-parameter optimization (FS-SISSO) [197] (cf., Section 2.5), and the tolerance-based branch-and-bound algorithm (TB3) with GBDT and SISSO as feature-selection criterion (Section 4.1). Prediction performances were estimated using SISSO by means of 10-fold cross-validation (cf., Section 5.1). Shown are the mean $\bar{x}$ and standard deviation Δx as $\bar{x} \pm \Delta x$ of the Pearson's coefficient of determination $x = R^2$ [101] and the root-mean-squared error (x = RMSE). The RMSE is in the unit of gigapascal (1 GPa = 6.24×10^{-3} eV/Å^3).

Similar to the octet-binary compound semiconductor data set, there is a clear trend towards lower prediction errors with increasing numbers of data samples and larger feature-subset cardinalities. While all symbolic-regression models of the feature-identification methods (TCMI, RFECV, FS-GBDT, FS-SISSO, TB3-GBDT, and TB3-SISSO) lead to similar prediction errors for the equilibrium lattice constant, prediction errors of symbolic-regression models with TCMI are consistently 30% larger in the case of the bulk modulus. A detailed analysis reveals that the poor prediction performance is due to inherent issues of TCMI with discrete features (cf., Section 3.3.3), which affect more strongly the prediction of the bulk modulus than the prediction of the equilibrium lattice constant (cf., Fig. 5.9).

Since the perovskite structure is further restricted to the cubic crystal structure, the chemical formula uniquely determines the materials properties of the perovskite compounds as a non-linear function of the three (atomic) features: the nuclear charges Z of the cations A/B and the ionic charge c(A) [410]. However, feature-identification methods show a high variability in the identification of relevant features, indicating multiple statistical relationships between the features and the property of interests. For example, FS-GBDT identifies about 3–13, RFECV and TCMI about 6–8, and FS-SISSO identifies about 4–12 features as relevant. The variability is also evident in the optimal and redundant feature subsets of the developed tolerance-based branch-and-bound algorithm (TB3): whereas minimal non-redundant feature subsets of TB3-GBDT or TB3-SISSO consist of 3–8 relevant features, redundant feature subsets of the TB3-algorithm include almost all features of the data set. Although the nuclear and ionic charges are not identified as relevant in all feature-identification methods, TB3-GBDT and TB3-SISSO identify them as one of the sub-optimal minimally non-redundant feature subsets, which are also present in feature subsets of larger feature-subset cardinalities.

A comparison of all methods shows: In terms of prediction performance and feature-subset cardinality (cf., Section 3.5), RFECV, FS-GBDT, and TB3-SISSO provide the smallest feature subsets, while TB3-SISSO and FS-SISSO result in the lowest prediction errors overall. Therefore, the prediction performance of the optimal feature subsets of TB3-SISSO are not only comparable to that of FS-SISSO, but also offer a much more compact materials representation for the prediction of the bulk modulus B_0 and the lattice constant a_0.

A frequency analysis (Fig. 5.9) shows: the most frequently identified relevant features for predicting the lattice constant a_0 are r_s(A), r_s(B), r_p(B), Z(A), Z(B), and EA(B), while the most frequently identified relevant features for the prediction of the bulk modulus B_0 are r_s(B), Z(B), EA(B), and c(A). From a physical point of view, the equilibrium lattice constant implicitly depends on the atomic radii, nuclear charge, and electron affinity: the atomic radii r_s and r_p describe the size of atoms and as such are related to the distance between atoms, the nuclear charges are strongly correlated to the size of the atoms, and the electron affinity determines the strength and type of bonding. Further, the nuclear and ionic charges are correlated to the chemical hardness of a material and hence to the elastic properties of the crystals and the bulk modulus. The bulk modulus, in turn, is known to have an explicit dependence on the equilibrium lattice constant as well as on the ionic charge of the perovskite oxides [410].

Tests confirm the dependence of the bulk modulus on the ionic charge and the equilibrium lattice constant (a_0). For example, an exclusion of the ionic charge in the model construction resulted in about 10% higher prediction errors, while the lattice constant as an additional feature of the data set had the effect that all investigated feature-identification methods identified a_0 as relevant in

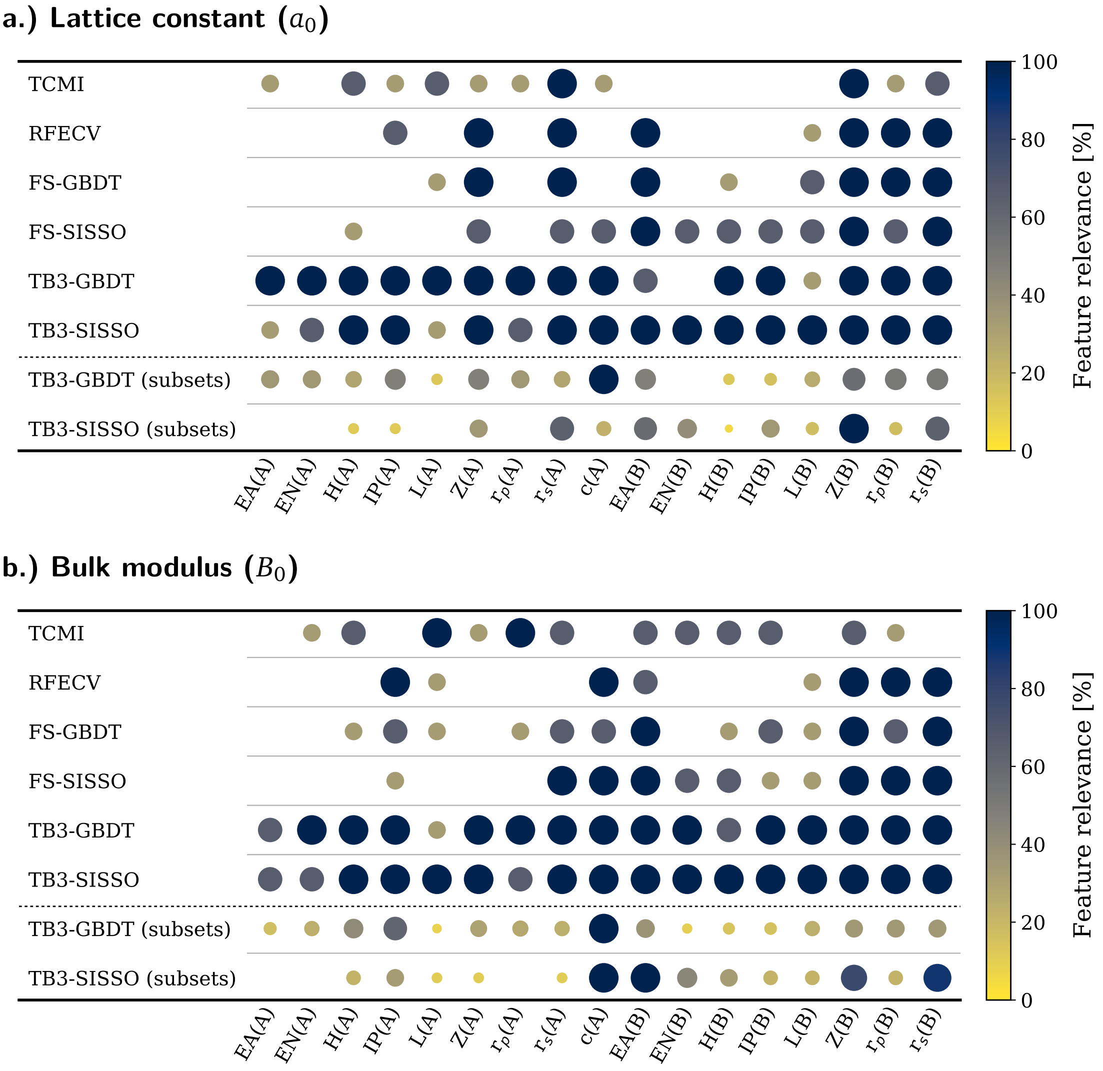


Fig. 5.9. Heat-map of most frequent identified (redundant) feature subsets of the perovskite oxides data set across all investigated feature-identification methods (cf., Section 5.1). The size and color reflect the frequency of the features (= relevance). A frequency analysis of identified minimally non-redundant feature subsets of the tolerance-based branch-and-bound algorithm (TB3) with GBDT and SISSO is shown at the bottom.

predicting the bulk modulus. The fact that the ionic charge and the most frequently identified features are consistently identified by the minimal non-redundant feature subsets of TB3 suggests that the developed framework among the other feature-identification methods have the potential to identify physically relevant features from a statistical analysis of the materials data.

Common to all feature-identification methods is the overall good prediction performance of the lattice constant ($R^2 \approx 0.87$, RMSE $\approx 60\,\text{mÅ}$), but only moderate prediction performance of the bulk modulus ($R^2 \leq 0.75$, RMSE $\geq 20\,\text{GPa}$).

The moderate prediction performance of the bulk modulus can be attributed to the multiple non-linear statistical relationships present in the data and the chosen model settings for estimating the

Performance metric	**Machine-learning algorithm**		
	GBDT	SISSO	SISSO$\|_{c(A)}$
a.) Atomic features only			
R^2	0.94 ± 0.02	0.75 ± 0.04	0.81 ± 0.08
RMSE [GPa]	9.8 ± 1.7	20.6 ± 1.3	16.0 ± 2.6
b.) Atomic features + estimated lattice constant ($\hat{a}_0$)			
R^2	0.93 ± 0.02	0.85 ± 0.04	0.87 ± 0.06
RMSE [GPa]	10.9 ± 1.7	15.7 ± 1.9	13.2 ± 4.0
c.) Atomic features + lattice constant (a_0)			
R^2	0.93 ± 0.02	0.84 ± 0.03	0.90 ± 0.05
RMSE [GPa]	10.9 ± 1.8	16.5 ± 1.5	11.7 ± 3.2

Tab. 5.7. Comparison of different machine-learning algorithms and techniques for the quantitative prediction of the perovskite's bulk modulus B_0, namely the gradient-boosting decision trees algorithm (GBDT) [249–252, 267, 268], the sure-independence screening and sparsifying operator algorithm (SISSO) [197], and the construction of independent machine-learning models for each value of the ionic charge with SISSO (SISSO$|_{c(A)}$). Shown are the prediction performances in terms of the Pearson's coefficient of determination R^2 [101] and the root-mean-squared error (RMSE) of 10-fold cross-validated machine-learning models using (a) the atomic features only, (b) the atomic features and an estimation of the lattice constant with the GBDT algorithm, or (c) atomic features including the equilibrium lattice constant from DFT. The RMSE is in units of gigapascal (1 GPa = 6.24×10^{-3} eV/Å^3).

final prediction performance of the different feature-identification methods. For example, a machine-learning model of the bulk modulus can be constructed with the same prediction performance as the equilibrium lattice constant (Tab. 5.7) without requiring additional features. This is done by either using the computed lattice constants from DFT ($\hat{a}$) or an estimation of the equilibrium lattice constants ($\hat{a}_0$) based on the 16 atomic features (cf., Eqs. 5.2 and 5.3). In addition, the prediction error of the bulk modulus can be further reduced by creating independent machine-learning models for each oxidation state as given by the ionic charge (SISSO$|_{c(A)}$)[18], by increasing the complexity of the symbolic-regression model (i.e., by increasing the number of terms, the number of operators or feature combinations in the symbolic expression), or by recursively applying SISSO [411] on the same set of features at the cost of much higher computational requirements.

The moderate prediction performance can be further attributed to the combination of a GBDT machine-learning algorithm used for the feature identification with a SISSO algorithm used for model construction. For instance, as for the TB3-algorithm, the prediction error of the optimal minimally non-redundant feature subsets of TB3-GBDT are larger than that of TB3-SISSO and larger than that of their redundant feature subsets. First, because there is a larger variance in the prediction performance of minimally non-redundant and redundant feature subsets of TB3-GBDT than for TB3-SISSO. And

[18] SISSO creates very different symbolic-regression models for each ionic charge. A multi-task approach by simultaneously modeling a symbolic-regression model (with the same symbolic expressions, but different coefficients [176]) for all oxidation states therefore results in higher prediction errors (atomic features of the bulk modulus (cf., Tab. 5.7): R^2 = 0.59 ± 0.09, RMSE = 25.9 ± 2.2, atomic features + $\hat{a}_0$: R^2 = 0.74 ± 0.06, RMSE = 20.6 ± 1.9, atomic features + a_0: R^2 = 0.76 ± 0.06, RMSE = 20.0 ± 1.3) and hence was further not considered in the analysis.

second, because the feature subsets of TB3-GBDT have a smaller feature-subset cardinality than feature subsets of TB3-SISSO. That is, the construction of a symbolic-regression model requires a higher model complexity as the construction of a GBDT machine-learning model. This can be exemplified by the feature subset {Z(A), Z(B), c(A)}. By using the GBDT machine-learning algorithm, the feature subset {Z(A), Z(B), c(A)} is found to be among the best-performing feature-subsets, which are statistically equivalent to a GBDT model constructed on the full set of features,

$$\begin{aligned}
&\textit{GBDT} \\
&\qquad \textbf{lattice constant } (a_0): \qquad R^2 = 0.99 \pm 0.00 \qquad \mathrm{RMSE} = (14 \pm 2)\ \mathrm{m\AA}\,, \\
&\qquad \textbf{bulk modulus } (B_0): \qquad R^2 = 0.94 \pm 0.02 \qquad \mathrm{RMSE} = (9.5 \pm 1.8)\ \mathrm{GPa}\,.
\end{aligned} \tag{5.2}$$

The same feature subset however is not among the best-performing feature subsets, when using a SISSO model (cf., Tabs. 5.5 and 5.6),

$$\begin{aligned}
&\textit{SISSO} \\
&\qquad \textbf{lattice constant } (a_0): \qquad R^2 = 0.75 \pm 0.04 \qquad \mathrm{RMSE} = (89 \pm 4)\ \mathrm{m\AA}\,, \\
&\qquad \textbf{bulk modulus } (B_0): \qquad R^2 = 0.47 \pm 0.07 \qquad \mathrm{RMSE} = (30.0 \pm 3.2)\ \mathrm{GPa}\,.
\end{aligned} \tag{5.3}$$

As such, an identification of all relevant features with TB3-GBDT is hampered by the fact that a piecewise-constant machine-learning algorithm performs better on the perovskite data set as a symbolic-regression model: the TB3-GBDT algorithm identifies subsets of smaller feature-subset cardinality and hence not all features required for a highly predictive symbolic-regression model with SISSO (Figs. 5.10 and 5.11).

Feature dependences

Feature-dependence maps generated with TB3-GBDT and TB3-SISSO reveal a block-like structure of feature interactions between features of the same atomic species (Fig. 5.12). Whereas feature interactions of the TB3-GBDT algorithm indicate a one-to-one correspondence between features (i.e., there is an exact statistical relationship between the features such that each feature can be used to accurately predict all the other features of the same atomic species), not all features of the same atomic species are statistically related by TB3-SISSO. This is due to the fact that feature subsets with a larger feature-subset cardinality than the optimal feature subset are not identified by the TB3-algorithm and hence are not considered as dependent at a confidence level of $\alpha = 0.95$ in the feature-dependence maps (Section 4.2).

The oxidation state of atomic species A can be exactly predicted with any (atomic) feature of A, but conversely, the oxidation state is not sufficient to estimate the diversity of atomic properties in each materials class of perovskites oxides with a GBDT model. Although there potentially exists a relationship between atomic properties of atomic species A and the ionic charge, it requires a much larger set of features as compared to the minimally non-redundant feature subset as found by the TB3-GBDT algorithm. It is therefore likely that the GBDT feature-dependence map simply does not capture this relationship, as other strong feature interactions prevent the identification of weaker statistical relationships in the data set as compared to the TB3-SISSO algorithm (Fig. 5.12).

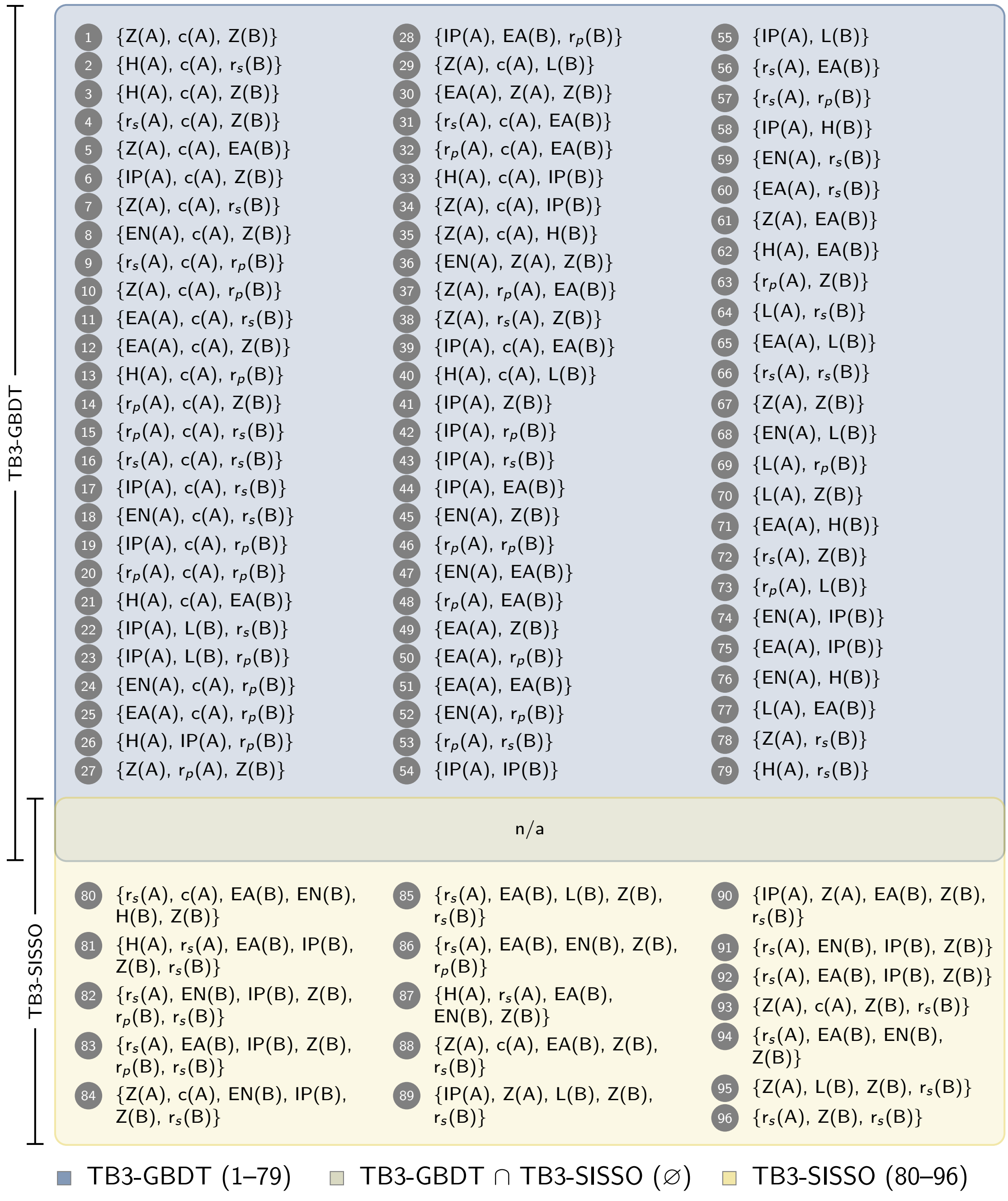
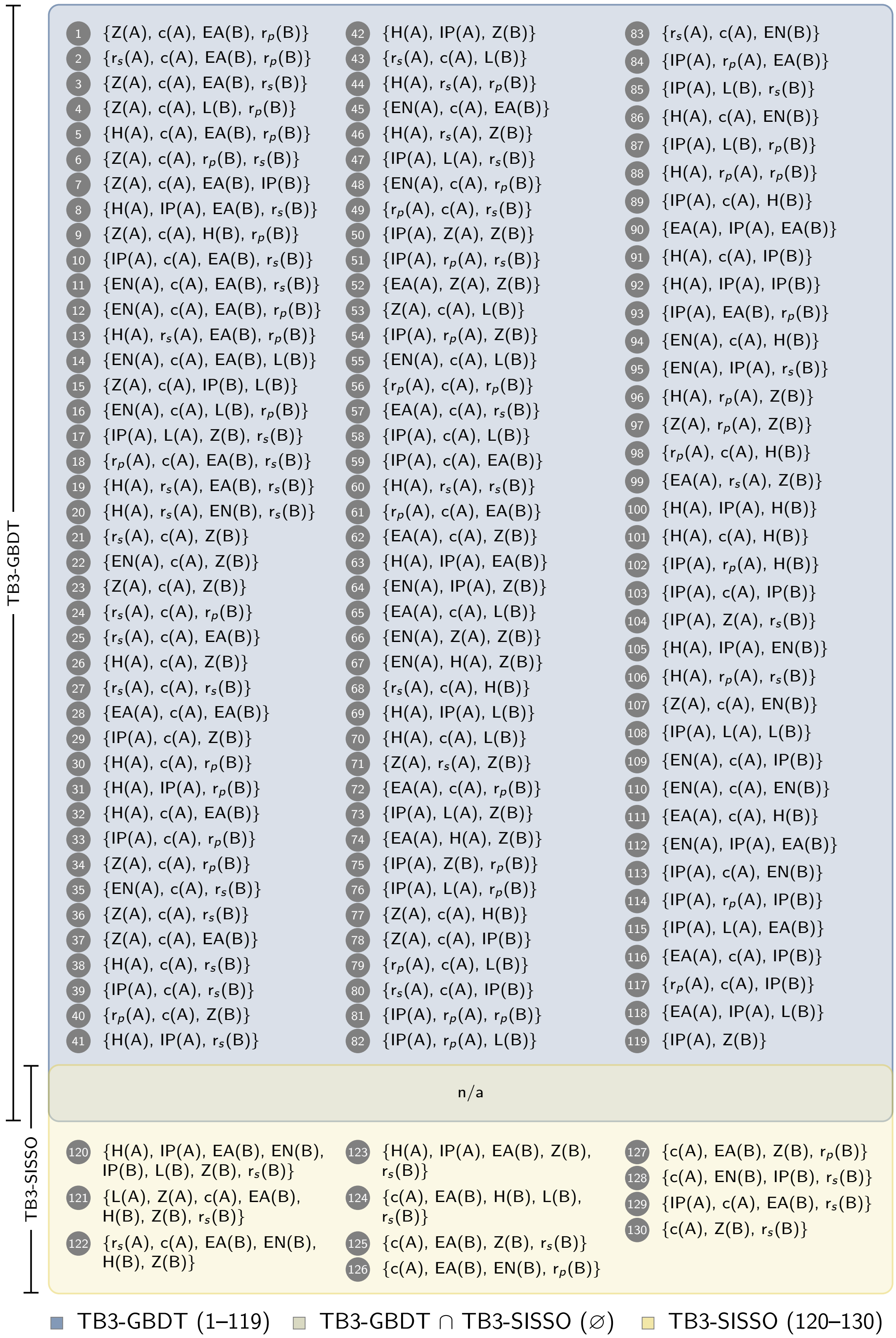

TB3-GBDT
TB3-SISSO
1 {Z(A), c(A), Z(B)}
2 {H(A), c(A), r_s(B)}
3 {H(A), c(A), Z(B)}
4 {r_s(A), c(A), Z(B)}
5 {Z(A), c(A), EA(B)}
6 {IP(A), c(A), Z(B)}
7 {Z(A), c(A), r_s(B)}
8 {EN(A), c(A), Z(B)}
9 {r_s(A), c(A), r_p(B)}
10 {Z(A), c(A), r_p(B)}
11 {EA(A), c(A), r_s(B)}
12 {EA(A), c(A), Z(B)}
13 {H(A), c(A), r_p(B)}
14 {r_p(A), c(A), Z(B)}
15 {r_p(A), c(A), r_s(B)}
16 {r_s(A), c(A), r_s(B)}
17 {IP(A), c(A), r_s(B)}
18 {EN(A), c(A), r_s(B)}
19 {IP(A), c(A), r_p(B)}
20 {r_p(A), c(A), r_p(B)}
21 {H(A), c(A), EA(B)}
22 {IP(A), L(B), r_s(B)}
23 {IP(A), L(B), r_p(B)}
24 {EN(A), c(A), r_p(B)}
25 {EA(A), c(A), r_p(B)}
26 {H(A), IP(A), r_p(B)}
27 {Z(A), r_p(A), Z(B)}
28 {IP(A), EA(B), r_p(B)}
29 {Z(A), c(A), L(B)}
30 {EA(A), Z(A), Z(B)}
31 {r_s(A), c(A), EA(B)}
32 {r_p(A), c(A), EA(B)}
33 {H(A), c(A), IP(B)}
34 {Z(A), c(A), IP(B)}
35 {Z(A), c(A), H(B)}
36 {EN(A), Z(A), Z(B)}
37 {Z(A), r_p(A), EA(B)}
38 {Z(A), r_s(A), Z(B)}
39 {IP(A), c(A), EA(B)}
40 {H(A), c(A), L(B)}
41 {IP(A), Z(B)}
42 {IP(A), r_p(B)}
43 {IP(A), r_s(B)}
44 {IP(A), EA(B)}
45 {EN(A), Z(B)}
46 {r_p(A), r_p(B)}
47 {EN(A), EA(B)}
48 {r_p(A), EA(B)}
49 {EA(A), Z(B)}
50 {EA(A), r_p(B)}
51 {EA(A), EA(B)}
52 {EN(A), r_p(B)}
53 {r_p(A), r_s(B)}
54 {IP(A), IP(B)}
55 {IP(A), L(B)}
56 {r_s(A), EA(B)}
57 {r_s(A), r_p(B)}
58 {IP(A), H(B)}
59 {EN(A), r_s(B)}
60 {EA(A), r_s(B)}
61 {Z(A), EA(B)}
62 {H(A), EA(B)}
63 {r_p(A), Z(B)}
64 {L(A), r_s(B)}
65 {EA(A), L(B)}
66 {r_s(A), r_s(B)}
67 {Z(A), Z(B)}
68 {EN(A), L(B)}
69 {L(A), r_p(B)}
70 {L(A), Z(B)}
71 {EA(A), H(B)}
72 {r_s(A), Z(B)}
73 {r_p(A), L(B)}
74 {EN(A), IP(B)}
75 {EA(A), IP(B)}
76 {EN(A), H(B)}
77 {L(A), EA(B)}
78 {Z(A), r_s(B)}
79 {H(A), r_s(B)}
n/a
80 {r_s(A), c(A), EA(B), EN(B), H(B), Z(B)}
81 {H(A), r_s(A), EA(B), IP(B), Z(B), r_s(B)}
82 {r_s(A), EN(B), IP(B), Z(B), r_p(B), r_s(B)}
83 {r_s(A), EA(B), IP(B), Z(B), r_p(B), r_s(B)}
84 {Z(A), c(A), EN(B), IP(B), Z(B), r_s(B)}
85 {r_s(A), EA(B), L(B), Z(B), r_s(B)}
86 {r_s(A), EA(B), EN(B), Z(B), r_p(B)}
87 {H(A), r_s(A), EA(B), EN(B), Z(B)}
88 {Z(A), c(A), EA(B), Z(B), r_s(B)}
89 {IP(A), Z(A), L(B), Z(B), r_s(B)}
90 {IP(A), Z(A), EA(B), Z(B), r_s(B)}
91 {r_s(A), EN(B), IP(B), Z(B)}
92 {r_s(A), EA(B), IP(B), Z(B)}
93 {Z(A), c(A), Z(B), r_s(B)}
94 {r_s(A), EA(B), EN(B), Z(B)}
95 {Z(A), L(B), Z(B), r_s(B)}
96 {r_s(A), Z(B), r_s(B)}
■ TB3-GBDT (1–79)
□ TB3-GBDT ∩ TB3-SISSO (∅)
□ TB3-SISSO (80–96)

Fig. 5.10. Identified minimally non-redundant (sub-optimal) feature subsets for the prediction of the equilibrium lattice constant a_0 using the tolerance-based branch-and-bound algorithm (TB3) and GBDT or SISSO as the feature-selection criterion (Section 4.1). Shown is the intersection and the union of the two feature-identification methods corresponding to a Jaccard similarity coefficient (Eq. 5.1) of zero. It should be noted that due to the different machine-learning algorithms, a direct comparison between TB3-GBDT and TB3-SISSO may not be appropriate. The numbering is used for referencing purposes.



TB3-GBDT

1 {Z(A), c(A), EA(B), r_p(B)}
2 {r_s(A), c(A), EA(B), r_p(B)}
3 {Z(A), c(A), EA(B), r_s(B)}
4 {Z(A), c(A), L(B), r_p(B)}
5 {H(A), c(A), EA(B), r_p(B)}
6 {Z(A), c(A), r_p(B), r_s(B)}
7 {Z(A), c(A), EA(B), IP(B)}
8 {H(A), IP(A), EA(B), r_s(B)}
9 {Z(A), c(A), H(B), r_p(B)}
10 {IP(A), c(A), EA(B), r_s(B)}
11 {EN(A), c(A), EA(B), r_s(B)}
12 {EN(A), c(A), EA(B), r_p(B)}
13 {H(A), r_s(A), EA(B), r_p(B)}
14 {EN(A), c(A), EA(B), L(B)}
15 {Z(A), c(A), IP(B), L(B)}
16 {EN(A), c(A), L(B), r_p(B)}
17 {IP(A), L(A), Z(B), r_s(B)}
18 {r_p(A), c(A), EA(B), r_s(B)}
19 {H(A), r_s(A), EA(B), r_s(B)}
20 {H(A), r_s(A), EN(B), r_s(B)}
21 {r_s(A), c(A), Z(B)}
22 {EN(A), c(A), Z(B)}
23 {Z(A), c(A), Z(B)}
24 {r_s(A), c(A), r_p(B)}
25 {r_s(A), c(A), EA(B)}
26 {H(A), c(A), Z(B)}
27 {r_s(A), c(A), r_s(B)}
28 {EA(A), c(A), EA(B)}
29 {IP(A), c(A), Z(B)}
30 {H(A), c(A), r_p(B)}
31 {H(A), IP(A), r_p(B)}
32 {H(A), c(A), EA(B)}
33 {IP(A), c(A), r_p(B)}
34 {Z(A), c(A), r_p(B)}
35 {EN(A), c(A), r_s(B)}
36 {Z(A), c(A), r_s(B)}
37 {Z(A), c(A), EA(B)}
38 {H(A), c(A), r_s(B)}
39 {IP(A), c(A), r_s(B)}
40 {r_p(A), c(A), Z(B)}
41 {H(A), IP(A), r_s(B)}
42 {H(A), IP(A), Z(B)}
43 {r_s(A), c(A), L(B)}
44 {H(A), r_s(A), r_p(B)}
45 {EN(A), c(A), EA(B)}
46 {H(A), r_s(A), Z(B)}
47 {IP(A), L(A), r_s(B)}
48 {EN(A), c(A), r_p(B)}
49 {r_p(A), c(A), r_s(B)}
50 {IP(A), Z(A), Z(B)}
51 {IP(A), r_p(A), r_s(B)}
52 {EA(A), Z(A), Z(B)}
53 {Z(A), c(A), L(B)}
54 {IP(A), r_p(A), Z(B)}
55 {EN(A), c(A), L(B)}
56 {r_p(A), c(A), r_p(B)}
57 {EA(A), c(A), r_s(B)}
58 {IP(A), c(A), L(B)}
59 {IP(A), c(A), EA(B)}
60 {H(A), r_s(A), r_s(B)}
61 {r_p(A), c(A), EA(B)}
62 {EA(A), c(A), Z(B)}
63 {H(A), IP(A), EA(B)}
64 {EN(A), IP(A), Z(B)}
65 {EA(A), c(A), L(B)}
66 {EN(A), Z(A), Z(B)}
67 {EN(A), H(A), Z(B)}
68 {r_s(A), c(A), H(B)}
69 {H(A), IP(A), L(B)}
70 {H(A), c(A), L(B)}
71 {Z(A), r_s(A), Z(B)}
72 {EA(A), c(A), r_p(B)}
73 {IP(A), L(A), Z(B)}
74 {EA(A), H(A), Z(B)}
75 {IP(A), Z(B), r_p(B)}
76 {IP(A), L(A), r_p(B)}
77 {Z(A), c(A), H(B)}
78 {Z(A), c(A), IP(B)}
79 {r_p(A), c(A), L(B)}
80 {r_s(A), c(A), IP(B)}
81 {IP(A), r_p(A), r_p(B)}
82 {IP(A), r_p(A), L(B)}
83 {r_s(A), c(A), EN(B)}
84 {IP(A), r_p(A), EA(B)}
85 {IP(A), L(B), r_s(B)}
86 {H(A), c(A), EN(B)}
87 {IP(A), L(B), r_p(B)}
88 {H(A), r_p(A), r_p(B)}
89 {IP(A), c(A), H(B)}
90 {EA(A), IP(A), EA(B)}
91 {H(A), c(A), IP(B)}
92 {H(A), IP(A), IP(B)}
93 {IP(A), EA(B), r_p(B)}
94 {EN(A), c(A), H(B)}
95 {EN(A), IP(A), r_s(B)}
96 {H(A), r_p(A), Z(B)}
97 {Z(A), r_p(A), Z(B)}
98 {r_p(A), c(A), H(B)}
99 {EA(A), r_s(A), Z(B)}
100 {H(A), IP(A), H(B)}
101 {H(A), c(A), H(B)}
102 {IP(A), r_p(A), H(B)}
103 {IP(A), c(A), IP(B)}
104 {IP(A), Z(A), r_s(B)}
105 {H(A), IP(A), EN(B)}
106 {H(A), r_p(A), r_s(B)}
107 {Z(A), c(A), EN(B)}
108 {IP(A), L(A), L(B)}
109 {EN(A), c(A), IP(B)}
110 {EN(A), c(A), EN(B)}
111 {EA(A), c(A), H(B)}
112 {EN(A), IP(A), EA(B)}
113 {IP(A), c(A), EN(B)}
114 {IP(A), r_p(A), IP(B)}
115 {IP(A), L(A), EA(B)}
116 {EA(A), c(A), IP(B)}
117 {r_p(A), c(A), IP(B)}
118 {EA(A), IP(A), L(B)}
119 {IP(A), Z(B)}

n/a

TB3-SISSO

120 {H(A), IP(A), EA(B), EN(B), IP(B), L(B), Z(B), r_s(B)}
121 {L(A), Z(A), c(A), EA(B), H(B), Z(B), r_s(B)}
122 {r_s(A), c(A), EA(B), EN(B), H(B), Z(B)}
123 {H(A), IP(A), EA(B), Z(B), r_s(B)}
124 {c(A), EA(B), H(B), L(B), r_s(B)}
125 {c(A), EA(B), Z(B), r_s(B)}
126 {c(A), EA(B), EN(B), r_p(B)}
127 {c(A), EA(B), Z(B), r_p(B)}
128 {c(A), EN(B), IP(B), r_s(B)}
129 {IP(A), c(A), EA(B), r_s(B)}
130 {c(A), Z(B), r_s(B)}

■ TB3-GBDT (1–119) □ TB3-GBDT ∩ TB3-SISSO (∅) □ TB3-SISSO (120–130)

Fig. 5.11. Identified minimally non-redundant (sub-optimal) feature subsets for the prediction of the bulk modulus B_0 using the tolerance-based branch-and-bound algorithm (TB3) and GBDT or SISSO as the feature-selection criterion (Section 4.1). Shown is the intersection and the union of the two feature-identification methods corresponding to a Jaccard similarity coefficient (Eq. 5.1) of zero. It should be noted that due to the different machine-learning algorithms, a direct comparison between TB3-GBDT and TB3-SISSO may not be appropriate. The numbering is used for referencing purposes.

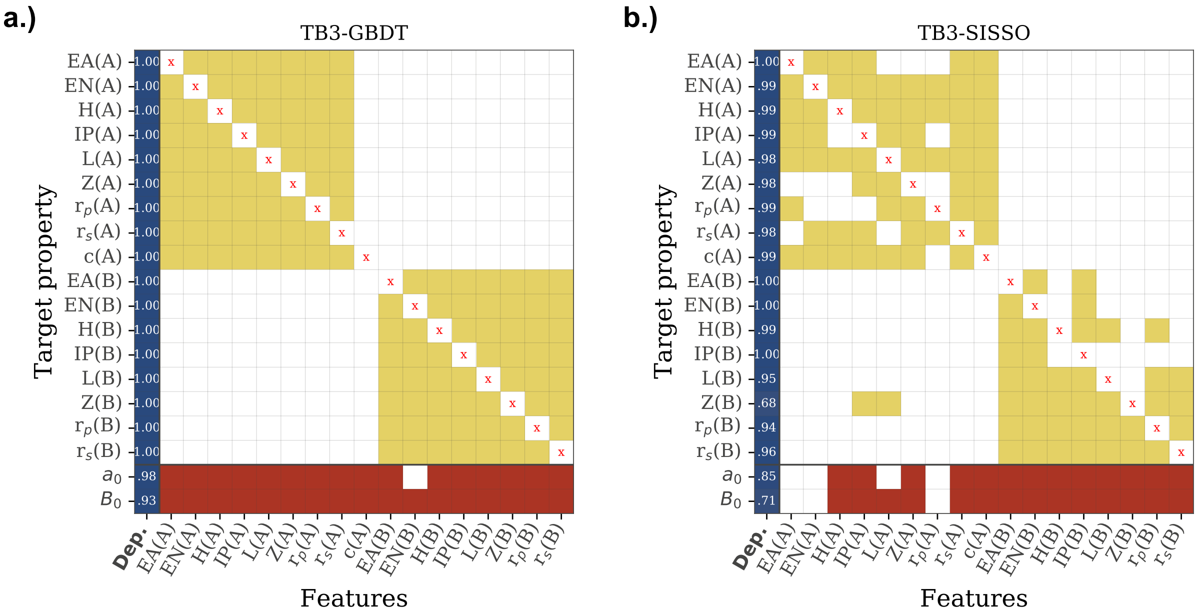


Fig. 5.12. Feature-dependence maps of the perovskite oxides data set created with the tolerance-based branch-and-bound algorithm (TB3) and two different machine-learning algorithms: (a) gradient-boosting decision trees (GBDT) [249–252, 267, 268] and (b) the sure-independence screening and sparsifying operator (SISSO) [197]. Feature-dependence maps were created at a confidence level of $\alpha = 0.95$ and a convergence threshold of $\epsilon = 0.01$ (cf., Section 4.2). The score (=strength) of the dependence is shown in the first column (Dep.). The score is the Pearson's R coefficient of determination [101] of a 10-fold cross-validated machine-learning model and by using the identified features of the TB3-algorithm (colored boxes). The dependence of the equilibrium lattice constant (a_0) and bulk modulus (B_0) are given in the last two rows.

Both the bulk modulus (B_0) and the nuclear charge Z(B) are characterized by relatively low dependence scores. In the case of the nuclear charge Z(B), for example, the feature-dependence map of the TB3-SISSO algorithm identifies a statistical dependency with the ionization potential (IP) and the lowest-unoccupied Kohn-Sham level (L) of atomic species A. Since there is no physically known relationship between the atomic properties of two atomic species and the features of the atomic species B are not strongly related to the other features of the atomic species A, it can be assumed that the low dependence scores are due to an incorrect identification of dependent features at a confidence level of $\alpha = 0.95$ (Section 4.2).

The block-like structure of feature interactions shows that all features of each atomic species can be expressed as an implicit dependence on the nuclear charges Z. Hence, a machine-learning model for predicting the equilibrium lattice constant a_0 or the bulk modulus B_0 requires at least one feature from each atomic species A/B. Most of these features are related by physical and statistical relationships as like in the octet-binary compound semiconductor data set (Section 5.2.1). For example, there is a strong statistical dependence between the atomic radii r$_s$ and r$_p$, the electron affinity (EA) is physically related to the electronegativity (EN) and the ionization potential (IP), and the electron affinity can also be expressed as a function of the ionization potential (IP) and the highest-occupied Kohn-Sham level (H).

#	**Features**	**GBDT**		**SISSO**	
	+ Model	R^2	RMSE [mÅ]	R^2	RMSE [mÅ]
1	$\{\mathbf{r}_s\mathbf{(A)}, \mathbf{Z(B)}\}$ $a_0 = 3.66\cdot\sqrt[3]{\sqrt[3]{Z(B)}}\exp\left(\frac{r_s(A)}{Z(B)}\right) + 0.183\cdot\left\lvert\log(Z(B)^3) - \left(Z(B)+\sqrt[3]{Z(B)}\right)\right\rvert + 0.93\cdot\left(\sqrt[3]{Z(B)}\sqrt{Z(B)} - \log(Z(B))\sqrt{Z(B)}\right) - 1.51$	0.96±0.02	32±7	0.76±0.05	86±4
2	$\{\mathbf{Z(A)}, \mathbf{r}_s\mathbf{(A)}, \mathbf{Z(B)}\}$ $a_0 = 0.139\cdot\left(\frac{\exp(r_s(A))}{\sqrt{Z(B)}} - \left(Z(A)^{-1} - \sqrt{Z(B)}\right)\right) + 0.0695\cdot\frac{\sqrt[3]{\exp\{Z(B)\}}}{Z(B)^2\cdot Z(B)^3} - 5.74\cdot\frac{\sqrt[3]{\exp\{Z(B)\}}}{(Z(B)^2)^3} + 2.89$	0.98±0.01	21±4	0.75±0.04	88±6
3	$\{\mathbf{r}_s\mathbf{(A)}, \mathbf{c(A)}, \mathbf{Z(B)}\}$ $a_0 = 2.46\text{e-}05\cdot\frac{\sqrt[3]{\exp\{Z(B)\}}}{(Z(B)^2)^2} + 0.0115\cdot\left(\frac{Z(B)^2}{\sqrt[3]{Z(B)}} - \frac{Z(B)^2}{\log(Z(B))}\right) + 0.0083\cdot\frac{r_s(A)^3\cdot r_s(A)}{c(A)}\left\lvert c(A) - \lvert c(A) - r_s(A)\rvert\right\rvert + 3.47$	0.99±0.00	15±4	0.75±0.04	88±7
4	$\{\mathbf{IP(A)}, \mathbf{c(A)}, \mathbf{Z(B)}\}$ $a_0 = 0.0988\cdot\frac{\sqrt{\exp(Z(B))}}{(Z(B)^3)^3} + 40.2\cdot\frac{\exp(-c(A))}{\sqrt{IP(A)}\exp(IP(A)/c(A))} + 0.0116\cdot\left(\frac{Z(B)^2}{\sqrt{Z(B)}} - \frac{Z(B)^2}{\log(Z(B))}\right) + 3.42$	0.99±0.00	17±4	0.74±0.05	89±8
5	$\{\mathbf{IP(A)}, \mathbf{Z(B)}\}$ $a_0 = -0.406\cdot\frac{Z(B)^2\sqrt{IP(A)}}{\exp\left(\sqrt{Z(B)}\right)} + 47.3\cdot\left(\frac{\sqrt{IP(A)}}{\sqrt{Z(B)}} - \frac{\sqrt{IP(A)}}{\log(Z(B))}\right) + 0.0771\cdot\frac{\sqrt{\exp(Z(B))}}{(Z(B)^3)^3} + 4.3$	0.98±0.01	21±4	0.73±0.05	91±9
6	$\{\mathbf{Z(A)}, \mathbf{c(A)}, \mathbf{Z(B)}\}$ $a_0 = -0.249\cdot\left(\left(Z(A)^{-1} - \sqrt{Z(B)}\right) - \frac{Z(A)}{Z(B)}\exp(-c(A))\right) + 0.048\cdot\frac{\sqrt{\exp(Z(B))}}{(Z(B)^3)^3} - 8.07\text{e-}06\cdot\left(\left(Z(A)+Z(B)\right)^2 - \frac{Z(A)^3}{\lvert Z(A)-Z(B)\rvert}\right) + 3.03$	0.99±0.00	14±3	0.72±0.06	93±7

Tab. 5.8. Ranked list of statistical equivalent symbolic-regression models for the prediction of the equilibrium lattice constant a_0 in ascending order of the SISSO [197] prediction errors (RMSE - ΔRMSE). Shown is the prediction performance of identified, best-performing, non-redundant, sub-optimal feature subsets of the tolerance-based branch-and-bound algorithm (TB3) using the gradient-boosting decision trees (GBDT) algorithm [249–252, 267, 268] as feature-selection criterion Sections 3.1.2 and 4.1). For comparison, 10-fold cross-validated (cf., Section 5.1) prediction performance of SISSO and the gradient-boosting decision trees (GBDT) algorithm [249–252, 267, 268] are reported in terms of the mean $\bar{x}$ and standard deviation Δx as $\bar{x} \pm \Delta x$ of the Pearson's coefficient of determination $x = R^2$ [101] and the root-mean-squared error (x = RMSE). The RMSE is in the unit of milli angstrom (mÅ).

#	**Features**	**GBDT**		**SISSO**	
	+ Model	R^2	RMSE [mÅ]	R^2	RMSE [mÅ]
1	{**H(A), r$_s$(A), EA(B), IP(B), Z(B), r$_s$(B)**} $a_0 = 1.82 \cdot \frac{r_s(A)}{Z(B)} \frac{EA(B) - IP(B)}{\exp(r_s(A))} \sqrt[3]{Z(B)} + 0.0209 \cdot \left\lvert \frac{Z(B)EA(B)}{IP(B)} - \left(\exp(H(A)) - \exp(r_s(B))\right) \right\rvert - 0.272 \cdot \left(\left\lvert \frac{\exp(-IP(B))}{\log(r_s(B))} \right\rvert - \frac{EA(B)}{H(A)} \log(r_s(B)) \right) + 4.25$	0.99±0.01	20±4	0.87±0.03	62±6
2	{**r$_s$(A), EN(B), IP(B), Z(B), r$_p$(B), r$_s$(B)**} $a_0 = 0.291 \cdot \left(\frac{\exp(r_p(B))}{Z(B)} + \sqrt[3]{Z(B)} + \frac{r_s(A)}{r_p(B)} \right) + 0.306 \cdot \left\lvert \sqrt{\exp(r_s(B))} - \left(\frac{IP(B)}{EN(B)} - \frac{EN(B)}{IP(B)} \right) \right\rvert - 0.00453 \cdot \lvert \log(r_s(B))Z(B)EN(B) - \exp(r_s(B))(EN(B) + IP(B)) \rvert + 2.55$	0.98±0.01	26±6	0.87±0.03	63±4
3	{**r$_s$(A), EA(B), EN(B), Z(B)**} $a_0 = 3.22 \cdot \frac{r_s(A)}{Z(B)} \frac{EA(B) - EN(B)}{\exp(r_s(A)) \log(Z(B))} - 1.26 \cdot \left\lvert \sqrt[3]{\log(EN(B))} - \frac{\log(Z(B))}{\sqrt[3]{Z(B)}} \right\rvert + 0.0123 \cdot \left\lvert \frac{Z(B)EA(B)}{\log(Z(B))} - \left((EA(B) - EN(B)) - (EN(B) - EA(B))\right) \right\rvert + 4.27$	0.97±0.01	27±6	0.87±0.03	63±4
4	{**r$_s$(A), EN(B), IP(B), Z(B)**} $a_0 = 3.01 \cdot \frac{r_s(A)}{Z(B)} \frac{EN(B) - IP(B)}{\exp(r_s(A))/\log(Z(B))} + 0.0146 \cdot \left\lvert (Z(B) - \sqrt{Z(B)}) - \frac{Z(B)EN(B)}{IP(B) - EN(B)} \right\rvert - 1.23 \cdot \left\lvert \sqrt[3]{\log(EN(B))} - \frac{\log(Z(B))}{\sqrt[3]{Z(B)}} \right\rvert + 4.24$	0.97±0.01	29±5	0.87±0.02	64±3
5	{**H(A), r$_s$(A), EA(B), EN(B), Z(B)**} $a_0 = 3 \cdot \frac{r_s(A)}{Z(B)} \frac{EA(B) - EN(B)}{\exp(r_s(A))/\log(Z(B))} + 0.0152 \cdot \left\lvert \frac{Z(B)EA(B)}{EA(B) - EN(B)} - \left(\sqrt{Z(B)} - \frac{H(A)}{EN(B)} \right) \right\rvert - 1.25 \cdot \left\lvert \sqrt[3]{\log(EN(B))} - \frac{\log(Z(B))}{\sqrt[3]{Z(B)}} \right\rvert + 4.24$	0.98±0.01	21±4	0.87±0.03	64±4
6	{**r$_s$(A), EA(B), IP(B), Z(B), r$_p$(B), r$_s$(B)**} $a_0 = 4.48 \cdot \frac{EA(B) - IP(B)}{\exp(r_s(A))} \frac{r_s(A)}{Z(B)} \sqrt[3]{r_p(B)} + 0.0125 \cdot \left\lvert \frac{Z(B)r_s(B)}{\exp(r_p(B))} - \exp(r_s(B))\left(r_p(B) + r_s(B)\right) \right\rvert + 1.01 \cdot \left\lvert \exp\left(\frac{r_s(B)}{Z(B)} \right) - \frac{EA(B)}{EA(B) - IP(B)} \right\rvert + 4.09$	0.98±0.01	25±6	0.87±0.04	63±7

(continues on next page)

(continued from previous page)

7	{**r**$_s$**(A), EA(B), IP(B), Z(B)**}	0.97±0.01	27±6	0.85±0.03	67±7
	$a_0 = 1.78 \cdot \frac{EA(B) - IP(B)}{\exp(r_s(A))} \frac{r_s(A)}{Z(B)} \sqrt{IP(B)} - 0.258 \cdot \frac{EA(B)^{-1}}{IP(B)^3 \big(EA(B) + (IP(B)/Z(B))\big)} + 0.00565 \cdot \left\lvert \frac{Z(B)EA(B)}{\log(IP(B))} - \Big(\big(EA(B) - IP(B)\big) - \big(IP(B) - EA(B)\big)\Big) \right\rvert + 4.15$				
8	{**r**$_s$**(A), EA(B), EN(B), Z(B), r**$_p$**(B)**}	0.98±0.01	25±6	0.85±0.03	68±5
	$a_0 = 0.00974 \cdot \frac{Z(B)r_s(A)}{EA(B) + EN(B)} \big(\exp(EA(B)) + \exp\big(-r_p(B)\big)\big) - 143 \cdot \frac{\sqrt[3]{EN(B)}}{Z(B)^2} \left\lvert \frac{r_s(A)}{EA(B)} - \frac{r_s(A)}{EN(B)} \right\rvert^{-1} - 0.000141 \cdot \frac{Z(B)^3}{\exp(EN(B))} ((EA(B) - EN(B)) - Z(B)EA(B))^{-1} + 3.77$				
9	{**r**$_s$**(A), EA(B), L(B), Z(B), r**$_s$**(B)**}	0.98±0.01	24±6	0.85±0.03	68±7
	$a_0 = -0.083 \cdot \left(\left(\frac{EA(B)}{L(B)} - \sqrt{Z(B)}\right) - \frac{\exp(r_s(A))}{\log(Z(B))}\right) + 0.00294 \cdot \left\lvert \exp(r_s(B))^2 - \frac{Z(B)EA(B)}{L(B) - EA(B)} \right\rvert + 0.133 \cdot \left\lvert \exp(r_s(B)) \log(r_s(B))) - \frac{L(B)}{L(B) - EA(B)} \right\rvert + 3.09$				
10	{**IP(A), Z(A), EA(B), Z(B), r**$_s$**(B)**}	0.99±0.00	18±3	0.86±0.04	66±8
	$a_0 = -0.121 \cdot \left((\sqrt[3]{Z(A)} - \sqrt{Z(B)}) - \frac{\sqrt{(}Z(A))}{\log(IP(A))}\right) + 16.8 \cdot \left(\sqrt[3]{\exp(-Z(B))} - \frac{\exp(-EA(B))}{Z(B)^2}\right) - 1.26\text{e-}05 \cdot Z(B)^2 \log(r_s(B)) \left((EA(B))^{-1} + \frac{Z(B)}{IP(A)}\right) + 3.19$				

Tab. 5.9. Ranked list of statistical equivalent symbolic-regression models for the prediction of the equilibrium lattice constant a_0 in ascending order of the SISSO [197] prediction errors (RMSE - ΔRMSE). Shown is the prediction performance of identified, best-performing, non-redundant, sub-optimal feature subsets of the tolerance-based branch-and-bound algorithm (TB3) using the sure-independence screening and sparsifying operator (SISSO) [197] as feature-selection criterion Sections 3.1.2 and 4.1). For comparison, 10-fold cross-validated (cf., Section 5.1) prediction performance of SISSO and the gradient-boosting decision trees (GBDT) algorithm [249–252, 267, 268] are reported in terms of the mean $\bar{x}$ and standard deviation Δx as $\bar{x} \pm \Delta x$ of the Pearson's coefficient of determination $x = R^2$ [101] and the root-mean-squared error (x = RMSE). The RMSE is in the unit of milli angstrom (mÅ).

#	Features + Model	GBDT		SISSO	
		R^2	RMSE [GPa]	R^2	RMSE [GPa]
1	**{IP(A), c(A), EA(B), r$_s$(B)}** $B_0 = 0.607 \cdot \frac{\sqrt[3]{c(A)}\log(r_s(B))}{\sqrt[3]{\log(r_s(B))}} - 0.701 \cdot \left\lvert \frac{EA(B)}{EA(B) - IP(A)} - \exp(EA(B))\log(r_s(B)) \right\rvert$ $+ 0.0239 \cdot \frac{\lvert c(A) - r_s(B)\rvert}{IP(A)^2}\big(EA(B) + IP(A)\big)^3 + 0.809$	0.95±0.02	9.3±1.6	0.69±0.05	22.8±1.1
2	**{Z(A), c(A), EA(B), r$_s$(B)}** $B_0 = -2.09 \cdot \frac{r_s(B)^3}{\sqrt{c(A)}\exp(r_s(B)^3)} - 0.0244 \cdot r_s(B)^3\exp(EA(B))\lvert\exp(r_s(B)) - \exp(-EA(B))\rvert$ $- 0.0552 \cdot \left(\frac{\log(c(A))}{\exp(EA(B))} - \frac{\exp(c(A))}{\sqrt[3]{Z(A)}}\right) + 1.41$	0.95±0.02	9.2±1.5	0.68±0.07	23.2±1.9
3	**{c(A), EA(B), r$_p$(B)}** $B_0 = -0.0688 \cdot \Big(\big(r_p(B)^2\exp(EA(B))\big) - \big(r_p(B)(c(A) + r_p(B))\big)\Big)$ $- 0.0551 \cdot \frac{EA(B)^3(EA(B)r_p(B))}{\lvert\exp\big(r_p(B)\big) - \exp(-EA(B))\rvert} - 0.0886 \cdot \frac{\exp\big(EA(B)^{-1}\big)}{(c(A) - r_p(B)) - r_p(B)} + 0.77$	0.91±0.02	12.2±1.7	0.67±0.06	23.7±1.7
4	**{r$_s$(A), c(A), EA(B), r$_p$(B)}** $B_0 = 2.38 \cdot \frac{c(A)r_s(A)}{\exp(c(A))}\frac{1}{\exp(EA(B)) - \exp\big(r_p(B)\big)} + 0.126 \cdot \left\lvert\frac{EA(B)}{\log\big(r_p(B)\big)} - \frac{EA(B)}{r_p(B)}(c(A) + r_p(B))\right\rvert$ $- 0.131 \cdot \left\lvert\frac{\log\big(r_p(B)\big)}{\exp(EA(B))} - \lvert\exp(EA(B)) - \exp(-EA(B))\rvert\right\rvert + 1.35$	0.95±0.02	8.9±1.4	0.66±0.06	23.8±1.7
5	**{c(A), EA(B), r$_s$(B)}** $B_0 = -3.09 \cdot \frac{r_s(B)^2}{\sqrt{c(A)}\exp(r_s(B)^2)} - 0.0532 \cdot (r_s(B)^2 r_s(B)^3)\lvert\exp(EA(B)) - \exp(-r_s(B))\rvert$ $- 0.0961 \cdot \big(\exp(-EA(B)) - \exp(r_s(B))\big)\left\lvert\sqrt{c(A)} - \sqrt{r_s(B)}\right\rvert + 1.87$	0.91±0.02	12.2±1.7	0.66±0.06	23.9±1.4
6	**{c(A), EA(B), r$_s$(B)}** $B_0 = -6.9 \cdot \frac{r_s(B)}{\exp(r_s(B))\sqrt[3]{c(A) + r_s(B)}} - 0.0204 \cdot \big(r_s(B)^3\exp(EA(B))\big)^2$ $+ 0.225 \cdot \frac{r_s(B)^2(c(A)EA(B))}{\exp(r_s(B)^3)} + 2.85$	0.91±0.02	12.2±1.7	0.67±0.08	23.5±2.1
7	**{EN(A), c(A), EA(B), r$_p$(B)}** $B_0 = -0.0538 \cdot \Big(\big(r_p(B)^2\exp(EA(B))\big) - \big(r_p(B)(c(A) + r_p(B))\big)\Big)$ $- 0.0255 \cdot \frac{\exp(-EA(B)) - \exp\big(r_p(B)\big)}{\big\lvert r_p(B) - \lvert c(A) - r_p(B)\rvert\big\rvert} - 0.000465 \cdot \frac{EN(A)^3 r_p(B)^2}{(c(A) - r_p(B)) - r_p(B)} + 0.669$	0.94±0.02	9.5±1.5	0.64±0.05	24.7±2.0

(continues on next page)

(continued from previous page)

8	{**Z(A), c(A), EA(B), r$_p$(B)**}	0.95±0.02	8.9±1.5	0.64±0.05	24.9±2.2
	$B_0 = 0.884 \cdot \left(\frac{\exp(c(A))\sqrt{r_p(B)}}{\exp(c(A)) + \exp(EA(B))}\right) + 0.0659 \cdot \left(\frac{\exp(c(A))}{\sqrt(Z(A))} - \big(\exp(EA(B)) + \exp(-EA(B))\big)\right) - 0.0676 \cdot \left(\left\|\sqrt[3]{Z(A)} + \frac{r_p(B)}{c(A)}\right\| - \|\exp\big(r_p(B)\big) - \exp(-EA(B))\|\right) + 0.198$				
9	{**IP(A), L(A), Z(B), r$_s$(B)**}	0.94±0.02	9.8±1.7	0.65±0.07	24.8±2.5
	$B_0 = -0.182 \cdot \frac{r_s(B)^3\sqrt{Z(B)}}{\exp(r_s(B)^3)} - 3.05 \cdot \frac{\log(IP(A) + L(A))}{(IP(A)/L(A)) - \sqrt[3]{Z(B)}} + 9.28\text{e-}06 \cdot \exp\left(\sqrt{Z(B)r_s(B)}\right) + 0.726$				
10	{**H(A), IP(A), EA(B), r$_s$(B)**}	0.95±0.02	9.3±1.5	0.65±0.08	24.5±2.1
	$B_0 = -3.87 \cdot \frac{r_s(B)^3}{(IP(A) + H(A))\exp(r_s(B)^3)} + 0.000551 \cdot \frac{r_s(B)^{-1}}{\log(r_s(B))\log(EA(B)^2)} - 0.137 \cdot \left(\left\|\frac{IP(A) - EA(B)}{EA(B) + IP(A)}\right\| - \left\|\exp(r_s(B)) - \exp(-EA(B))\right\|\right) + 1.57$				

Tab. 5.10. Ranked list of statistical equivalent symbolic-regression models for the prediction of the bulk modulus B_0 in ascending order of the SISSO [197] prediction errors (RMSE - ΔRMSE). Shown is the prediction performance of identified, best-performing, non-redundant, sub-optimal feature subsets of the tolerance-based branch-and-bound algorithm (TB3) using the gradient-boosting decision trees (GBDT) algorithm [249–252, 267, 268] as feature-selection criterion Sections 3.1.2 and 4.1). For comparison, 10-fold cross-validated (cf., Section 5.1) prediction performance of SISSO and the gradient-boosting decision trees (GBDT) algorithm [249–252, 267, 268] are reported in terms of the mean $\bar{x}$ and standard deviation Δx as $\bar{x} \pm \Delta x$ of the Pearson's coefficient of determination $x = R^2$ [101] and the root-mean-squared error (x = RMSE). The RMSE is in units of gigapascal (GPa).

#	**Features**	**GBDT**		**SISSO**	
	+ Model	R^2	RMSE [GPa]	R^2	RMSE [GPa]
1	{**H(A), IP(A), EA(B), EN(B), L(B), Z(B), r$_s$(B)**}	0.94±0.02	9.5±1.8	0.74±0.05	21.1±2.5
	$B_0 = -2.8 \cdot \frac{r_s(B)^3}{(IP(A) + H(A))\exp(r_s(B)^3)} - 0.128 \cdot \left\|\big(\exp(L(B)) - \exp(r_s(B))\big) - \frac{L(B)\sqrt[3]{Z(B)}}{EN(B)}\right\| + 0.242 \cdot \left\|\frac{EA(B) - IP(A)}{\sqrt[3]{Z(B)}} - \frac{H(A) + L(B)}{\exp(r_s(B))}\right\| + 1.36$				
2	{**L(A), Z(A), c(A), EA(B), H(B), Z(B), r$_s$(B)**}	0.93±0.02	10.2±1.9	0.73±0.05	21.5±1.5
	$B_0 = -22.5 \cdot \frac{r_s(B)^3}{\sqrt[3]{c(A)}(\exp(r_s(B)))^3} - 0.14 \cdot \left\|\exp(r_s(B))\log(r_s(B)) - \frac{EA(B)\sqrt{Z(B)}}{H(B)}\right\| - 0.227 \cdot \left(\frac{c(A)L(A)}{Z(B)r_s(B)} + \frac{EA(B)H(B)}{Z(A)L(A)}\right) + 1.97$				

(continues on next page)

(continued from previous page)

3	**{c(A), EA(B), Z(B), $\mathbf{r}_s$(B)}**	0.91±0.02	12.2±1.7	0.72±0.05	21.5±1.7
	$B_0 = 5.34 \cdot \left(\frac{r_s(B)\log(c(A))}{Z(B)} - \frac{r_s(B)}{\exp(r_s(B))}\right) - 0.0641 \cdot \left\lvert\frac{\sqrt{Z(B)}}{\exp(r_s(B))} - \frac{\exp(EA(B))}{\exp(-r_s(B))}\right\rvert + 0.00713 \cdot \left\lvert Z(B)\log(r_s(B)) - \frac{\exp(EA(B))}{\exp(-c(A))}\right\rvert + 2.83$				
4	**{c(A), EA(B), H(B), L(B), $\mathbf{r}_s$(B)}**	0.91±0.02	12.2±1.7	0.71±0.04	22.0±2.3
	$B_0 = -0.688 \cdot \left(\left(\exp(-c(A)) + \exp(L(B))\right) - \left(\log(r_s(B))\right)^2\right) - 0.22 \cdot \left\lvert\exp(r_s(B))\log(r_s(B)) - \frac{EA(B)+L(B)}{L(B)-EA(B)}\right\rvert - 0.0601 \cdot (c(A)H(B)\exp(L(B)) - r_s(B)(H(B)-L(B))) + 1.38$				
5	**{$\mathbf{r}_s$(A), c(A), EA(B), EN(B), H(B), Z(B)}**	0.95±0.02	9.1±1.8	0.73±0.07	21.2±2.1
	$B_0 = 3.59 \cdot \left(\frac{c(A)}{Z(B)}\log(EN(B)) - \frac{r_s(A)}{Z(B)}\exp(EA(B))\right) + 0.00279 \cdot \frac{Z(B)}{H(B)}\frac{EN(B)+H(B)}{(EA(B)-EN(B))-H(B)} - 7.9\text{e-}06 \cdot \frac{\exp(EA(B)H(B))}{\sqrt[3]{Z(B)} + (EN(B)/EA(B))} + 0.981$				
6	**{H(A), IP(A), EA(B), Z(B), $\mathbf{r}_s$(B)}**	0.94±0.02	9.4±1.7	0.69±0.05	23.0±2.0
	$B_0 = -3.23 \cdot \frac{r_s(B)^3}{(IP(A)+H(A))\exp(r_s(B)^3)} - 0.0466 \cdot \left\lvert\frac{\sqrt{Z(B)}}{\exp(r_s(B))} - \frac{\exp(EA(B))}{\exp(-r_s(B))}\right\rvert - 50.1 \cdot \left(\frac{EA(B)}{Z(B)}\frac{IP(A)+H(A)}{\left(IP(A)-H(A)\right) - Z(B)EA(B)}\right) + 1.33$				
7	**{c(A), EA(B), Z(B), $\mathbf{r}_p$(B)}**	0.91±0.02	12.6±1.7	0.67±0.05	23.4±2.4
	$B_0 = -0.614 \cdot \left(\frac{\sqrt[3]{Z(B)}}{\exp(r_p(B))} + \exp(-c(A))\exp(EA(B))\right) + 1.27 \cdot \frac{c(A)-r_p(B)}{Z(B)\exp\left(\sqrt[3]{EA(B)}\right)} - 3.95\text{e-}06 \cdot \frac{Z(B)^2(c(A)-r_p(B))}{\log\left(\lvert c(A)-r_p(B)\rvert\right)} + 1.46$				

Tab. 5.11. Ranked list of statistical equivalent symbolic-regression models for the prediction of the bulk modulus B_0 in ascending order of the SISSO [197] prediction errors (RMSE - ΔRMSE). Shown is the prediction performance of identified, best-performing, non-redundant, sub-optimal feature subsets of the tolerance-based branch-and-bound algorithm (TB3) using the sure-independence screening and sparsifying operator (SISSO) [197] as feature-selection criterion Sections 3.1.2 and 4.1). For comparison, 10-fold cross-validated (cf., Section 5.1) prediction performance of SISSO and the gradient-boosting decision trees (GBDT) algorithm [249–252, 267, 268] are reported in terms of the mean $\bar{x}$ and standard deviation Δx as $\bar{x} \pm \Delta x$ of the Pearson's coefficient of determination $x = R^2$ [101] and the root-mean-squared error (x = RMSE). The RMSE is in units of gigapascal (GPa).

Statistical models

The prediction performance of statistical equivalent (cf., Section 5.1) non-redundant sub-optimal feature subsets of TB3-GBDT and TB3-SISSO are summarized in (Tabs. 5.8 to 5.11). Specifically, feature subsets of all machine-learning models consists of at least one feature from each atomic species. Machine-learning models further show moderate to good prediction performances with comparatively small prediction errors of less than 10% of the range of prediction values.

Of all identified feature subsets, only 6 of 79 (TB3-GBDT) and 14 of 17 (TB3-SISSO) feature subsets are statistically equivalent to the optimal feature subset for the prediction of the equilibrium lattice constant (a_0), while 14 of 119 (TB3-GBDT) and 7 of 11 (TB3-SISSO) feature subsets are statistically equivalent to the optimal feature subset for the prediction of the bulk modulus (B_0). Consequently, there is an ensemble of competing machine-learning models of different feature subsets for estimating the equilibrium lattice constant and the bulk modulus (cf., Chapter 4). The most frequent occurring features are c(A), r_s(A), Z(B), EA(B), r_s(B), r_p(B). These features have been identified in almost all of the identified relevant features of the feature-identification methods (Tabs. 5.5 and 5.6). It can therefore be argued that TCMI, RFECV, FS-GBDT, and FS-SISSO identify more features as dependent as actually required for creating highly predictive machine-learning models, while the TB3-algorithm is able to further reduce the number of features without degrading the prediction performance of the machine-learning model.

Uncertainty estimation

Ensemble-based predictions strongly under-estimate the prediction intervals of the machine-learning models (Figs. 5.13 and 5.14). In contrast, conformal prediction [62–64] correctly estimates the prediction intervals at the specified confidence levels (Tab. 5.12). The validity (i.e., the percentage of actual values outside the prediction interval) is only violated at a confidence level of $\alpha = 0.50$ as a result of randomly halving the perovskite oxides data set and machine-learning models that become significantly dependent on the choice of training data.

Ensemble-based prediction intervals are smaller than conformal-based prediction intervals at the same confidence level α, but are as large as conformal-based prediction intervals at the same validity $\hat{\alpha}$ (Tab. 5.12). Therefore, conformal prediction is advantageous in estimating the target properties of the perovskite oxides data set by providing statistical guarantees for the uncertainties of the predictions [363, 366]. Although the maximum prediction intervals can be as large as the range of the property of interests and not all predictions may be within the 95% prediction bands, constructed (ensemble) machine-learning models clearly capture the underlying trend of the data. Overall, machine-learning models slightly under-estimate the lattice constant, while machine-learning models over-estimate perovskites with low and under-estimate perovskites with high bulk moduli (Figs. 5.13 and 5.14).

Due to the poor prediction performance of symbolic-regression models for the bulk modulus, both neodymium zinc oxide ($\mathrm{NdZnO_3}$) and neodymium cadmium oxide ($\mathrm{NdCdO_3}$) cannot be reliably predicted by the feature subsets of the TB3-GBDT algorithm with a symbolic-regression model (Fig. 5.14). Much tighter prediction intervals can be obtained by applying ensemble-based and conformal predictions to the same feature subsets as identified by TB3-GBDT/TB3-SISSO but using a piecewise-constant machine-learning models such as the gradient-boosting decision trees algorithm

SISSO

↲ **Algorithm (TB3-...)**		**Ensemble prediction**			**Conformal prediction**		
	Confidence level (α)	Validity ($\hat{\alpha}$)	Mean (Δ)	Max (Δ)	Validity ($\hat{\alpha}$)	Mean (Δ)	Max (Δ)
a.) Lattice constant (a_0)							
GBDT	0.50	*0.15*	0.02	0.11	*0.49*	0.06	0.20
	0.80	*0.26*	0.03	0.14	*0.81*	0.11	0.37
	0.95	*0.31*	0.04	0.19	*0.96*	0.25	0.80
SISSO	0.50	*0.34*	0.02	0.06	*0.47*	0.04	0.16
	0.80	*0.57*	0.05	0.10	*0.82*	0.08	0.30
	0.95	*0.71*	0.06	0.15	*0.96*	0.18	0.67
b.) Bulk modulus (B_0)							
GBDT	0.50	*0.28*	7.6	20.2	*0.47*	14.6	47.8
	0.80	*0.50*	13.8	41.5	*0.80*	31.1	99.8
	0.95	*0.67*	13.2	838.2	*0.97*	83.4	263.0
SISSO	0.50	*0.28*	7.0	21.1	*0.50*	13.4	26.2
	0.80	*0.49*	12.7	32.5	*0.80*	26.6	69.9
	0.95	*0.59*	16.2	39.5	*0.96*	62.0	157.5

Tab. 5.12. Validity of ensemble and conformal prediction obtained from TB3-GBDT and TB3-SISSO at three different confidence levels $\alpha = [50\%, 80\%, 95\%]$ using the sure-independence screening and sparsifying operator (SISSO) [197] as the machine-learning algorithm, where the confidence level of the ensemble prediction was computed as the αth-percentile of the ensemble predictions. The validity ($\hat{\alpha}$) specifies the probability ($\hat{\alpha}\% = 100\hat{\alpha}$) that the actual value x of the underlying process is within the prediction interval $x \in [\bar{x} - \bar{\Delta}, \bar{x} + \bar{\Delta}]$ (Eq. 4.22) of a (point) prediction $\bar{x}$ with uncertainty $\pm\bar{\Delta}$ ($\bar{\Delta} = \text{Mean}(\Delta) \leq \text{Max}(\Delta)$). Shown are the performance statistics of the lattice constant (a_0) in units of milliangstroms (a) and the bulk modulus (B_0) in units of gigapascals (b).

[249–252, 267, 268] (SISSO: Tab. 5.12, GBDT: Tab. 5.13) for the estimation of the equilibrium lattice constant or bulk modulus. Compared to symbolic-regression models (Figs. 5.13 and 5.14), choosing a GBDT machine-learning model would not only reduce the prediction errors by a factor of at least 2, but would also lead to highly predictive machine-learning models (Figs. 5.15 and 5.16).

Anomalous materials

An analysis of the perovskites compounds shows that there are no anomalous materials in the data set. Though, none of the perovskites are classified as anomalous, materials with low credibility scores, i.e., $\Pi < 0.2$, provide insights into potential weaknesses of the constructed machine-learning models. A more detailed investigation of these materials may therefore be desirable to improve the machine-learning predictions for an actual application.

As the credibility score and the prediction intervals (Δ) exhibit a weak monotonic relationship ($\rho^2 \approx 0.3$) in terms of the Spearman coefficient of determination, materials with low credibility score tend to have large prediction errors. For instance, perovskite compounds with comparatively low credibility scores such as zirconium-based perovskites (e.g., $LaZrO_3$, $LiZrO_3$, $LiZrO_3$, etc.) or

GBDT

↲ **Algorithm (TB3-. . .)**		**Ensemble prediction**			**Conformal prediction**		
	Confidence level (α)	Validity ($\hat{\alpha}$)	Mean (Δ)	Max (Δ)	Validity ($\hat{\alpha}$)	Mean (Δ)	Max (Δ)
a.) Lattice constant (a_0)							
GBDT	0.50	*0.25*	0.00	0.05	*0.49*	0.01	0.05
	0.80	*0.40*	0.01	0.07	*0.81*	0.02	0.14
	0.95	*0.49*	0.01	0.09	*0.97*	0.05	0.34
SISSO	0.50	*0.25*	0.01	0.05	*0.49*	0.01	0.05
	0.80	*0.50*	0.01	0.07	*0.82*	0.02	0.13
	0.95	*0.67*	0.01	0.08	*0.97*	0.06	0.31
b.) Bulk modulus (B_0)							
GBDT	0.50	*0.42*	3.6	15.6	*0.50*	4.2	20.0
	0.80	*0.59*	5.4	20.2	*0.80*	10.5	48.7
	0.95	*0.72*	6.7	23.0	*0.97*	32.2	138.6
SISSO	0.50	*0.32*	3.4	15.1	*0.50*	4.7	21.3
	0.80	*0.46*	4.6	17.8	*0.80*	10.9	46.1
	0.95	*0.57*	5.3	19.5	*0.96*	29.3	127.6

Tab. 5.13. Validity of ensemble and conformal prediction obtained from TB3-GBDT and TB3-SISSO at three different confidence levels $\alpha = [50\%, 80\%, 95\%]$ using the gradient-boosting decision trees (GBDT) algorithm [249–252, 267, 268] as the machine-learning algorithm, where the confidence level of the ensemble prediction was computed as the αth-percentile of the ensemble predictions. The validity ($\hat{\alpha}$) specifies the probability ($\hat{\alpha}\% = 100\hat{\alpha}$) that the actual value x of the underlying process is within the prediction interval $x \in [\bar{x} - \bar{\Delta}, \bar{x} + \bar{\Delta}]$ (Eq. 4.22) of a (point) prediction $\bar{x}$ with uncertainty $\pm\bar{\Delta}$ ($\bar{\Delta} = \mathrm{Mean}(\Delta) \leq \mathrm{Max}(\Delta)$). Shown are the performance statistics of the lattice constant (a_0) in units of milliangstroms (a) and the bulk modulus (B_0) in units of gigapascal (b).

ferroelectric perovskites (e.g., $NdFeO_3$) have widely dispersed lattice constants ($a_0 = [3.6\,, 4.3]$ Å) and medium to high bulk moduli ($B_0 = [90\,, 180]$ GPa) with prediction errors (ε) and intervals (Δ) larger than the mean of the full data set. Also, the credibility scores of constructed (ensemble of) machine-learning models with SISSO (Figs. 5.13 and 5.14) show that perovskite compounds containing bismuth ($CaBiO_3$, $CsBiO_3$, $NaBiO_3$, etc.) have larger prediction errors, probably as a result of being in the top 1% percentile of all perovskite compounds with a lattice constant greater than $a_0 > 4.2$ Å. Finally, perovskites compounds based on neodymium and cadmium (namely $NdCdO_3$, $NdZnO_3$, $CeCdO_3$, $LaCdO_3$) have lower credibility scores than other perovskites when using TB3-GBDT prior to building symbolic-regression models with SISSO: The features as identified by TB3-GBDT seem not to be sufficient to uniquely describe neodymium- and cadmium-based compounds.

Summary

The perovskite oxides data set is characterized by strong feature interactions between features for each of the atomic species A/B and moderate to good prediction performances of machine-learning models with comparatively small errors of less than 10% of the range of prediction values. In particular, it has been shown that TCMI, RFECV, FS-GBDT, and FS-SISSO identify more features as dependent than the tolerance-based branch-and-bound (TB3) algorithm. Hence, the TB3-algorithm can efficiently reduce the number of features without degrading the prediction performance of the generated machine-learning models.

By construction, perovskites materials are uniquely determined by the nuclear and ionic charges of the constituent elements: Z(A), Z(B), and c(A). Feature-dependence maps illustrate this finding and show that all features of each atomic species can be expressed as an implicit dependence on the nuclear charges Z. Therefore, a machine-learning model for the prediction of the equilibrium lattice constant a_0 or the bulk modulus B_0 requires at least one feature from each atomic species. Most of these features are related by physical or statistical relationships as like in the octet-binary compound semiconductor data set (Section 5.2.1). Though the nuclear and ionic charges are not identified as relevant in all feature-identification methods, TB3-GBDT and TB3-SISSO identify them as one of the sub-optimal minimally non-redundant feature subsets or as a part of feature subsets with larger feature-subset cardinalities.

The most frequently identified features with the TB3 algorithm for the prediction of the equilibrium lattice constant and the bulk modulus are c(A), r_s(A), Z(B), EA(B), r_s(B), and r_p(B). Notably, these features have been identified in almost all of the identified relevant features of the feature-identification methods (Tabs. 5.5 and 5.6). The fact that the ionic charge and the most frequently identified features are consistently identified by the minimal non-redundant feature subsets of the TB3-algorithm therefore suggests that the developed framework has the potential to identify physically relevant features from a statistical analysis of the materials data.

Common to all feature-identification methods is the good prediction performance of the lattice constant, but only moderate prediction performance of the bulk modulus. A deeper investigation reveals that a machine-learning model of the bulk modulus can reach a similar prediction performance as the equilibrium lattice constant (Tab. 5.7) without requiring additional features by either using the computed lattice constants from DFT ($\hat{a}$) or an estimation of the equilibrium lattice constants ($\hat{a}_0$) based on the 16 atomic features (Eq. 5.2). Furthermore, it has been demonstrated that the bulk modulus have an explicit dependence on the equilibrium lattice constant, the nuclear, and the ionic charges (cf., [410, 411]). Both, the lattice constant and the bulk modulus can be estimated with high accuracy by the 16 atomic features, although implicitly a machine-learning model for the lattice constant should be created first and this prediction used to estimate the bulk modulus. In addition, the developed framework identified an ensemble of competing machine-learning models of different features subsets for estimating the equilibrium lattice constant and the bulk modulus.

The ensemble of machine-learning models have been shown to strongly under-estimate the prediction intervals of machine-learning predictions. By way of contrast, conformal prediction correctly estimated the prediction intervals at the specified confidence levels (Tab. 5.12) and is therefore indispensable for a reliable uncertainty estimation of machine-learning predictions of the perovskite oxides data set.

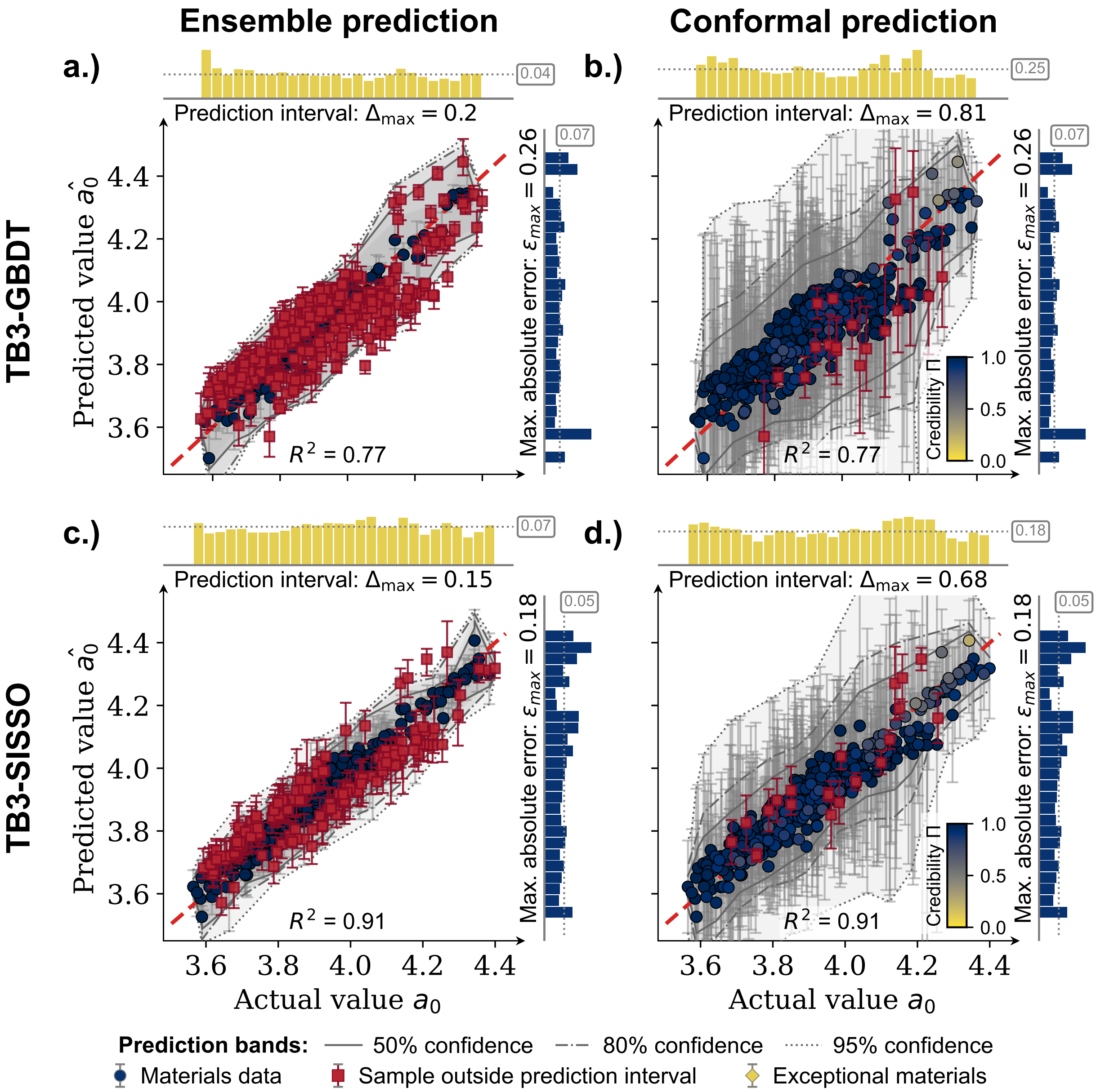


Fig. 5.13. Ensemble prediction performance of symbolic-regression models (SISSO [197]) constructed from feature subsets of TB3-GBDT and TB3-SISSO to estimate the equilibrium lattice constant a_0 of perovskites materials. Shown are the prediction bands (50th, 80th, 95th-percentiles of the model's predictions), the credibility Π (Eq. 4.29) at a confidence level of $\alpha = 0.95$, the distribution of the size of prediction intervals (diagram above the parity plot, Δ), and the errors of the machine-learning model (diagram right of the parity plot, $\varepsilon = |a_0 - \hat{a}_0|$). The numbers in the boxes display the mean values ($\bar{\Delta}$, $\bar{\varepsilon}$), while the maximum errors are given in the texts below or to the left of the diagrams. Units are in angstrom (Å). Predictions outside the prediction intervals are depicted as squares and anomalous materials as diamond-shape symbols.

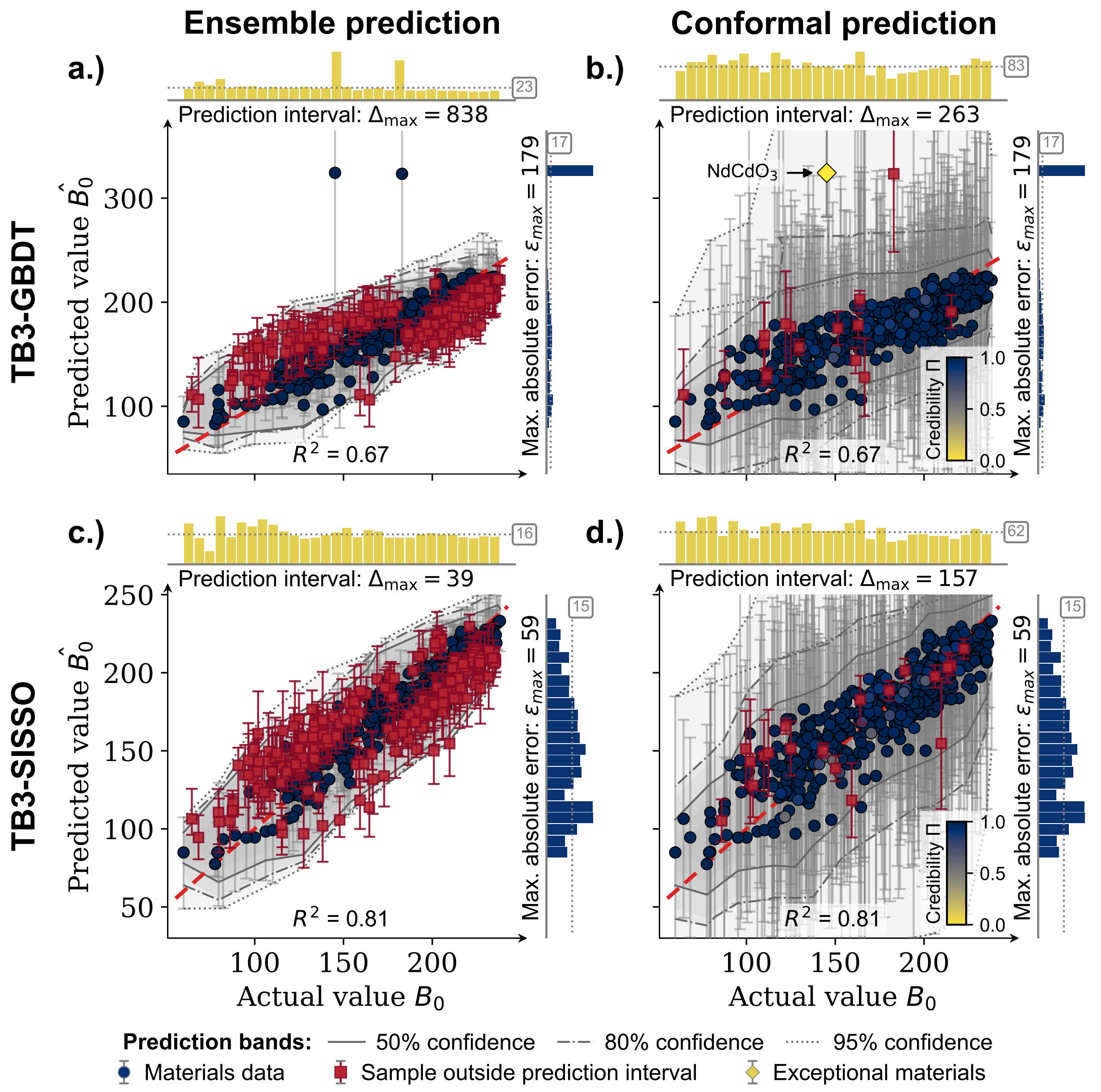


Fig. 5.14. Ensemble prediction performance of symbolic-regression models (SISSO [197]) constructed from feature subsets of TB3-GBDT and TB3-SISSO to estimate the bulk modulus B_0 of perovskites materials. Shown are the prediction bands (50th, 80th, 95th-percentiles of the model's predictions), the credibility Π (Eq. 4.29) at a confidence level of $\alpha = 0.95$, the distribution of the size of prediction intervals (diagram above the parity plot, Δ), and the errors of the machine-learning model (diagram right of the parity plot, $\varepsilon = |a_0 - \hat{a_0}|$). The numbers in the boxes display the mean values ($\bar{\Delta}$, $\bar{\varepsilon}$), while the maximum errors are given in the texts below or to the left of the diagrams. Units are in gigapascal (GPa). Predictions outside the prediction intervals are depicted as squares and anomalous materials as diamond-shape symbols.

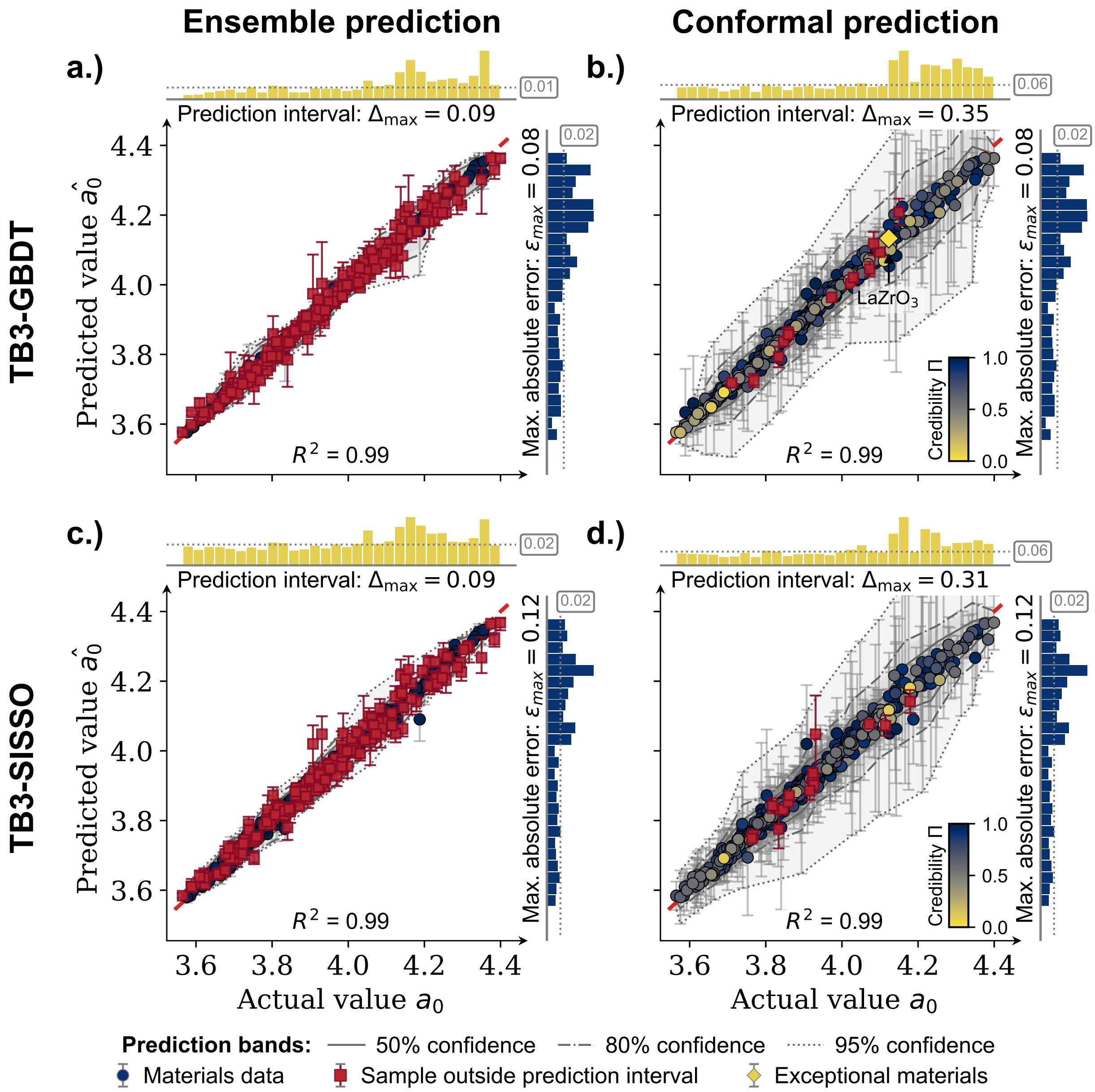


Fig. 5.15. Ensemble prediction performance of gradient-boosting decision tree (GBDT) models [249–252, 267, 268] constructed from feature subsets of TB3-GBDT and TB3-SISSO to estimate the equilibrium lattice constant a_0 of perovskites materials. Shown are the prediction bands (50th, 80th, 95th-percentiles of the model's predictions), the credibility Π (Eq. 4.29) at a confidence level of $\alpha = 0.95$, the distribution of the size of prediction intervals (diagram above the parity plot, Δ), and the errors of the machine-learning model (diagram right of the parity plot, $\varepsilon = |a_0 - \hat{a}_0|$). The numbers in the boxes display the mean values ($\bar{\Delta}$, $\bar{\varepsilon}$), while the maximum errors are given in the texts below or to the left of the diagrams. Units are in angstrom (Å). Predictions outside the prediction intervals are depicted as squares and anomalous materials as diamond-shape symbols.

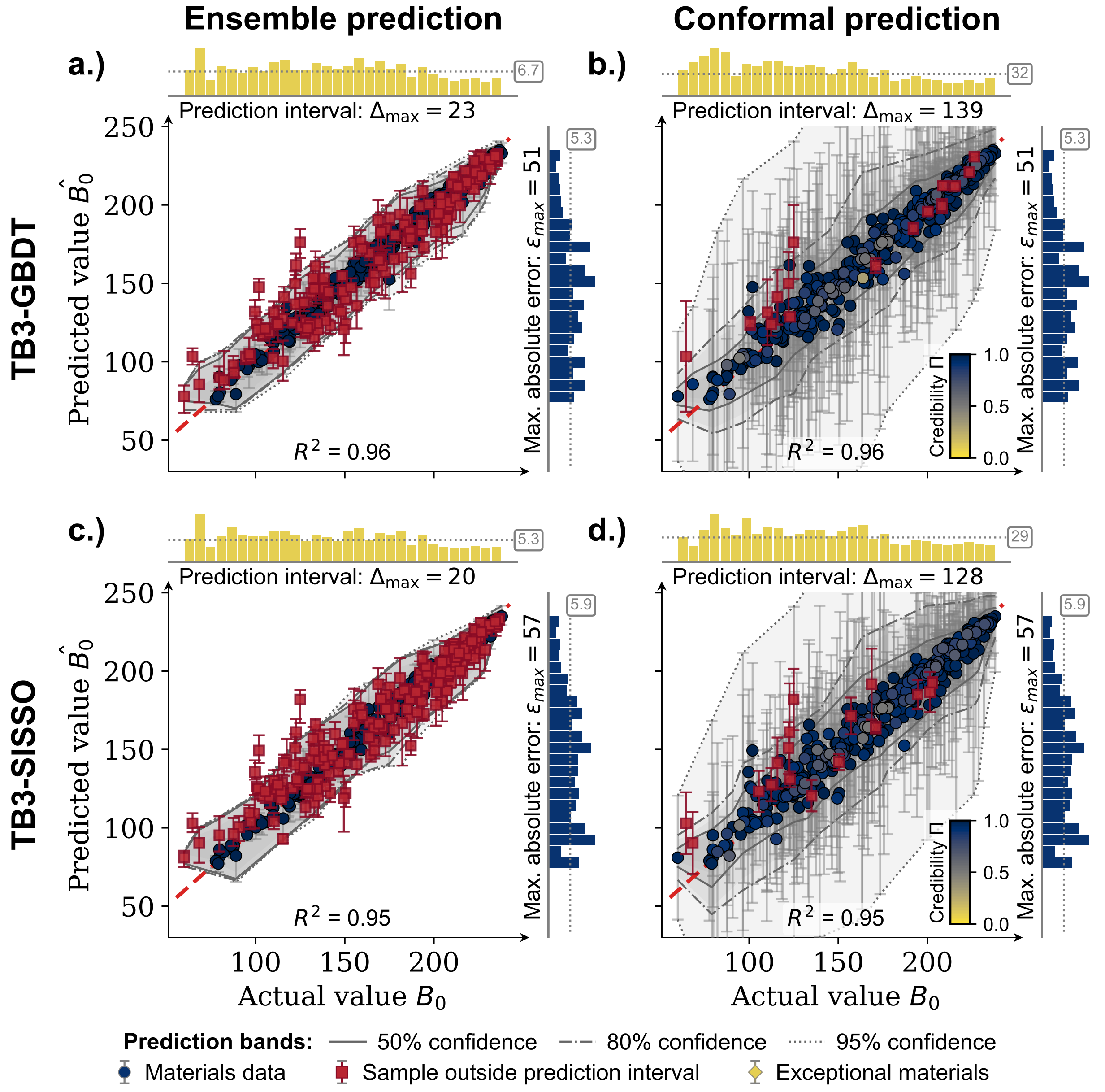


Fig. 5.16. Ensemble prediction performance of gradient-boosting decision tree (GBDT) models [249–252, 267, 268] constructed from feature subsets of TB3-GBDT and TB3-SISSO to estimate the bulk modulus B_0 of perovskites materials. Shown are the prediction bands (50th, 80th, 95th-percentiles of the model's predictions), the credibility Π (Eq. 4.29) at a confidence level of $\alpha = 0.95$, the distribution of the size of prediction intervals (diagram above the parity plot, Δ), and the errors of the machine-learning model (diagram right of the parity plot, $\varepsilon = |a_0 - \hat{a_0}|$). The numbers in the boxes display the mean values ($\bar{\Delta}$, $\bar{\varepsilon}$), while the maximum errors are given in the texts below or to the left of the diagrams. Units are in gigapascal (GPa). Predictions outside the prediction intervals are depicted as squares and anomalous materials as diamond-shape symbols.

5.2.3 Elastic property predictions of inorganic crystalline compounds

The elastic tensor of an inorganic crystalline compound describes the linear response of a material to external forces [412, 461]. As such, it is correlated to many thermal and mechanical properties [165, 462–467], inter-atomic bondings [468, 469], and forces [165]. However, the elastic tensor is computationally intensive [166, 412, 470] and is difficult to estimate[19]. Therefore, extensive studies have been conducted to estimate the elastic tensor and derived quantities using machine learning.

Two of the many interesting derived macroscopic mechanical properties from the elastic tensor are the bulk (K) and shear modulus (G). The bulk modulus describes the material's resistance to uniform compression under external loads and forces (cf., Section 5.2.2), whereas the shear modulus describes the stiffness of a material with respect to forces parallel to the surfaces (Fig. 5.17). Both quantities relate the stress to the strain of a material (Fig. 5.17) and can be used to investigate the mechanical stability of a material.

Given their important role in the screening and development of new materials with targeted structural properties (cf., Section 5.2.2), numerous efforts have been undertaken to establish links between the bulk and shear modulus of a crystal to their atomic properties. Because only a small fraction of all known inorganic compounds have been determined so far[20], numerous machine-learning techniques [165, 166, 186, 470, 473, 474] have been developed to both correct the elastic tensor by first-principles calculations and to relate the bulk and shear modulus of an inorganic crystalline compound to the atomic properties of the constituent elements. For example, de Jong, Chen *et al.* [165] used multivariate local polynomial regression [475] and gradient boosting [249] to estimate elastic properties of k-nary compounds of diverse chemistries and structures. Isayev, Oses *et al.* [166] introduced the concept of property-labeled material fragments to incorporate crystal-structure information for a variety of electro-chemical properties. And Wang, Yang *et al.* and co-workers [186, 470, 474, 476] employed convolutional neural networks [105] to estimate the bulk and shear modulus.

Most of these studies extensively investigated the applicability of machine learning to estimate the bulk and shear modulus of inorganic crystalline compounds. Feature identification, if at all, was discussed only on a simplified basis using either feature importance [249–252, 266–268] of decision-tree algorithms [166, 473], partial-dependence plots [166, 249, 429], or prior knowledge [473]. Therefore, extending previous efforts in identifying the relevant features for estimating the bulk and shear modulus, this thesis focuses on the systematic, statistical identification and characterization of these features considering atomic properties to create predictive machine-learning models. In particular, this thesis takes advantage of the feature-identification framework developed in Chapter 4 and Section 5.1 to directly link the feature identification to the prediction performance of the generated machine-learning models.

[19] The elastic tensor can be measured experimentally, for instance, through single-crystal Brillouin scattering [471, 472]

[20] As of today about 20, 000 inorganic crystalline compounds are available in the materials databases AFLOW [16, 17, 33] and Materials Project [13, 15, 35]. This represents about 10% of all materials stored in the Materials Project and < 1% of materials in the AFLOW database.

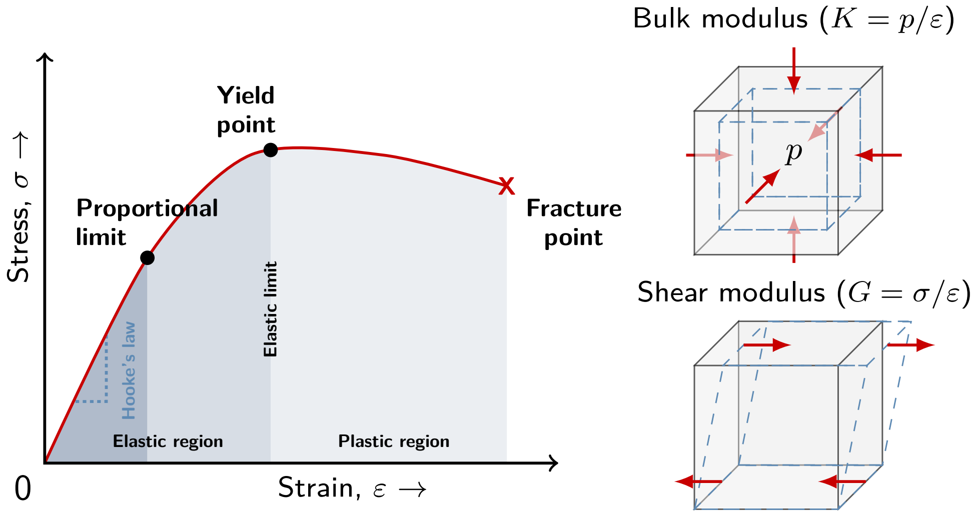


Fig. 5.17. Illustration of the stress-strain relationship (left) and the forces exerted on the material (right). Within the (linear) elastic regime, as given by Hooke's law, deformed materials return to their original configurations. In this regime, the bulk modulus describes the inverse compressibility of a material under uniform external loads (p - pressure, arrows), while the shear modulus characterizes the stiffness parallel to the surfaces (arrows). Beyond the elastic limit (i.e., the maximum stress or force per unit area) permanent deformation occurs, which causes a material to yield, flow, or fracture.

The dataset

The analysis is performed on the data set taken from Refs. [35, 412] using tabulated materials properties [151, 477–479] similar to Refs. [166, 473]. The data set consists of 1,181 inorganic crystalline compounds and includes metallic, small-band-gap binary oxides, and semiconductor compounds of varying numbers of compounds (1, . . . , 4). In total, the data set includes 63 elements, 87 space groups and six lattice systems (Fig. 5.18) – cubic (452), hexagonal (261), rhombohedral (37), tetragonal (193), orthorhombic (193), and monoclinic (45).

de Jong, Chen *et al.* used the projector-augmented wave method [480, 481] at the GGA level of the theory (cf., Section 2.2) and the PBE exchange-correlation functional [482] at zero temperature and pressure to compute the elastic tensor components from a linear fit of a generated set of distorted structures. As the bulk K and the shear G modulus are directional quantities, they determined the bulk and shear modulus with upper ($K \leq K_V, G \leq G_V$) and lower bounds ($K_R \leq K, G_R \leq G$) in the isotropic approximation assuming uniform strain (Voigt notation [483]) and stress (Reuss [484]) respectively[21]. They also calculated the arithmetic mean of the two quantities (the so-called Voigt-Reuss-Hill (VRH) notation [485]),

$$K_{\mathrm{VRH}} = \frac{1}{2}\left(K_V + K_R\right) \ , \quad G_{\mathrm{VRH}} = \frac{1}{2}\left(G_V + G_R\right) \ , \tag{5.4}$$

to compare and validate them with the experimental data. de Jong, Chen *et al.* demonstrated that the VRH estimates of the bulk and shear modulus were strongly linear related to the experimental

[21] In a single crystal, the Voigt notation [483] corresponds to the arithmetic mean of the stresses with respect to the given strain, while the Reuss notation [484] corresponds to the arithmetic mean of the strains with respect to the given stress.

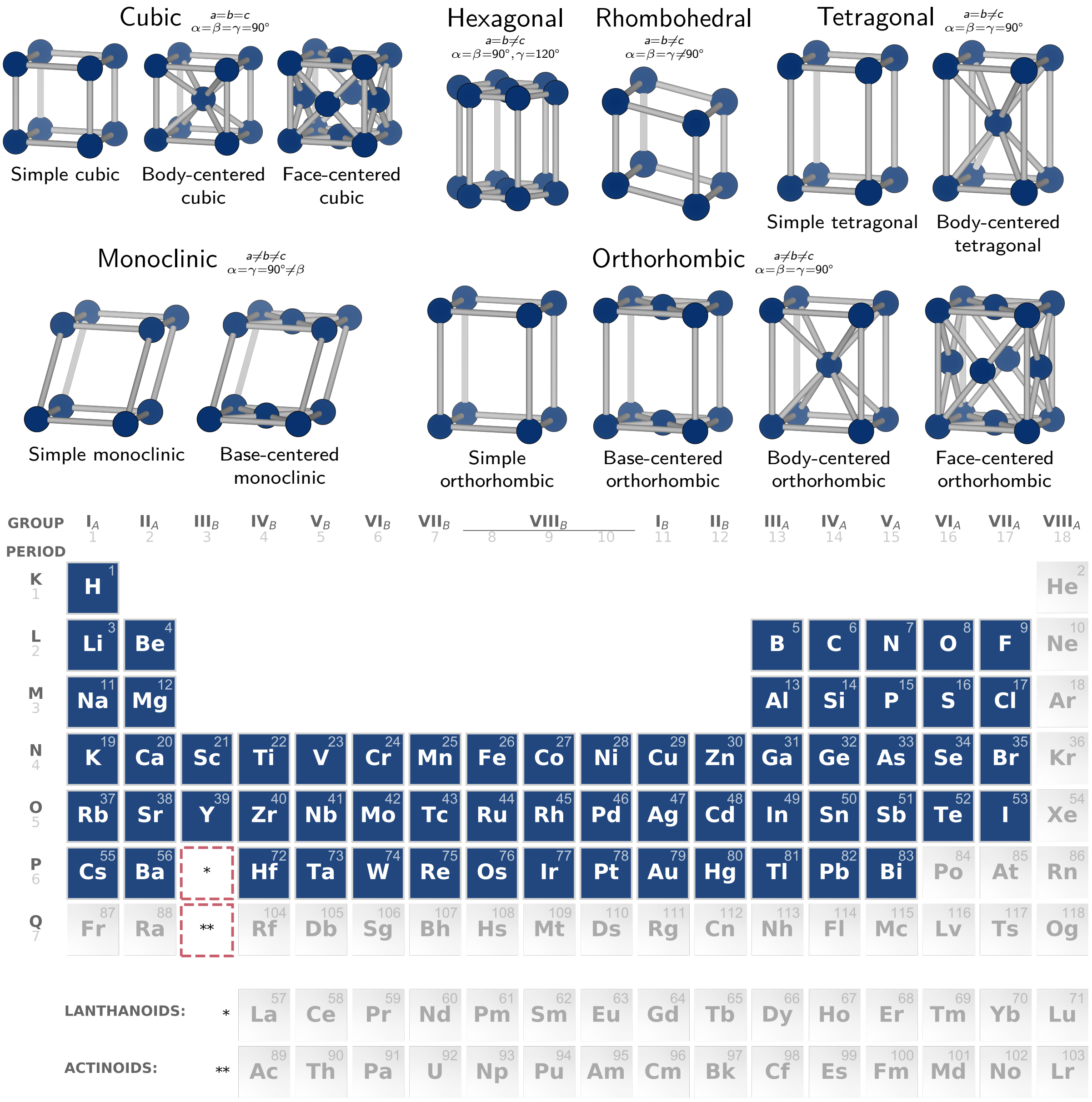

Cubic
$a=b=c$ $\alpha=\beta=\gamma=90°$
Simple cubic
Body-centered cubic
Face-centered cubic
Hexagonal
$a=b\neq c$ $\alpha=\beta=90°, \gamma=120°$
Rhombohedral
$a=b\neq c$ $\alpha=\beta=\gamma\neq 90°$
Tetragonal
$a=b\neq c$ $\alpha=\beta=\gamma=90°$
Simple tetragonal
Body-centered tetragonal
Monoclinic
$a\neq b\neq c$ $\alpha=\gamma=90°\neq\beta$
Simple monoclinic
Base-centered monoclinic
Orthorhombic
$a\neq b\neq c$ $\alpha=\beta=\gamma=90°$
Simple orthorhombic
Base-centered orthorhombic
Body-centered orthorhombic
Face-centered orthorhombic
GROUP
PERIOD
I$_A$ 1, II$_A$ 2, III$_B$ 3, IV$_B$ 4, V$_B$ 5, VI$_B$ 6, VII$_B$ 7, VIII$_B$ 8 9 10, I$_B$ 11, II$_B$ 12, III$_A$ 13, IV$_A$ 14, V$_A$ 15, VI$_A$ 16, VII$_A$ 17, VIII$_A$ 18
K 1: H 1, He 2
L 2: Li 3, Be 4, B 5, C 6, N 7, O 8, F 9, Ne 10
M 3: Na 11, Mg 12, Al 13, Si 14, P 15, S 16, Cl 17, Ar 18
N 4: K 19, Ca 20, Sc 21, Ti 22, V 23, Cr 24, Mn 25, Fe 26, Co 27, Ni 28, Cu 29, Zn 30, Ga 31, Ge 32, As 33, Se 34, Br 35, Kr 36
O 5: Rb 37, Sr 38, Y 39, Zr 40, Nb 41, Mo 42, Tc 43, Ru 44, Rh 45, Pd 46, Ag 47, Cd 48, In 49, Sn 50, Sb 51, Te 52, I 53, Xe 54
P 6: Cs 55, Ba 56, *, Hf 72, Ta 73, W 74, Re 75, Os 76, Ir 77, Pt 78, Au 79, Hg 80, Tl 81, Pb 82, Bi 83, Po 84, At 85, Rn 86
Q 7: Fr 87, Ra 88, **, Rf 104, Db 105, Sg 106, Bh 107, Hs 108, Mt 109, Ds 110, Rg 111, Cn 112, Nh 113, Fl 114, Mc 115, Lv 116, Ts 117, Og 118
LANTHANOIDS: * La 57, Ce 58, Pr 59, Nd 60, Pm 61, Sm 62, Eu 63, Gd 64, Tb 65, Dy 66, Ho 67, Er 68, Tm 69, Yb 70, Lu 71
ACTINOIDS: ** Ac 89, Th 90, Pa 91, U 92, Np 93, Pu 94, Am 95, Cm 96, Bk 97, Cf 98, Es 99, Fm 100, Md 101, No 102, Lr 103

Fig. 5.18. Crystal structures and used constituent elements of the 1,181 inorganic crystalline compounds in the data set [412].

values with an error of less than 15% [412]. Subsequent studies [165, 166, 473] reported similar agreements with the experimental values and demonstrated the applicability of machine learning to estimate the bulk and shear modulus based on the elemental properties of the inorganic crystalline compounds. The inorganic crystalline compounds data set is therefore ideally suited to investigate different feature-identification methods and to relate the statistical relationships in the data to the underlying elastic materials behavior through machine learning.

Although it is known that many elemental properties such as the thermal conductivity [165, 462, 463], chemical hardness [434, 486, 487], the atomic weight [488], and volume [165, 487, 489, 490] correlate well with the bulk and shear modulus, it is typically unclear which features sufficiently explain the diversity of the materials for a given data set. In general, promising features of the data set must be capable of both uniquely characterizing the diversity of compounds and relating the essential physics and chemistry to the properties of interest (cf., Section 2.4). For instance, de Jong, Chen *et al.* used 18 features of compositional (group number, row number, atomic mass, ...), structural (density, ...), and calculated properties from DFT calculations such as the cohesive and formation energy [165]. Furmanchuk, Agrawal *et al.* considered fundamental and experimentally measured properties of pure elements in their crystalline states [148, 149] and gradually extracted 50 out of 384 features [473]. Finally, Isayev, Oses *et al.* incorporated structural properties of the compound and utilized 21 elemental, measured, and derived elemental properties [479] to construct pairwise combinations of features and to compute a series of statistical attributes based on elemental properties (such as the mean absolute deviation, minimum, maximum, sum, and mean of the constituent elements [56, 188]) leading to a total of 2,494 features [166].

Analogous to Refs. [166, 473], the features of the investigated data set are based on well-tabulated fundamental, compositional, and structural properties for both uniquely characterizing the diverse range of compounds and estimating the bulk and shear modulus in the VRH notation [485] (Tab. 5.14). However, in contrast to Refs. [166, 473], the present analysis neither requires n-wise combinations of features nor features from experimentally demanding and computationally intensive first-principles calculations. In total, features of fundamental properties include the location of the element in the periodic table [188, 478, 479], the Mendeleev [491] (MN), group (g_P), and period numbers (p_P), the total number of valence electrons (n) and unfilled valence orbitals ($\bar{n}$) as well as those specified by their s-, p-, and d-character. The features of the compositional properties include calculated, measured, and derived properties of the compound such as the highest-occupied or lowest-unoccupied molecular orbitals (H, L) obtained from the non-spin polarized density-functional theory with the PBE exchange-correlation functional [482] (cf., Section 2.2).Measured properties include the atomic weight (m_atom) [477, 492, 493], electron affinity (EA) [477, 479, 494–497], absolute electronegativity (EN) [498], thermal conductivity (λ) [478, 479], heat capacity (C) [478, 479], heat of formation (H_form) [479], enthalpies of atomization (ΔH_at), fusion (ΔH_fusion) and vaporization (ΔH_vapor) [478], the first three ionization potentials ($IP_{1,2,3}$) [499], bulk (B) [151, 478, 489] and Young's modulus Y [151, 478]. And derived properties include the effective nuclear charge (Z_eff) [477, 500], molar volume (V_mol) [477], chemical hardness (η) [498, 501], covalent (r_cov) [502], absolute (r_abs) [503, 504], and van-der-Waals radii (r_vdw) [479], dipole polarizability (χ) [505]. In addition, features of structural properties were obtained from the crystal-structure of the inorganic compounds [412] such as the lattice parameters (a, b, c), ratios (a/b, b/c, a/c) and angles (α, β, γ), density (ρ), weight (m), volume (V), volume per atom (V_atom), number of atoms (N), number of species (X), total number of electrons (e), and number of electrons per atom (e_atom).

The features of the data set were constructed from the variable number of constituent elements of the compounds. As such, there were many different contraction and weighting schemes to map the elemental properties of the constituent elements to the one-dimensional features of the compounds

Property name	Symbol	Unit	Reference
Fundamental properties			
Mendeleev number	MN		[491]
Group and perdiod numbers	g_P, p_P		[188, 478, 479]
Total number of valence electrons	n		[188, 478, 479]
Total number of unfilled valence orbitals	$\bar{n}$		[188, 478, 479]
Number of valence electrons of the s-, p-, and d-orbitals	n_s, n_p, n_d		[188, 478, 479]
Number of unfilled valence s-, p-, and d-orbitals	$\bar{n}_s$, $\bar{n}_p$, $\bar{n}_d$		[188, 478, 479]
Compositional properties (weighted arithmetic mean of constituent elements)			
Highest-occupied molecular orbital	H	eV	DFT+PBE
Lowest-unoccupied molecular orbital	L	eV	DFT+PBE
Radius of the maximum electronic density for the valence s-, p-, and d-orbitals	r_s, r_p, r_d	Å	DFT+PBE
Atomic weight	m_atom	u	[477, 492, 493]
Atomic volume	V_atom	$\mathrm{Å}^3$	[477]
Atomic bulk modulus	B	GPa	[151, 478, 489]
Atomic Young's modulus	Y	GPa	[151, 478]
Electron affinity	EA	eV	[477, 479, 494–497]
Absolute electronegativity	EN	eV	[498]
Ionization potential	IP_1, IP_2, IP_3	eV	[499]
Heat of formation	H_form	$\mathrm{kJ\,mol^{-1}}$	[478, 479]
Atomization enthalpy	ΔH_at	$\mathrm{kJ\,mol^{-1}}$	[478, 479]
Evaporation enthalpy	ΔH_vapor	$\mathrm{kJ\,mol^{-1}}$	[478, 479]
Fusion enthalpy	ΔH_fusion	$\mathrm{kJ\,mol^{-1}}$	[478, 479]
Compositional properties (continued...)			
Thermal conductivity	λ	$\mathrm{W\,m^{-1}\,K^{-1}}$	[478, 479]
Heat capacity	C	$\mathrm{J\,kg^{-1}\,K^{-1}}$	[478, 479]
Effective nuclear charge	Z_eff	eV	[477, 500]
Chemical hardness	η	eV	[498, 501]
Dipole polarizability	χ	a.u.	[505]
Absolute radius	r_abs	Å	[503, 504]
Covalent radius	r_cov	Å	[502]
van-der-Waals radius	r_vdw	Å	[479]
Structural properties (extracted from the POSCAR files of [412])			
Lattice angles	α, β, γ	°	[412]
Lattice constants	a, b, c	Å	[412]
Lattice ratios	a/b, b/c, a/c		[412]
Density	ρ	$\mathrm{u/Å^3}$	[412]
Weight	m	u	[412]
Volume	V	$\mathrm{Å}^3$	[412]
Molar volume	V_mol	$\mathrm{Å^3/mol}$	[412]
Number of atoms	N		[412]
Number of atomic species	X		[412]
Total number of electrons	e		[412]
Number of electrons per atom	e_atom		[412]
Target properties			
Bulk modulus (VRH [485])	K_VRH	GPa	[412]
Shear modulus (VRH [485])	G_VRH	GPa	[412]

Tab. 5.14. List of all 54 features for the prediction of the bulk (K_VRH) and shear modulus (V_VRH) in the VRH notation. Fundamental and compositional features were taken from the Python package `mendeleev` [477], the weighted arithmetic mean were computed with the `matminer` package [189], and structural features were obtained from the `pymatgen` package [151]. Compositional features calculated from elemental properties were computed with the non-spin polarized density-functional theory and the PBE exchange-correlation functional (DFT+PBE) [482].

for each property. In this thesis, compositional properties were averaged and weighted[22] by the stoichiometry of the constituent elements [56, 166, 188][23]. Other contraction and weighting schemes were tested and added as features to the data set, i.e., harmonic, geometric, or quadratic means [506] and statistical properties such as the minimum or maximum value of the elemental property [56, 188] – the additional features resulted in an approximately quadratic increase in the computational requirements. It further led to larger feature-subset cardinalities of the identified feature subsets. However, the additional features did not substantially improve the prediction performance of the constructed statistical models, as many of the contraction and weighing schemes can be interrelated (e.g., the standard deviation with the mean of the atomic properties and the arithmetic mean together with

[22] Weighted arithmetic mean: $\mathrm{avg}(\vec{x}) = \sum_i x_i w_i / |\vec{w}|$, where $\vec{x}$ is a feature and $\vec{w}$ are the weights.

[23] The use of a contraction or weighting scheme destroys any relationship between the properties of interest and the atomic properties of the compound. Compositional properties, therefore, cannot be related to the atomic properties of the constituent elements and have no explicit dependence on the atomic properties.

harmonic mean to the geometric mean). Therefore, even though these additional features may be of interest for the statistical modeling of the elastic behavior of the materials, only the weighted arithmetic mean of the elemental properties is used in the following to focus on the statistical identification of relevant feature subsets with the framework developed in Chapter 4 and Sec. 5.1.

Although a feature identification with the developed framework is not limited to the 54 features presented above (Tab. 5.14), the number of features is essentially limited by the comparison to other feature-identification methods. In particular, the larger number of compounds, materials classes, and features as compared to the octet-binary compound semiconductors and perovskites data sets (cf., Sections 5.2.1 and 5.2.2) prevented a feature identification with the FS-SISSO algorithm and the construction of the symbolic-regression models for the assessment of the final prediction performance. The main challenge of the data set was therefore to reduce the large number of features (and hence the model complexity of the generated symbolic-regression models) prior to machine learning with any of the investigated feature-identification methods from Section 5.1.2.

Because higher computational requirements prevented the generation of symbolic-regression models from the identified subset of features to estimate the final prediction performance of the machine-learning models, the final prediction performance of the feature-identification methods were approximated as follows: Given an identified feature subset $\mathcal{X}^*$ and an empty feature candidate set $\mathfrak{X} = \varnothing$, a symbolic-regression (SISSO) model with $\mathfrak{X}_i = \mathfrak{X} \cup \{X_i\}$ was constructed for each of the identified features $X_i \in \mathcal{X}^*$. Then, the candidate set was augmented by the feature X_i ($\mathfrak{X} = \mathfrak{X}_i$), whose symbolic-regression model resulted in the lowest root-mean-squared error, and the procedure was repeated until the prediction error finally ceased to decrease. The approximation can be understood as a greedy algorithm that systematically reduces the prediction error by iteratively selecting the candidate features from an identified set of relevant features. However, prediction performances of the generated machine-learning models cannot be directly related to previous studies, as these use different machine-learning algorithms (piecewise-constant models [166], local-polynomial regression [165, 412]), different data sets (Materials Project [165, 412], AFLOW [166]), different data for creation and testing, and features of the data set[24]. As such, reported prediction performances are limited to the comparison of the investigated feature-identification methods.

Feature identification

To identify the relevant features for the bulk and shear modulus in the VRH notation, tests were performed on three randomly chosen subsets of 295, 590, and 1, 181 inorganic crystalline compounds, different feature-identification methods taken from Chapter 3 (TCMI, RFECV, FS-GBDT, and FS-SISSO), and the developed feature-identification framework from Sec. 5.1 (TB3-GBDT and TB3-SISSO). Machine-learning models were finally built with the SISSO algorithm using fixed hyper-

[24] Using a local-polynomial regression algorithm [475], de Jong, Chen *et al.* [165] achieve a prediction error of about $K_{\text{VRH}} = 18.3\,\text{GPa}$ for the bulk modulus and $G_{\text{VRH}} = 79.7\,\text{GPa}$ for the shear modulus with a model based on 40 features of which some were derived from density-functional theory. Using a piecewise-constant machine-learning algorithm, de Jong, Chen *et al.* [166] generated a set of 2,494 (fundamental, structural, and compositional) features and a model with a prediction error of $K_{\text{VRH}} = 14.3\,\text{GPa}$ and $G_{\text{VRH}} = 18.4\,\text{GPa}$, respectively. In comparison, the lowest prediction errors of the generated machine-learning models with the TB3-algorithm are $K_{\text{VRH}}16.9\,\text{GPa}$ for the bulk modulus and $G_{\text{VRH}} = 18.8\,\text{GPa}$) for the shear modulus using at most 50 features (cf., Tabs. 5.15 and 5.16).

#	Dependence	Features		Performance	
	measure	#	Relevant features	R^2	RMSE [GPa]
295 samples	TCMI	9	n_s, n_p, Y, IP$_1$, H_form, α, β, γ, X	0.80±0.06	34.2±5.9
	RFECV	11	$\bar{n}$, L, V_atom, B, Y, H_form, ΔH_at, ΔH_vapor, ΔH_fusion, ρ, V_mol	0.95±0.03	17.4±4.9
	FS-GBDT	12	$\bar{n}$, L, V_atom, B, Y, H_form, ΔH_at, ΔH_vapor, ΔH_fusion, a, ρ, V_mol	0.94±0.03	17.3±4.9
	TB3-GBDT	14	g_P, p_P, $\boldsymbol{n}$, n_d, $\bar{n}$, $\boldsymbol{m}_\mathbf{atom}$, H_form, $\boldsymbol{\Delta H}_\mathbf{at}$, ΔH_vapor, ΔH_fusion, C, Z_eff, ρ, $\boldsymbol{V}_\mathbf{mol}$	0.96±0.02/ **0.94±0.03**	14.1±2.7/ **17.3±2.1**
	TB3-SISSO	15	$\boldsymbol{p}_\mathbf{P}$, n, n_d, $\bar{n}$, $\boldsymbol{m}_\mathbf{atom}$, V_atom, B, Y, H_form, $\boldsymbol{\Delta H}_\mathbf{at}$, ΔH_vapor, C, r_cov, $\boldsymbol{\rho}$, V_mol	0.96±0.01/ **0.95±0.01**	14.3±1.6/ **16.2±1.1**
590 samples	TCMI	5	n_p, $\bar{n}$, $\bar{n}_p$, B, X	0.73±0.05	38.7±4.3
	RFECV	12	V_atom, B, Y, EN, H_form, ΔH_at, ΔH_vapor, ΔH_fusion, λ, ρ, V, V_mol	0.93±0.02	20.2±2.9
	FS-GBDT	14	n, $\bar{n}_d$, V_atom, B, Y, H_form, ΔH_at, ΔH_vapor, ΔH_fusion, λ, Z_eff, χ, ρ, V_mol	0.93±0.02	19.0±2.7
	TB3-GBDT	35	MN, g_P, $\boldsymbol{p}_\mathbf{P}$, $\boldsymbol{n}$, n_p, n_d, $\bar{n}$, $\bar{n}_s$, $\bar{\boldsymbol{n}}_{\boldsymbol{p}}$, $\bar{n}_d$, H, r$_s$, r$_p$, r$_d$, m_atom, V_atom, B, EA, IP$_1$, IP$_3$, H_form, ΔH_at, ΔH_vapor, ΔH_fusion, λ, $\boldsymbol{C}$, Z_eff, r_abs, r_cov, r_vdw, a, ρ, V, $\boldsymbol{V}_\mathbf{mol}$, e_atom	0.95±0.02/ **0.92±0.03**	16.0±3.1/ **21.0±3.9**
	TB3-SISSO	17	$\boldsymbol{p}_\mathbf{P}$, n, n_d, r$_s$, m_atom, $\boldsymbol{V}_\mathbf{atom}$, B, H_form, ΔH_at, $\boldsymbol{\Delta H}_\mathbf{vapor}$, ΔH_fusion, C, η, r_abs, r_vdw, ρ, $\boldsymbol{V}_\mathbf{mol}$	0.95±0.01/ **0.93±0.02**	16.1±2.4/ **20.0±2.4**
1181 samples	TCMI	5	n_s, L, IP$_2$, ΔH_at, Z_eff	0.72±0.06	38.3±4.2
	RFECV	11	V_atom, B, Y, H_form, ΔH_at, ΔH_vapor, ΔH_fusion, λ, C, ρ, V_mol	0.93±0.01	18.7±2.6
	FS-GBDT	17	n, H, V_atom, B, Y, IP$_2$, IP$_3$, H_form, ΔH_at, ΔH_vapor, ΔH_fusion, λ, η, r_abs, a, ρ, V_mol	0.94±0.01	18.3±2.2
	TB3-GBDT	38	MN, g_P, p_P, n, n_s, n_p, $\boldsymbol{n}_{\boldsymbol{d}}$, $\bar{n}$, $\bar{n}_s$, $\bar{n}_p$, $\bar{n}_d$, H, L, r$_s$, r$_p$, r$_d$, m_atom, V_atom, B, Y, EA, EN, IP$_1$, IP$_2$, IP$_3$, H_form, $\boldsymbol{\Delta H}_\mathbf{at}$, ΔH_vapor, ΔH_fusion, λ, C, $\boldsymbol{Z}_\mathbf{eff}$, η, r_abs, r_cov, r_vdw, $\boldsymbol{\rho}$, $\boldsymbol{V}_\mathbf{mol}$	0.94±0.03/ **0.93±0.02**	17.2±4.7/ **19.5±2.4**
	TB3-SISSO	20	$\boldsymbol{p}_\mathbf{P}$, n, n_p, n_d, $\bar{n}_p$, $\boldsymbol{m}_\mathbf{atom}$, V_atom, B, EN, IP$_3$, H_form, $\boldsymbol{\Delta H}_\mathbf{at}$, ΔH_vapor, ΔH_fusion, C, η, r_abs, r_vdw, $\boldsymbol{\rho}$, V_mol	0.95±0.01/ **0.93±0.02**	16.9±1.9/ **18.8±2.0**
Reference [166]				0.97	14.3
Bulk modulus			Stats: K_VRH = [6.5 , 435.7] GPa	mean = 136.3 GPa	std = 72.9 GPa

Tab. 5.15. Prediction performance of identified redundant feature subsets and optimal non-redundant feature subsets (bold) of different feature-identification methods for estimating the bulk modulus K_VRH in the Voigt-Reuss-Hill notation [485]: total cumulative mutual information (TCMI, Section 3.3.3), recursive feature elimination [263] using random forest (RFECV) [266], gradient-boosting decision trees using permutation feature importance (FS-GBDT) [252, 266, 338], and the tolerance-based branch-and-bound algorithm (TB3) with GBDT as feature-selection criterion (Section 4.1). Prediction performances were estimated with SISSO by means of 10-fold cross-validation (cf., Section 5.1) using the approximation procedure as described in Section 5.2.3. Shown are the mean $\bar{x}$ and standard deviation Δx as $\bar{x} \pm \Delta x$ of the Pearson's coefficient of determination $x = R^2$ [101] and the root-mean-squared error (x = RMSE). The RMSE is in units of gigapascal (1 GPa = 10×10^9 N m^{-2}). The reported prediction performance from Ref. [166] is also shown.

#	**Dependence measure**	**Features** #	Relevant features	**Performance** R^2	RMSE [GPa]
295 samples	TCMI	7	IP_2, IP_3, H_form, α, β, γ, X	0.60±0.15	29.0±5.1
	RFECV	49	MN, g_P, p_P, n, n_s, n_p, n_d, $\bar{n}$, $\bar{n}_s$, $\bar{n}_p$, $\bar{n}_d$, H, L, r_s, r_p, r_d, m_atom, V_atom, B, Y, EA, EN, IP_1, IP_2, IP_3, H_form, ΔH_at, ΔH_vapor, ΔH_fusion, λ, C, Z_eff, η, χ, r_abs, r_cov, r_vdw, a, b, c, a/c, b/c, ρ, m, V, V_mol, N, e, e_atom	0.80±0.11	19.0±5.0
	FS-GBDT	10	$\bar{n}_d$, Y, H_form, ΔH_at, ΔH_fusion, λ, c, m, V, V_mol	0.80±0.10	19.6±4.2
	TB3-GBDT	40	MN, g_P, p_P, n, n_s, n_p, n_d, $\bar{n}$, $\boldsymbol{\bar{n}_p}$, $\boldsymbol{\bar{n}_d}$, H, L, r_s, r_p, r_d, m_atom, $\boldsymbol{V_\mathrm{atom}}$, B, Y, EA, EN, IP_1, IP_2, H_form, ΔH_at, ΔH_vapor, ΔH_fusion, $\boldsymbol{\lambda}$, C, $\boldsymbol{Z_\mathrm{eff}}$, η, χ, r_abs, r_cov, $\boldsymbol{r_\mathrm{vdw}}$, a, b, ρ, V, $\boldsymbol{V_\mathrm{mol}}$	0.81±0.10/ **0.71±0.05**	18.5±4.0/ **28.3±7.9**
	TB3-SISSO	46	MN, g_P, p_P, n, n_s, n_p, n_d, $\bar{n}$, $\bar{n}_s$, $\boldsymbol{\bar{n}_p}$, $\bar{n}_d$, H, L, r_s, r_d, V_atom, B, Y, EA, EN, IP_1, IP_2, IP_3, $\boldsymbol{H_\mathrm{form}}$, ΔH_at, ΔH_vapor, $\boldsymbol{\Delta H_\mathrm{fusion}}$, C, Z_eff, η, r_abs, r_cov, r_vdw, a, b, c, a/b, a/c, b/c, ρ, m, V, $\boldsymbol{V_\mathrm{mol}}$, N, e, e_atom	0.77±0.14/ **0.81±0.09**	20.7±5.3/ **19.3±3.6**
590 samples	TCMI	4	H_form, C, α, e_atom	0.58±0.11	28.5±5.0
	RFECV	19	MN, n, n_d, $\bar{n}$, $\bar{n}_d$, r_p, V_atom, Y, EN, IP_2, H_form, ΔH_at, ΔH_vapor, ΔH_fusion, λ, c, ρ, V, V_mol	0.80±0.08	19.4±3.6
	FS-GBDT	20	MN, n_s, n_d, $\bar{n}$, $\bar{n}_d$, r_p, m_atom, V_atom, Y, EN, IP_2, H_form, ΔH_at, ΔH_vapor, ΔH_fusion, λ, c, m, V, V_mol	0.80±0.08	19.6±3.5
	TB3-GBDT	42	MN, g_P, $\boldsymbol{p_\mathrm{P}}$, n, n_s, $\boldsymbol{n_p}$, n_d, $\boldsymbol{\bar{n}}$, $\boldsymbol{\bar{n}_s}$, $\bar{n}_p$, $\bar{n}_d$, H, L, r_s, r_p, r_d, m_atom, V_atom, B, $\boldsymbol{Y}$, EA, EN, IP_1, $\mathbf{IP}_\mathbf{2}$, IP_3, H_form, ΔH_at, ΔH_vapor, $\boldsymbol{\Delta H_\mathrm{fusion}}$, λ, C, η, r_abs, r_cov, r_vdw, $\boldsymbol{c}$, a/b, a/c, b/c, ρ, V, $\boldsymbol{V_\mathrm{mol}}$	0.83±0.05/ **0.76±0.07**	18.4±2.5/ **21.5±2.5**
	TB3-SISSO	51	MN, g_P, $\boldsymbol{p_\mathrm{P}}$, n, n_s, n_p, $\boldsymbol{n_d}$, $\bar{n}$, $\bar{n}_s$, $\bar{n}_p$, $\boldsymbol{\bar{n}_d}$, H, L, r_s, r_p, r_d, m_atom, V_atom, B, Y, EA, EN, IP_1, IP_2, IP_3, H_form, ΔH_at, ΔH_vapor, ΔH_fusion, λ, C, Z_eff, η, χ, r_abs, r_cov, α, β, γ, a, a/b, a/c, b/c, ρ, m, V, $\boldsymbol{V_\mathrm{mol}}$, N, X, e, e_atom	0.78±0.09/ **0.79±0.05**	21.3±6.3/ **20.0±3.1**
1181 samples	TCMI	6	L, IP_2, ΔH_at, α, β, γ	0.49±0.06	31.7±3.9
	RFECV	20	MN, n_d, $\bar{n}$, $\bar{n}_d$, L, r_p, V_atom, B, Y, EN, IP_2, H_form, ΔH_at, ΔH_fusion, λ, r_abs, a, c, ρ, V_mol	0.82±0.04	19.0±2.8
	FS-GBDT	6	$\bar{n}_d$, V_atom, Y, ΔH_at, ΔH_fusion, V_mol	0.77±0.05	20.8±2.0
	TB3-GBDT	50	**MN**, g_P, p_P, n, n_s, n_p, $\boldsymbol{n_d}$, $\bar{n}$, $\bar{n}_s$, $\bar{n}_p$, $\bar{n}_d$, H, L, r_s, r_p, r_d, m_atom, V_atom, B, $\boldsymbol{Y}$, EA, EN, IP_1, IP_2, IP_3, H_form, $\boldsymbol{\Delta H_\mathrm{at}}$, ΔH_vapor, ΔH_fusion, λ, C, Z_eff, $\boldsymbol{\eta}$, χ, r_cov, r_vdw, a, b, $\boldsymbol{c}$, a/b, a/c, b/c, $\boldsymbol{\rho}$, m, V, $\boldsymbol{V_\mathrm{mol}}$, $\boldsymbol{N}$, X, e, e_atom	0.82±0.03/ **0.77±0.04**	18.8±2.2/ **21.0±3.1**
	TB3-SISSO	50	MN, g_P, p_P, n, n_s, n_p, n_d, $\bar{n}$, $\bar{n}_s$, $\bar{n}_p$, $\bar{n}_d$, H, L, r_s, r_p, r_d, m_atom, $\boldsymbol{V_\mathrm{atom}}$, B, Y, EA, EN, IP_1, IP_2, IP_3, $\boldsymbol{H_\mathrm{form}}$, ΔH_at, ΔH_vapor, ΔH_fusion, λ, $\boldsymbol{C}$, Z_eff, η, χ, r_abs, r_cov, r_vdw, α, β, γ, b, c, a/b, a/c, ρ, m, V, $\boldsymbol{V_\mathrm{mol}}$, X, e	0.82±0.03/ **0.77±0.05**	18.8±2.2/ **20.9±1.8**
Reference [166]				0.88	18.4
Shear modulus	Stats: G_VRH = [2.7 , 523] GPa		mean = 67.6 GPa		std = 44.6 GPa

Tab. 5.16. Prediction performance of identified redundant feature subsets and optimal non-redundant feature subsets (bold) of different feature-identification methods for estimating the shear modulus G_VRH in the Voigt-Reuss-Hill notation [485]: total cumulative mutual information (TCMI, Section 3.3.3), recursive feature elimination [263] using random forest (RFECV) [266], gradient-boosting decision trees using permutation feature importance (FS-GBDT) [252, 266, 338], and the tolerance-based branch-and-bound algorithm (TB3) with GBDT as feature-selection criterion (Section 4.1). Prediction performances were estimated with SISSO by means of 10-fold cross-validation (cf., Section 5.1) using the approximation procedure as described in Section 5.2.3. Shown are the mean $\bar{x}$ and standard deviation Δx as $\bar{x} \pm \Delta x$ of the Pearson's coefficient of determination $x = R^2$ [101] and the root-mean-squared error (x = RMSE). The RMSE is in units of gigapascal (1 GPa = $10 \times 10^9\,\mathrm{N\,m^{-2}}$). The reported prediction performance from Ref. [166] is also shown.

parameter settings [197] (cf., Sec. 5.1.2 and Appendix A.1). Results are reported in Tables 5.15 and 5.16.

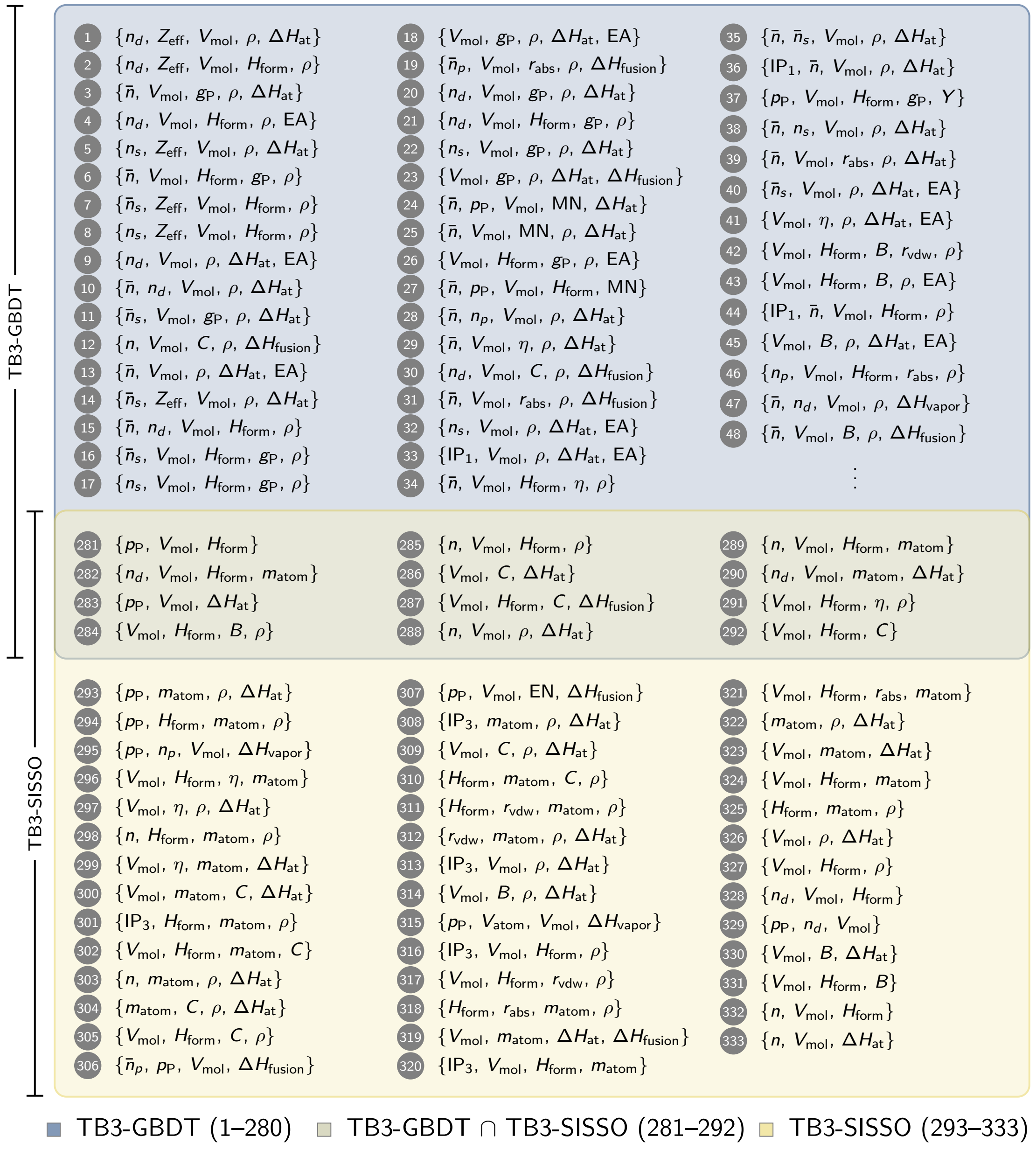
Of all the feature-identification methods, TCMI, FS-SISSO, and TB3-SISSO were particularly computer intensive. Whereas TB3-SISSO completed within one week using 32 nodes[25] (cf., Sections 5.2.1 and 5.2.2), but required a limitation of the maximum cardinality of the feature subsets (number of features ≤ 4, cf., Section 3.4) as like TCMI, FS-SISSO failed to complete due to high computer and memory requirements. Technically, the limitation of the search depth may be a viable tool to adjust the computational requirements of the developed feature-identification framework and the underlying branch-and-bound algorithm – especially when other feature-identification methods are no longer applicable as in the case of TB3-SISSO. However, as can be seen from the example of TCMI, limiting the search depth can also lead to much larger prediction errors and a potential incorrect identification of relevant features.

It is to be noted that the transferability of identified (sub-optimal) feature subsets is limited when different machine-learning algorithms are applied for the search and model construction. For example, a comparison shows that in all 292 non-redundant feature subsets identified by TB3-GBDT and 53 non-redundant feature subsets identified by TB3-SISSO, only 12 of the identified feature subsets are found by both methods (Figs. 5.19 and 5.20). Due to different cardinalities of the feature subsets, this corresponds to a Jaccard similarity coefficient of 0.04. Again, the low Jaccard similarity coefficient is indicative of strong multivariate dependencies in the data set and a model-dependent identification of feature subsets (cf., Sections 5.2.1 and 5.2.2).

Results further show that all feature-identification methods are characterized by a high variability in the identification of relevant features. For example, TCMI identified about 5–9 features, RFECV 11–49 features, and FS-GBDT 6–20 features, i.e., they identified a few to almost all features as relevant for K_{VRH} or G_{VRH}. While redundant feature subsets of TB3-GBDT and TB3-SISSO also varied significantly (TB3-GBDT: 14–50 features, TB3-SISSO: 15–51), the optimal non-redundant minimally feature subsets were more robust to the varying number of data samples (4–8 features).

A frequency analysis (Fig. 5.21) shows that the atomization enthalpy (ΔH_{at}), the density (ρ), and the molar volume (V_{mol}) are the most frequently identified features for the bulk modulus K_{VRH}. Likewise, the most frequently identified features for the shear modulus G_{VRH} are the atomization enthalpy (ΔH_{at}), the atomic volume (V_{atom}), Young's modulus (Y), and V_{mol}. That is, taken together, the frequency analyses are in agreement with most of the trends reported in prior works on investigating the bulk and shear modulus of specific families of materials (cf., [165, 166, 473]). For example, the averaged Bulk (K) and Young's modulus (Y) of the constituent elements are known to be physical related to K_{VRH} and G_{VRH}. It is also known that the bulk and shear modulus are strongly statistical related to the atomic volume [507], the bond length [434, 508, 509], and to the lattice constants [510]. And it is known that the atomic properties, that are logically linked to the chemical bonding of the materials (molar volume [166, 473] and heat of a material), are relevant and strongly correlated to the bulk and shear modulus [166].

[25] Feature identification was performed on 8 nodes, each with two Intel Xeon E5-2698 v3 processors (= 32 cores/node) without hyper-threading.



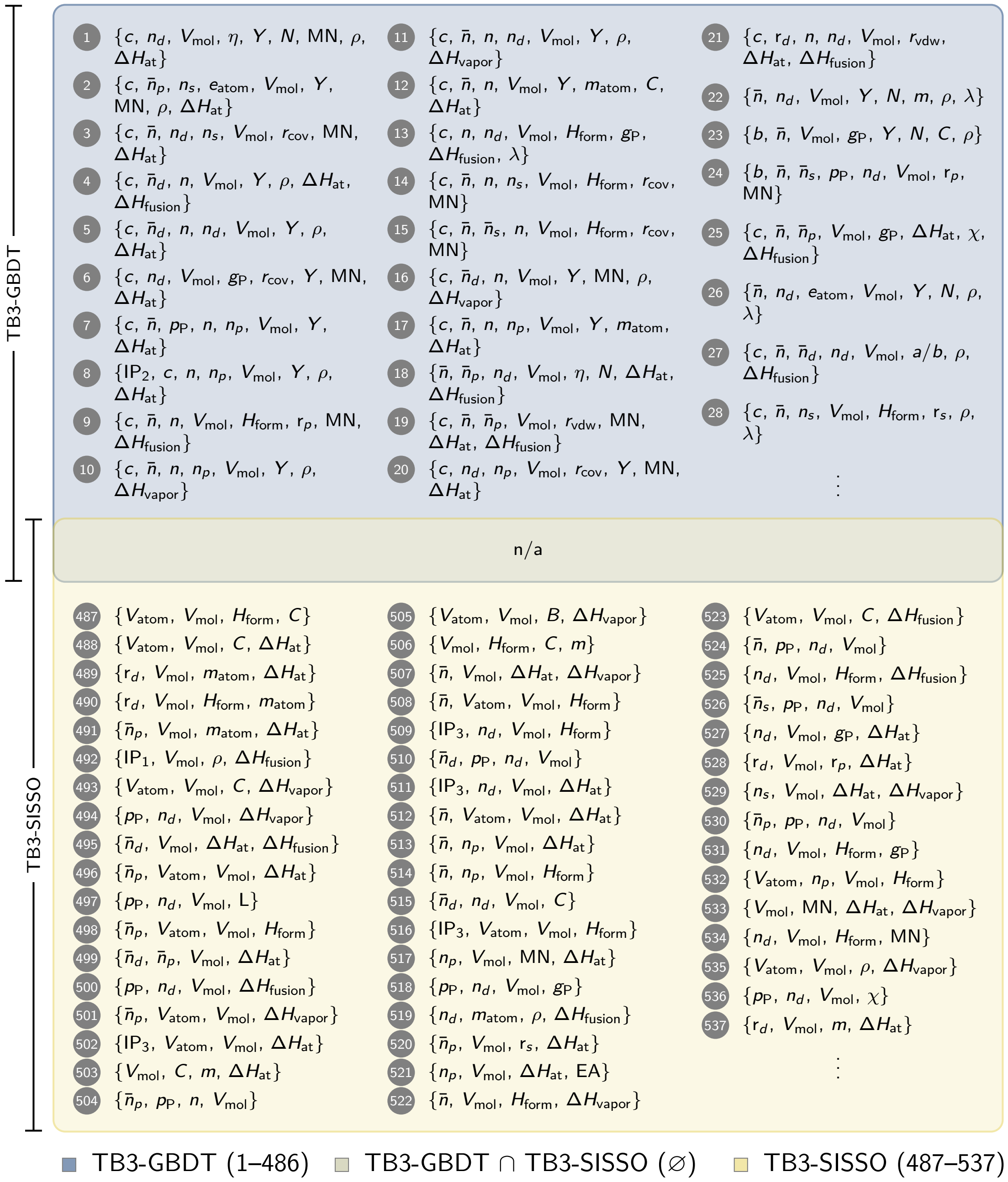
1 $\{n_d, Z_{\mathrm{eff}}, V_{\mathrm{mol}}, \rho, \Delta H_{\mathrm{at}}\}$
2 $\{n_d, Z_{\mathrm{eff}}, V_{\mathrm{mol}}, H_{\mathrm{form}}, \rho\}$
3 $\{\bar{n}, V_{\mathrm{mol}}, g_{\mathrm{P}}, \rho, \Delta H_{\mathrm{at}}\}$
4 $\{n_d, V_{\mathrm{mol}}, H_{\mathrm{form}}, \rho, \mathrm{EA}\}$
5 $\{n_s, Z_{\mathrm{eff}}, V_{\mathrm{mol}}, \rho, \Delta H_{\mathrm{at}}\}$
6 $\{\bar{n}, V_{\mathrm{mol}}, H_{\mathrm{form}}, g_{\mathrm{P}}, \rho\}$
7 $\{\bar{n}_s, Z_{\mathrm{eff}}, V_{\mathrm{mol}}, H_{\mathrm{form}}, \rho\}$
8 $\{n_s, Z_{\mathrm{eff}}, V_{\mathrm{mol}}, H_{\mathrm{form}}, \rho\}$
9 $\{n_d, V_{\mathrm{mol}}, \rho, \Delta H_{\mathrm{at}}, \mathrm{EA}\}$
10 $\{\bar{n}, n_d, V_{\mathrm{mol}}, \rho, \Delta H_{\mathrm{at}}\}$
11 $\{\bar{n}_s, V_{\mathrm{mol}}, g_{\mathrm{P}}, \rho, \Delta H_{\mathrm{at}}\}$
12 $\{n, V_{\mathrm{mol}}, C, \rho, \Delta H_{\mathrm{fusion}}\}$
13 $\{\bar{n}, V_{\mathrm{mol}}, \rho, \Delta H_{\mathrm{at}}, \mathrm{EA}\}$
14 $\{\bar{n}_s, Z_{\mathrm{eff}}, V_{\mathrm{mol}}, \rho, \Delta H_{\mathrm{at}}\}$
15 $\{\bar{n}, n_d, V_{\mathrm{mol}}, H_{\mathrm{form}}, \rho\}$
16 $\{\bar{n}_s, V_{\mathrm{mol}}, H_{\mathrm{form}}, g_{\mathrm{P}}, \rho\}$
17 $\{n_s, V_{\mathrm{mol}}, H_{\mathrm{form}}, g_{\mathrm{P}}, \rho\}$
18 $\{V_{\mathrm{mol}}, g_{\mathrm{P}}, \rho, \Delta H_{\mathrm{at}}, \mathrm{EA}\}$
19 $\{\bar{n}_p, V_{\mathrm{mol}}, r_{\mathrm{abs}}, \rho, \Delta H_{\mathrm{fusion}}\}$
20 $\{n_d, V_{\mathrm{mol}}, g_{\mathrm{P}}, \rho, \Delta H_{\mathrm{at}}\}$
21 $\{n_d, V_{\mathrm{mol}}, H_{\mathrm{form}}, g_{\mathrm{P}}, \rho\}$
22 $\{n_s, V_{\mathrm{mol}}, g_{\mathrm{P}}, \rho, \Delta H_{\mathrm{at}}\}$
23 $\{V_{\mathrm{mol}}, g_{\mathrm{P}}, \rho, \Delta H_{\mathrm{at}}, \Delta H_{\mathrm{fusion}}\}$
24 $\{\bar{n}, p_{\mathrm{P}}, V_{\mathrm{mol}}, \mathrm{MN}, \Delta H_{\mathrm{at}}\}$
25 $\{\bar{n}, V_{\mathrm{mol}}, \mathrm{MN}, \rho, \Delta H_{\mathrm{at}}\}$
26 $\{V_{\mathrm{mol}}, H_{\mathrm{form}}, g_{\mathrm{P}}, \rho, \mathrm{EA}\}$
27 $\{\bar{n}, p_{\mathrm{P}}, V_{\mathrm{mol}}, H_{\mathrm{form}}, \mathrm{MN}\}$
28 $\{\bar{n}, n_p, V_{\mathrm{mol}}, \rho, \Delta H_{\mathrm{at}}\}$
29 $\{\bar{n}, V_{\mathrm{mol}}, \eta, \rho, \Delta H_{\mathrm{at}}\}$
30 $\{n_d, V_{\mathrm{mol}}, C, \rho, \Delta H_{\mathrm{fusion}}\}$
31 $\{\bar{n}, V_{\mathrm{mol}}, r_{\mathrm{abs}}, \rho, \Delta H_{\mathrm{fusion}}\}$
32 $\{n_s, V_{\mathrm{mol}}, \rho, \Delta H_{\mathrm{at}}, \mathrm{EA}\}$
33 $\{\mathrm{IP}_1, V_{\mathrm{mol}}, \rho, \Delta H_{\mathrm{at}}, \mathrm{EA}\}$
34 $\{\bar{n}, V_{\mathrm{mol}}, H_{\mathrm{form}}, \eta, \rho\}$
35 $\{\bar{n}, \bar{n}_s, V_{\mathrm{mol}}, \rho, \Delta H_{\mathrm{at}}\}$
36 $\{\mathrm{IP}_1, \bar{n}, V_{\mathrm{mol}}, \rho, \Delta H_{\mathrm{at}}\}$
37 $\{p_{\mathrm{P}}, V_{\mathrm{mol}}, H_{\mathrm{form}}, g_{\mathrm{P}}, Y\}$
38 $\{\bar{n}, n_s, V_{\mathrm{mol}}, \rho, \Delta H_{\mathrm{at}}\}$
39 $\{\bar{n}, V_{\mathrm{mol}}, r_{\mathrm{abs}}, \rho, \Delta H_{\mathrm{at}}\}$
40 $\{\bar{n}_s, V_{\mathrm{mol}}, \rho, \Delta H_{\mathrm{at}}, \mathrm{EA}\}$
41 $\{V_{\mathrm{mol}}, \eta, \rho, \Delta H_{\mathrm{at}}, \mathrm{EA}\}$
42 $\{V_{\mathrm{mol}}, H_{\mathrm{form}}, B, r_{\mathrm{vdw}}, \rho\}$
43 $\{V_{\mathrm{mol}}, H_{\mathrm{form}}, B, \rho, \mathrm{EA}\}$
44 $\{\mathrm{IP}_1, \bar{n}, V_{\mathrm{mol}}, H_{\mathrm{form}}, \rho\}$
45 $\{V_{\mathrm{mol}}, B, \rho, \Delta H_{\mathrm{at}}, \mathrm{EA}\}$
46 $\{n_p, V_{\mathrm{mol}}, H_{\mathrm{form}}, r_{\mathrm{abs}}, \rho\}$
47 $\{\bar{n}, n_d, V_{\mathrm{mol}}, \rho, \Delta H_{\mathrm{vapor}}\}$
48 $\{\bar{n}, V_{\mathrm{mol}}, B, \rho, \Delta H_{\mathrm{fusion}}\}$
⋮

281 $\{p_{\mathrm{P}}, V_{\mathrm{mol}}, H_{\mathrm{form}}\}$
282 $\{n_d, V_{\mathrm{mol}}, H_{\mathrm{form}}, m_{\mathrm{atom}}\}$
283 $\{p_{\mathrm{P}}, V_{\mathrm{mol}}, \Delta H_{\mathrm{at}}\}$
284 $\{V_{\mathrm{mol}}, H_{\mathrm{form}}, B, \rho\}$
285 $\{n, V_{\mathrm{mol}}, H_{\mathrm{form}}, \rho\}$
286 $\{V_{\mathrm{mol}}, C, \Delta H_{\mathrm{at}}\}$
287 $\{V_{\mathrm{mol}}, H_{\mathrm{form}}, C, \Delta H_{\mathrm{fusion}}\}$
288 $\{n, V_{\mathrm{mol}}, \rho, \Delta H_{\mathrm{at}}\}$
289 $\{n, V_{\mathrm{mol}}, H_{\mathrm{form}}, m_{\mathrm{atom}}\}$
290 $\{n_d, V_{\mathrm{mol}}, m_{\mathrm{atom}}, \Delta H_{\mathrm{at}}\}$
291 $\{V_{\mathrm{mol}}, H_{\mathrm{form}}, \eta, \rho\}$
292 $\{V_{\mathrm{mol}}, H_{\mathrm{form}}, C\}$

293 $\{p_{\mathrm{P}}, m_{\mathrm{atom}}, \rho, \Delta H_{\mathrm{at}}\}$
294 $\{p_{\mathrm{P}}, H_{\mathrm{form}}, m_{\mathrm{atom}}, \rho\}$
295 $\{p_{\mathrm{P}}, n_p, V_{\mathrm{mol}}, \Delta H_{\mathrm{vapor}}\}$
296 $\{V_{\mathrm{mol}}, H_{\mathrm{form}}, \eta, m_{\mathrm{atom}}\}$
297 $\{V_{\mathrm{mol}}, \eta, \rho, \Delta H_{\mathrm{at}}\}$
298 $\{n, H_{\mathrm{form}}, m_{\mathrm{atom}}, \rho\}$
299 $\{V_{\mathrm{mol}}, \eta, m_{\mathrm{atom}}, \Delta H_{\mathrm{at}}\}$
300 $\{V_{\mathrm{mol}}, m_{\mathrm{atom}}, C, \Delta H_{\mathrm{at}}\}$
301 $\{\mathrm{IP}_3, H_{\mathrm{form}}, m_{\mathrm{atom}}, \rho\}$
302 $\{V_{\mathrm{mol}}, H_{\mathrm{form}}, m_{\mathrm{atom}}, C\}$
303 $\{n, m_{\mathrm{atom}}, \rho, \Delta H_{\mathrm{at}}\}$
304 $\{m_{\mathrm{atom}}, C, \rho, \Delta H_{\mathrm{at}}\}$
305 $\{V_{\mathrm{mol}}, H_{\mathrm{form}}, C, \rho\}$
306 $\{\bar{n}_p, p_{\mathrm{P}}, V_{\mathrm{mol}}, \Delta H_{\mathrm{fusion}}\}$
307 $\{p_{\mathrm{P}}, V_{\mathrm{mol}}, \mathrm{EN}, \Delta H_{\mathrm{fusion}}\}$
308 $\{\mathrm{IP}_3, m_{\mathrm{atom}}, \rho, \Delta H_{\mathrm{at}}\}$
309 $\{V_{\mathrm{mol}}, C, \rho, \Delta H_{\mathrm{at}}\}$
310 $\{H_{\mathrm{form}}, m_{\mathrm{atom}}, C, \rho\}$
311 $\{H_{\mathrm{form}}, r_{\mathrm{vdw}}, m_{\mathrm{atom}}, \rho\}$
312 $\{r_{\mathrm{vdw}}, m_{\mathrm{atom}}, \rho, \Delta H_{\mathrm{at}}\}$
313 $\{\mathrm{IP}_3, V_{\mathrm{mol}}, \rho, \Delta H_{\mathrm{at}}\}$
314 $\{V_{\mathrm{mol}}, B, \rho, \Delta H_{\mathrm{at}}\}$
315 $\{p_{\mathrm{P}}, V_{\mathrm{atom}}, V_{\mathrm{mol}}, \Delta H_{\mathrm{vapor}}\}$
316 $\{\mathrm{IP}_3, V_{\mathrm{mol}}, H_{\mathrm{form}}, \rho\}$
317 $\{V_{\mathrm{mol}}, H_{\mathrm{form}}, r_{\mathrm{vdw}}, \rho\}$
318 $\{H_{\mathrm{form}}, r_{\mathrm{abs}}, m_{\mathrm{atom}}, \rho\}$
319 $\{V_{\mathrm{mol}}, m_{\mathrm{atom}}, \Delta H_{\mathrm{at}}, \Delta H_{\mathrm{fusion}}\}$
320 $\{\mathrm{IP}_3, V_{\mathrm{mol}}, H_{\mathrm{form}}, m_{\mathrm{atom}}\}$
321 $\{V_{\mathrm{mol}}, H_{\mathrm{form}}, r_{\mathrm{abs}}, m_{\mathrm{atom}}\}$
322 $\{m_{\mathrm{atom}}, \rho, \Delta H_{\mathrm{at}}\}$
323 $\{V_{\mathrm{mol}}, m_{\mathrm{atom}}, \Delta H_{\mathrm{at}}\}$
324 $\{V_{\mathrm{mol}}, H_{\mathrm{form}}, m_{\mathrm{atom}}\}$
325 $\{H_{\mathrm{form}}, m_{\mathrm{atom}}, \rho\}$
326 $\{V_{\mathrm{mol}}, \rho, \Delta H_{\mathrm{at}}\}$
327 $\{V_{\mathrm{mol}}, H_{\mathrm{form}}, \rho\}$
328 $\{n_d, V_{\mathrm{mol}}, H_{\mathrm{form}}\}$
329 $\{p_{\mathrm{P}}, n_d, V_{\mathrm{mol}}\}$
330 $\{V_{\mathrm{mol}}, B, \Delta H_{\mathrm{at}}\}$
331 $\{V_{\mathrm{mol}}, H_{\mathrm{form}}, B\}$
332 $\{n, V_{\mathrm{mol}}, H_{\mathrm{form}}\}$
333 $\{n, V_{\mathrm{mol}}, \Delta H_{\mathrm{at}}\}$

■ TB3-GBDT (1–280) ■ TB3-GBDT ∩ TB3-SISSO (281–292) ■ TB3-SISSO (293–333)

Fig. 5.19. Identified minimally non-redundant (sub-optimal) feature subsets for the prediction of the bulk modulus K_{VRH} in the Voigt-Reuss-Hill notation [485] using the tolerance-based branch-and-bound algorithm (TB3) and GBDT or SISSO as the feature-selection criterion (Section 4.1). Shown is the intersection and the union of the two feature-identification methods corresponding to a Jaccard similarity coefficient (Eq. 5.1) of 0.04. It should be noted that due to the different machine-learning algorithms, a direct comparison between TB3-GBDT and TB3-SISSO may not be appropriate. The numbering is used for referencing purposes.



1 $\{c,\ n_d,\ V_{\mathrm{mol}},\ \eta,\ Y,\ N,\ \mathrm{MN},\ \rho,\ \Delta H_{\mathrm{at}}\}$
2 $\{c,\ \bar{n}_p,\ n_s,\ e_{\mathrm{atom}},\ V_{\mathrm{mol}},\ Y,\ \mathrm{MN},\ \rho,\ \Delta H_{\mathrm{at}}\}$
3 $\{c,\ \bar{n},\ n_d,\ n_s,\ V_{\mathrm{mol}},\ r_{\mathrm{cov}},\ \mathrm{MN},\ \Delta H_{\mathrm{at}}\}$
4 $\{c,\ \bar{n}_d,\ n,\ V_{\mathrm{mol}},\ Y,\ \rho,\ \Delta H_{\mathrm{at}},\ \Delta H_{\mathrm{fusion}}\}$
5 $\{c,\ \bar{n}_d,\ n,\ n_d,\ V_{\mathrm{mol}},\ Y,\ \rho,\ \Delta H_{\mathrm{at}}\}$
6 $\{c,\ n_d,\ V_{\mathrm{mol}},\ g_{\mathrm{P}},\ r_{\mathrm{cov}},\ Y,\ \mathrm{MN},\ \Delta H_{\mathrm{at}}\}$
7 $\{c,\ \bar{n},\ p_{\mathrm{P}},\ n,\ n_p,\ V_{\mathrm{mol}},\ Y,\ \Delta H_{\mathrm{at}}\}$
8 $\{\mathrm{IP}_2,\ c,\ n,\ n_p,\ V_{\mathrm{mol}},\ Y,\ \rho,\ \Delta H_{\mathrm{at}}\}$
9 $\{c,\ \bar{n},\ n,\ V_{\mathrm{mol}},\ H_{\mathrm{form}},\ r_p,\ \mathrm{MN},\ \Delta H_{\mathrm{fusion}}\}$
10 $\{c,\ \bar{n},\ n,\ n_p,\ V_{\mathrm{mol}},\ Y,\ \rho,\ \Delta H_{\mathrm{vapor}}\}$
11 $\{c,\ \bar{n},\ n,\ n_d,\ V_{\mathrm{mol}},\ Y,\ \rho,\ \Delta H_{\mathrm{vapor}}\}$
12 $\{c,\ \bar{n},\ n,\ V_{\mathrm{mol}},\ Y,\ m_{\mathrm{atom}},\ C,\ \Delta H_{\mathrm{at}}\}$
13 $\{c,\ n,\ n_d,\ V_{\mathrm{mol}},\ H_{\mathrm{form}},\ g_{\mathrm{P}},\ \Delta H_{\mathrm{fusion}},\ \lambda\}$
14 $\{c,\ \bar{n},\ n,\ n_s,\ V_{\mathrm{mol}},\ H_{\mathrm{form}},\ r_{\mathrm{cov}},\ \mathrm{MN}\}$
15 $\{c,\ \bar{n},\ \bar{n}_s,\ n,\ V_{\mathrm{mol}},\ H_{\mathrm{form}},\ r_{\mathrm{cov}},\ \mathrm{MN}\}$
16 $\{c,\ \bar{n}_d,\ n,\ V_{\mathrm{mol}},\ Y,\ \mathrm{MN},\ \rho,\ \Delta H_{\mathrm{vapor}}\}$
17 $\{c,\ \bar{n},\ n,\ n_p,\ V_{\mathrm{mol}},\ Y,\ m_{\mathrm{atom}},\ \Delta H_{\mathrm{at}}\}$
18 $\{\bar{n},\ \bar{n}_p,\ n_d,\ V_{\mathrm{mol}},\ \eta,\ N,\ \Delta H_{\mathrm{at}},\ \Delta H_{\mathrm{fusion}}\}$
19 $\{c,\ \bar{n},\ \bar{n}_p,\ V_{\mathrm{mol}},\ r_{\mathrm{vdw}},\ \mathrm{MN},\ \Delta H_{\mathrm{at}},\ \Delta H_{\mathrm{fusion}}\}$
20 $\{c,\ n_d,\ n_p,\ V_{\mathrm{mol}},\ r_{\mathrm{cov}},\ Y,\ \mathrm{MN},\ \Delta H_{\mathrm{at}}\}$
21 $\{c,\ r_d,\ n,\ n_d,\ V_{\mathrm{mol}},\ r_{\mathrm{vdw}},\ \Delta H_{\mathrm{at}},\ \Delta H_{\mathrm{fusion}}\}$
22 $\{\bar{n},\ n_d,\ V_{\mathrm{mol}},\ Y,\ N,\ m,\ \rho,\ \lambda\}$
23 $\{b,\ \bar{n},\ V_{\mathrm{mol}},\ g_{\mathrm{P}},\ Y,\ N,\ C,\ \rho\}$
24 $\{b,\ \bar{n},\ \bar{n}_s,\ p_{\mathrm{P}},\ n_d,\ V_{\mathrm{mol}},\ r_p,\ \mathrm{MN}\}$
25 $\{c,\ \bar{n},\ \bar{n}_p,\ V_{\mathrm{mol}},\ g_{\mathrm{P}},\ \Delta H_{\mathrm{at}},\ \chi,\ \Delta H_{\mathrm{fusion}}\}$
26 $\{\bar{n},\ n_d,\ e_{\mathrm{atom}},\ V_{\mathrm{mol}},\ Y,\ N,\ \rho,\ \lambda\}$
27 $\{c,\ \bar{n},\ \bar{n}_d,\ n_d,\ V_{\mathrm{mol}},\ a/b,\ \rho,\ \Delta H_{\mathrm{fusion}}\}$
28 $\{c,\ \bar{n},\ n_s,\ V_{\mathrm{mol}},\ H_{\mathrm{form}},\ r_s,\ \rho,\ \lambda\}$
⋮

n/a

487 $\{V_{\mathrm{atom}},\ V_{\mathrm{mol}},\ H_{\mathrm{form}},\ C\}$
488 $\{V_{\mathrm{atom}},\ V_{\mathrm{mol}},\ C,\ \Delta H_{\mathrm{at}}\}$
489 $\{r_d,\ V_{\mathrm{mol}},\ m_{\mathrm{atom}},\ \Delta H_{\mathrm{at}}\}$
490 $\{r_d,\ V_{\mathrm{mol}},\ H_{\mathrm{form}},\ m_{\mathrm{atom}}\}$
491 $\{\bar{n}_p,\ V_{\mathrm{mol}},\ m_{\mathrm{atom}},\ \Delta H_{\mathrm{at}}\}$
492 $\{\mathrm{IP}_1,\ V_{\mathrm{mol}},\ \rho,\ \Delta H_{\mathrm{fusion}}\}$
493 $\{V_{\mathrm{atom}},\ V_{\mathrm{mol}},\ C,\ \Delta H_{\mathrm{vapor}}\}$
494 $\{p_{\mathrm{P}},\ n_d,\ V_{\mathrm{mol}},\ \Delta H_{\mathrm{vapor}}\}$
495 $\{\bar{n}_d,\ V_{\mathrm{mol}},\ \Delta H_{\mathrm{at}},\ \Delta H_{\mathrm{fusion}}\}$
496 $\{\bar{n}_p,\ V_{\mathrm{atom}},\ V_{\mathrm{mol}},\ \Delta H_{\mathrm{at}}\}$
497 $\{p_{\mathrm{P}},\ n_d,\ V_{\mathrm{mol}},\ \mathrm{L}\}$
498 $\{\bar{n}_p,\ V_{\mathrm{atom}},\ V_{\mathrm{mol}},\ H_{\mathrm{form}}\}$
499 $\{\bar{n}_d,\ \bar{n}_p,\ V_{\mathrm{mol}},\ \Delta H_{\mathrm{at}}\}$
500 $\{p_{\mathrm{P}},\ n_d,\ V_{\mathrm{mol}},\ \Delta H_{\mathrm{fusion}}\}$
501 $\{\bar{n}_p,\ V_{\mathrm{atom}},\ V_{\mathrm{mol}},\ \Delta H_{\mathrm{vapor}}\}$
502 $\{\mathrm{IP}_3,\ V_{\mathrm{atom}},\ V_{\mathrm{mol}},\ \Delta H_{\mathrm{at}}\}$
503 $\{V_{\mathrm{mol}},\ C,\ m,\ \Delta H_{\mathrm{at}}\}$
504 $\{\bar{n}_p,\ p_{\mathrm{P}},\ n,\ V_{\mathrm{mol}}\}$
505 $\{V_{\mathrm{atom}},\ V_{\mathrm{mol}},\ B,\ \Delta H_{\mathrm{vapor}}\}$
506 $\{V_{\mathrm{mol}},\ H_{\mathrm{form}},\ C,\ m\}$
507 $\{\bar{n},\ V_{\mathrm{mol}},\ \Delta H_{\mathrm{at}},\ \Delta H_{\mathrm{vapor}}\}$
508 $\{\bar{n},\ V_{\mathrm{atom}},\ V_{\mathrm{mol}},\ H_{\mathrm{form}}\}$
509 $\{\mathrm{IP}_3,\ n_d,\ V_{\mathrm{mol}},\ H_{\mathrm{form}}\}$
510 $\{\bar{n}_d,\ p_{\mathrm{P}},\ n_d,\ V_{\mathrm{mol}}\}$
511 $\{\mathrm{IP}_3,\ n_d,\ V_{\mathrm{mol}},\ \Delta H_{\mathrm{at}}\}$
512 $\{\bar{n},\ V_{\mathrm{atom}},\ V_{\mathrm{mol}},\ \Delta H_{\mathrm{at}}\}$
513 $\{\bar{n},\ n_p,\ V_{\mathrm{mol}},\ \Delta H_{\mathrm{at}}\}$
514 $\{\bar{n},\ n_p,\ V_{\mathrm{mol}},\ H_{\mathrm{form}}\}$
515 $\{\bar{n}_d,\ n_d,\ V_{\mathrm{mol}},\ C\}$
516 $\{\mathrm{IP}_3,\ V_{\mathrm{atom}},\ V_{\mathrm{mol}},\ H_{\mathrm{form}}\}$
517 $\{n_p,\ V_{\mathrm{mol}},\ \mathrm{MN},\ \Delta H_{\mathrm{at}}\}$
518 $\{p_{\mathrm{P}},\ n_d,\ V_{\mathrm{mol}},\ g_{\mathrm{P}}\}$
519 $\{n_d,\ m_{\mathrm{atom}},\ \rho,\ \Delta H_{\mathrm{fusion}}\}$
520 $\{\bar{n}_p,\ V_{\mathrm{mol}},\ r_s,\ \Delta H_{\mathrm{at}}\}$
521 $\{n_p,\ V_{\mathrm{mol}},\ \Delta H_{\mathrm{at}},\ \mathrm{EA}\}$
522 $\{\bar{n},\ V_{\mathrm{mol}},\ H_{\mathrm{form}},\ \Delta H_{\mathrm{vapor}}\}$
523 $\{V_{\mathrm{atom}},\ V_{\mathrm{mol}},\ C,\ \Delta H_{\mathrm{fusion}}\}$
524 $\{\bar{n},\ p_{\mathrm{P}},\ n_d,\ V_{\mathrm{mol}}\}$
525 $\{n_d,\ V_{\mathrm{mol}},\ H_{\mathrm{form}},\ \Delta H_{\mathrm{fusion}}\}$
526 $\{\bar{n}_s,\ p_{\mathrm{P}},\ n_d,\ V_{\mathrm{mol}}\}$
527 $\{n_d,\ V_{\mathrm{mol}},\ g_{\mathrm{P}},\ \Delta H_{\mathrm{at}}\}$
528 $\{r_d,\ V_{\mathrm{mol}},\ r_p,\ \Delta H_{\mathrm{at}}\}$
529 $\{n_s,\ V_{\mathrm{mol}},\ \Delta H_{\mathrm{at}},\ \Delta H_{\mathrm{vapor}}\}$
530 $\{\bar{n}_p,\ p_{\mathrm{P}},\ n_d,\ V_{\mathrm{mol}}\}$
531 $\{n_d,\ V_{\mathrm{mol}},\ H_{\mathrm{form}},\ g_{\mathrm{P}}\}$
532 $\{V_{\mathrm{atom}},\ n_p,\ V_{\mathrm{mol}},\ H_{\mathrm{form}}\}$
533 $\{V_{\mathrm{mol}},\ \mathrm{MN},\ \Delta H_{\mathrm{at}},\ \Delta H_{\mathrm{vapor}}\}$
534 $\{n_d,\ V_{\mathrm{mol}},\ H_{\mathrm{form}},\ \mathrm{MN}\}$
535 $\{V_{\mathrm{atom}},\ V_{\mathrm{mol}},\ \rho,\ \Delta H_{\mathrm{vapor}}\}$
536 $\{p_{\mathrm{P}},\ n_d,\ V_{\mathrm{mol}},\ \chi\}$
537 $\{r_d,\ V_{\mathrm{mol}},\ m,\ \Delta H_{\mathrm{at}}\}$
⋮

Fig. 5.20. Identified minimally non-redundant (sub-optimal) feature subsets for the prediction of the shear modulus K_{VRH} in the Voigt-Reuss-Hill notation [485] using the tolerance-based branch-and-bound algorithm (TB3) and GBDT or SISSO as the feature-selection criterion (Section 4.1). Shown is the intersection and the union of the two feature-identification methods corresponding to a Jaccard similarity coefficient (Eq. 5.1) of zero. It should be noted that due to the different machine-learning algorithms, a direct comparison between TB3-GBDT and TB3-SISSO may not be appropriate. The numbering is used for referencing purposes.

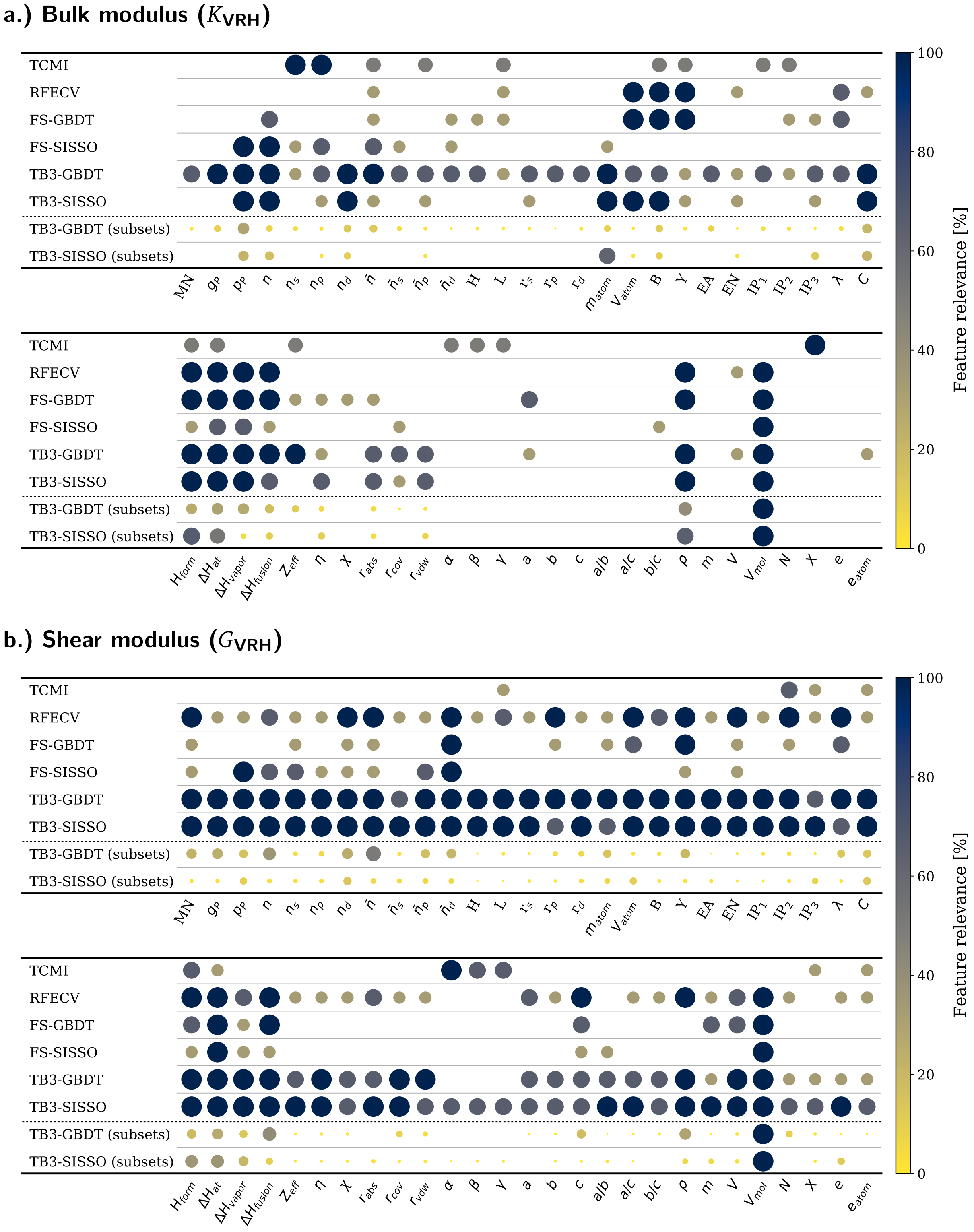


Fig. 5.21. Heat-map of most frequent identified (redundant) feature subsets of the elastic data set across all investigated feature-identification methods (cf., Section 5.1). The size and color reflect the frequency of the features (= relevance). A frequency analysis of identified minimal non-redundant feature subsets of the tolerance-based branch-and-bound algorithm (TB3) and GBDT is provided in the last line of each target property.

Feature identification further hints at the existence of a statistical relationship between the number of unfilled valence orbitals ($\bar{n}$) and the bulk and shear moduli. There has been much speculation about the relevance of unfilled valence orbitals [509, 511]. However, although $\bar{n}$ can be considered statistically as a rough characterization of the bulk and shear modulus, a physical relationship has yet to be provided.

In contrast to the octet-binary compound semiconductors and perovskites data sets (cf., Sections 5.2.1 and 5.2.2), there is no trend towards lower prediction errors of symbolic-regression models with increasing numbers of data samples and of identified relevant features. In particular, TB3-GBDT machine-learning models of the redundant feature subsets have lower prediction errors than models based on the optimal minimal non-redundant feature subsets. Although the effect is marginal, it is possible that identified feature-subsets of the TB3-GBDT algorithm either do not identify some of the relevant features for the estimation of K_{VRH} or G_{VRH} or the model construction requires a higher flexibility in the symbolic-regression models in the assessment of the final prediction performance (Section 5.2.2).

With respect to the estimation of the bulk and shear modulus, RFECV, FS-GBDT, and TB3 have similar prediction errors, whereas the prediction errors of TCMI are about 50% to twice as large for the bulk and shear modulus (Tabs. 5.15 and 5.16). In terms of feature identification, the non-redundant optimal feature subsets of TB3 result in the highest prediction performance with the least number of features, closely followed by RFECV and FS-GBDT. As such, the developed feature-identification framework (cf., Section 3.5) is efficient in identifying the relevant features of the elastic data set and in creating statistical models without significantly reducing the prediction performance of the generated machine-learning models.

Feature dependences

A feature-dependence map generated with the TB3-GBDT algorithm[26] reveals multiple statistical relationships between the features and the bulk or shear modulus (Fig. 5.22). In particular, the feature-dependence map exhibits a block-like structure of feature interactions, of which features of each type (fundamental, compositional, and structural properties) are more closely related to each other than to the features of the other types. This leads to a partitioning of the feature-dependence map into sub-blocks of feature interactions. Though, not all of the features of each sub-block are found to be dependent in the feature-dependence map due to the high redundancy of the features (cf., Section 2.4 and Chapter 3), features within these sub-blocks can be used interchangeably without decreasing the prediction performance of a machine-learning model. As such, machine-learning models are expected to require at least one feature from each sub-block to estimate the bulk or shear modulus of the inorganic crystalline compound data set with the same accuracy as compared to a machine-learning model constructed on the full set of features (cf., Tabs. 5.15 and 5.16).

There are a few features that are either independent or have very few feature interactions, such as the atomic weight (m_{atom}, m), molar volume (V_{mol}, V), density (ρ), number of electrons (e_{atom}, e), electron affinity (EA), chemical hardness (η), or the formation (H_{form}) and atomization enthalpy

[26] A feature-dependence map generated with the TB3-SISSO algorithm was attempted, but failed due to the higher computational requirements of the underlying SISSO algorithm.

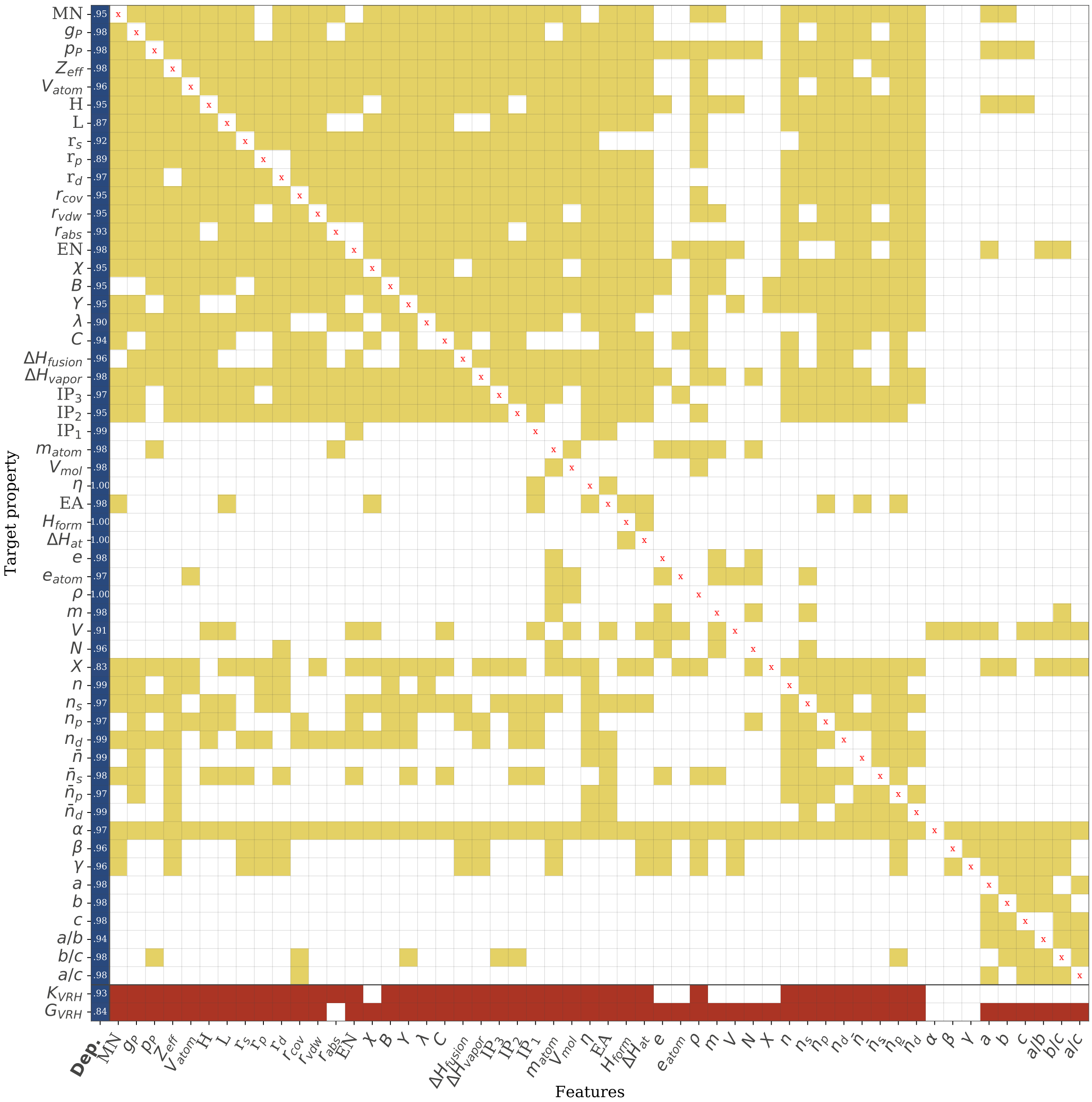


Fig. 5.22. Feature-dependence map of the crystalline inorganic compounds data set created with the tolerance-based branch-and-bound algorithm (TB3) and the gradient-boosting decision trees algorithm (GBDT) [249–252, 267, 268]. The feature-dependence map was created with a convergence threshold of $\epsilon = 0.01$ (cf., Section 4.2) at a confidence level of $\alpha = 0.95$. The score (=strength) of the dependence is shown in the first column (Dep.). The score was determined by the Pearson's R coefficient of determination [101] of a 10-fold cross-validated machine-learning model using the identified features of the TB3-GBDT algorithm as feature subsets (colored boxes). The dependences of the bulk (K_{VRH}) and shear modulus (G_{VRH}) in the Voigt-Reuss-Hill notation [485] are shown in the last two rows.

Features: V_{mol}, p_{P}, g_{P}, ΔH_{at}, H_{form}, $\Delta H_{\mathrm{fusion}}$, ρ, B, Y, $\bar{n}$, n

a.) Bulk modulus (K_{VRH})

Statistics: $R^2 = 0.94\pm0.01$ RMSE = (17.0 ± 1.7) GPa

Regression model:

$$K_{\mathrm{VRH}} = 55.9 \cdot \frac{p_{\mathrm{P}}\Delta H_{\mathrm{at}}}{V_{\mathrm{mol}}^2} \frac{\log(V_{\mathrm{mol}})}{\sqrt[3]{H_{\mathrm{form}}}} - 2.9 \cdot \left(\left|(Y/B) - |p_{\mathrm{P}} - n|\right| - \left((p_{\mathrm{P}} + g_{\mathrm{P}}) + |p_{\mathrm{P}} - \bar{n}|\right)\right)$$

$$- 8.62\mathrm{e}{+}03 \cdot \frac{|(n - p_{\mathrm{P}}) - (n/p_{\mathrm{P}})|}{V_{\mathrm{mol}}^4/\Delta H_{\mathrm{fusion}}} - 36.9$$

b.) Shear modulus (G_{VRH})

Statistics: $R^2 = 0.82\pm0.05$ RMSE = (18.5 ± 2.0) GPa

Regression model:

$$G_{\mathrm{VRH}} = 6.93 \cdot \frac{p_{\mathrm{P}}\Delta H_{\mathrm{at}}}{V_{\mathrm{mol}}^2} \sqrt{\frac{n}{g_{\mathrm{P}}}} + 0.00447 \cdot \frac{Y^3}{H_{\mathrm{form}} V_{\mathrm{mol}}} \frac{1}{(p_{\mathrm{P}}/g_{\mathrm{P}}) + (n - g_{\mathrm{P}})}$$

$$- 0.383 \cdot \frac{B\rho}{V_{\mathrm{mol}}^2} |\bar{n} + \bar{n} - g_{\mathrm{P}}| + 12.2$$

Fig. 5.23. Performance statistics and 10-fold cross-validated symbolic-regression models (SISSO [197]) of 1,181 inorganic crystalline compounds based on the eleven most frequently identified features from the non-redundant sub-optimal feature subsets of TB3-GBDT (cf., Tabs. 5.17 and 5.18). The prediction performance (cf., Section 5.1) is reported in terms of the mean $\bar{x}$ and standard deviation Δx as $\bar{x} \pm \Delta x$ of the Pearson's coefficient of determination $x = R^2$ [101] and the root-mean-squared error (x = RMSE). The RMSE is in the unit of gigapascals (GPa) and was approximated as described in Section 5.2.3.

(ΔH_{at}). Some of these features (namely the mass, molar volume, the formation and the atomization enthalpy) are strongly statistically related to the bulk and shear modulus and therefore may play an important role in the construction of a machine-learning model in the prediction of K_{VRH} and G_{VRH}.

Statistical models

The prediction performance of statistical equivalent (cf., Section 5.1) non-redundant sub-optimal feature subsets of TB3-GBDT are summarized in (Tabs. 5.17 and 5.18). In total, 75 out of 292 (K_{VRH}) and 201 out of 465 feature subsets (G_{VRH}) are found to be statistically equivalent at a confidence level of $\alpha = 0.95$. Thus, as in the octet-binary compound semiconductors and perovskites data set, there is an ensemble of competing machine-learning models of different feature subsets but statistically equivalent prediction performances.

A frequency analysis suggests that both the bulk and shear modulus can be described by the same set of features as given by the elastic tensor (Fig. 5.21). By far the most frequently identified features are the molar volume (V_{mol}), the mean group (g_{P}), the mean period (p_{P}), heat of formation (H_{form}), fusion ($\Delta H_{\mathrm{fusion}}$) and atomization enthalpy (ΔH_{at}), followed by the density (ρ), the total number of valence electrons (n), the total number of unfilled valence orbitals ($\bar{n}$), and the averaged atomic bulk (B) and Young's modulus (Y).

Logically, a symbolic-regression model based on the eleven features outperforms the investigated feature-identification methods in terms of feature-subset size and prediction performance (cf., Tabs. 5.15 and 5.16). That is, the tolerance-based branch-and-bound algorithm identifies the smallest possible materials representation as compared to TCMI, RFECV, and FS-GBDT: namely, 4–5 features for estimating the bulk modulus K_{VRH}, 6 features for estimating the shear modulus G_{VRH}, and a total of 11 features for estimating both properties.

The framework further relates the input features to the prediction performance of the generated machine-learning models, while providing a means to investigate the statistical relationships using deterministic, symbolic-regression models (Fig. 5.23). Although such models (and machine-learning models in general) only capture an overall statistical trend in the data and therefore cannot be regarded as a physical law[27] [55], several interesting relationships can be identified. For instance, bulk and shear moduli can be considered intrinsically related to the atomic bulk and Young's modulus. Second, compounds that require higher energies for breaking the inter-atomic bonds (ΔH_{at}) seem to be more resistant to external stresses and forces than compounds whose atoms are only weakly bonded. And third, compounds with a low molar volume (V_{mol}) of the constituent elements show to have a higher bulk and shear modulus in general.

Quantitatively, both symbolic-regression models (Fig. 5.24) correctly model the trends of the bulk and shear modulus as a function of the marginal contributions of the mean atomization enthalpy (ΔH_{at}) and the molar volume (V_{mol}), but fail to model the mean fusion enthalpy ($\Delta H_{\mathrm{fusion}}$), heat of formation ($H_{\mathrm{form}}$), bulk ($B$) and Young's ($Y$) modulus. The inaccurate modeling of the relevant features $\Delta H_{\mathrm{fusion}}$, H_{form}, B, and Y severely limits the applicability of the generated symbolic-regression models. However, both symbolic-regression models have been shown to be capable of estimating the bulk and shear modulus for a wide range of k-nary compounds of different chemical compositions and structures (Fig. 5.23), while expressing the statistical relationships in simple analytical terms. Therefore, even though the models cannot fully describe the material's behavior, they may still prove useful in specific materials-science applications or in the pre-screening over larger databases.

Uncertainty estimation

The inaccurate modeling of some of the statistical trends (Fig. 5.24) prevents a direct application of the generated symbolic-regression models. Without knowing the error made in the material-property estimations, the difference between the estimated and the actual value of the target properties can be quite large. Uncertainty estimates account for the model's prediction errors and quantify the probability that the actual target properties will be within the predicted range of the generated symbolic-regression models. In Section 4.3 two kinds of uncertainty methods were discussed: ensemble [52] and conformal prediction [60–64]. Both methods are known to be more robust [416] and reliable in terms of estimating the model's variance in the material-property predictions [360, 363, 378] and of providing a more comprehensive understanding of the machine-learning model (cf., [358]). More specifically, they take advantage of the fact that the prediction mean and uncertainty can be computed

[27] At the microscopic level, bulk and shear modulus are determined by the inter-atomic potential of crystalline materials.

#	**Features**	**GBDT**		**SISSO**	
	+ Model	R^2	RMSE [GPa]	R^2	RMSE [GPa]
1	$\{\boldsymbol{\bar{n}}, \boldsymbol{\bar{n}_s}, \boldsymbol{p_{\mathrm{P}}}, \boldsymbol{V_{\mathrm{mol}}}, \boldsymbol{\Delta H_{\mathrm{at}}}\}$	0.93±0.03	18.2±3.7	0.94±0.01	17.3±1.9
	$K_{\mathrm{VRH}} = 55.9 \cdot \frac{p_{\mathrm{P}} \Delta H_{\mathrm{at}} \log(V_{\mathrm{mol}})}{V_{\mathrm{mol}}^2 \sqrt[3]{\Delta H_{\mathrm{at}}}} - 0.00145 \cdot \frac{\Delta H_{\mathrm{at}}^2}{\exp(p_{\mathrm{P}})} \left\lvert (\bar{n} - p_{\mathrm{P}}) - \sqrt[3]{p_{\mathrm{P}}} \right\rvert - 24.1 \cdot \left\lvert \frac{\bar{n}}{\log V_{\mathrm{mol}}} - \left(\bar{n}_s + \frac{\bar{n}}{p_{\mathrm{P}}}\right) \right\rvert + 10.7$				
2	$\{\boldsymbol{p_{\mathrm{P}}}, \boldsymbol{V_{\mathrm{mol}}}, \boldsymbol{\Delta H_{\mathrm{at}}}, \boldsymbol{\Delta H_{\mathrm{fusion}}}\}$	0.93±0.03	19.2±3.5	0.94±0.01	18.0±1.5
	$K_{\mathrm{VRH}} = 27 \cdot \frac{p_{\mathrm{P}}}{V_{\mathrm{mol}}} \frac{\Delta H_{\mathrm{at}} + \Delta H_{\mathrm{fusion}}}{\log(V_{\mathrm{mol}}) \sqrt[3]{\Delta H_{\mathrm{at}}}} - 0.121 \cdot \left\lvert \frac{\Delta H_{\mathrm{at}}}{\log(p_{\mathrm{P}})} - \left((\Delta H_{\mathrm{at}} - \Delta H_{\mathrm{fusion}}) - p_{\mathrm{P}} \Delta H_{\mathrm{fusion}}\right) \right\rvert + 4.13\mathrm{e}{+}03 \cdot \frac{p_{\mathrm{P}} \Delta H_{\mathrm{fusion}} \exp(-\Delta H_{\mathrm{fusion}})}{(V_{\mathrm{mol}} / \Delta H_{\mathrm{fusion}})^3} - 8.43$				
3	$\{\boldsymbol{p_{\mathrm{P}}}, \boldsymbol{V_{\mathrm{mol}}}, \boldsymbol{H_{\mathrm{form}}}, \boldsymbol{\Delta H_{\mathrm{fusion}}}\}$	0.93±0.03	19.1±3.6	0.94±0.01	18.1±1.5
	$K_{\mathrm{VRH}} = 27 \cdot \frac{(p_{\mathrm{P}}/V_{\mathrm{mol}})(H_{\mathrm{form}} + \Delta H_{\mathrm{fusion}})}{\log(V_{\mathrm{mol}}) \sqrt[3]{H_{\mathrm{form}}}} - 0.121 \cdot \left\lvert \frac{H_{\mathrm{form}}}{\log(p_{\mathrm{P}})} - \left((H_{\mathrm{form}} - \Delta H_{\mathrm{fusion}}) - p_{\mathrm{P}} \Delta H_{\mathrm{fusion}}\right) \right\rvert + 4.16\mathrm{e}{+}03 \cdot \frac{p_{\mathrm{P}} \Delta H_{\mathrm{fusion}} \exp(-\Delta H_{\mathrm{fusion}})}{(V_{\mathrm{mol}} / \Delta H_{\mathrm{fusion}})^3} - 8.65$				
4	$\{\boldsymbol{p_{\mathrm{P}}}, \boldsymbol{V_{\mathrm{mol}}}, \boldsymbol{H_{\mathrm{form}}}, \boldsymbol{g_{\mathrm{P}}}, \boldsymbol{Y}\}$	0.94±0.03	17.9±3.9	0.94±0.01	18.0±2.1
	$K_{\mathrm{VRH}} = 40.3 \cdot \frac{p_{\mathrm{P}} H_{\mathrm{form}} V_{\mathrm{mol}}^2}{\log(H_{\mathrm{form}}) \log(V_{\mathrm{mol}})} - 0.0586 \cdot \left(\frac{Y \log(H_{\mathrm{form}})}{p_{\mathrm{P}}} - \frac{Y g_{\mathrm{P}}}{\log(H_{\mathrm{form}})}\right) + 554 \cdot \exp\left(-\frac{Y}{g_{\mathrm{P}}}\right) \frac{Y^2}{V_{\mathrm{mol}}^3} + 7.67$				
5	$\{\boldsymbol{\bar{n}}, \boldsymbol{p_{\mathrm{P}}}, \boldsymbol{V_{\mathrm{mol}}}, \mathbf{MN}, \boldsymbol{\Delta H_{\mathrm{at}}}\}$	0.94±0.03	17.8±3.9	0.94±0.01	18.2±1.9
	$K_{\mathrm{VRH}} = 56 \cdot \frac{p_{\mathrm{P}} \Delta H_{\mathrm{at}}}{V_{\mathrm{mol}}^2} \frac{\log(V_{\mathrm{mol}})}{\sqrt[3]{\Delta H_{\mathrm{at}}}} - 0.00135 \cdot \frac{\Delta H_{\mathrm{at}}^2}{\exp(p_{\mathrm{P}})} \cdot \left\lvert (\bar{n} - p_{\mathrm{P}}) - \sqrt[3]{p_{\mathrm{P}}} \right\rvert - 4.02\mathrm{e}{+}03 \cdot \left\lvert \frac{\log(\mathrm{MN})}{\mathrm{MN} \cdot V_{\mathrm{mol}}} - \frac{\exp(-p_{\mathrm{P}})}{p_{\mathrm{P}} V_{\mathrm{mol}}} \right\rvert + 16.1$				
6	$\{\boldsymbol{\bar{n}}, \boldsymbol{p_{\mathrm{P}}}, \boldsymbol{V_{\mathrm{mol}}}, \boldsymbol{H_{\mathrm{form}}}, \mathbf{MN}\}$	0.94±0.03	17.9±3.9	0.94±0.01	18.0±1.8
	$K_{\mathrm{VRH}} = 9.24 \cdot \frac{p_{\mathrm{P}} \sqrt{H_{\mathrm{form}}}}{V_{\mathrm{mol}} \sqrt[3]{V_{\mathrm{mol}}/\mathrm{MN}}} + 1.79 \cdot \left(p_{\mathrm{P}} H_{\mathrm{form}} - \frac{H_{\mathrm{form}}}{\mathrm{MN} \exp\left(\sqrt{V_{\mathrm{mol}}}\right)}\right) - 0.00112 \cdot \frac{H_{\mathrm{form}}^2}{\exp(p_{\mathrm{P}})} \left((\bar{n} - p_{\mathrm{P}}) - \sqrt[3]{p_{\mathrm{P}}}\right) + 6.13$				

(continues on next page)

(continued from previous page)

7	$\{\boldsymbol{p_\mathrm{P}}, \boldsymbol{Z_\mathrm{eff}}, \boldsymbol{V_\mathrm{mol}}, \boldsymbol{\Delta H_\mathrm{at}}\}$	0.93±0.03	18.5±3.6	0.94±0.01	18.3±1.7
	$K_\mathrm{VRH} = 38.6 \cdot \frac{p_\mathrm{P} \Delta H_\mathrm{at} \log(V_\mathrm{mol})}{\log(\Delta H_\mathrm{at}) V_\mathrm{mol}^2} - 0.00368 \cdot \frac{\Delta H_\mathrm{at}^2 \sqrt[3]{\Delta H_\mathrm{at}}}{p_\mathrm{P}^3 \exp(Z_\mathrm{eff})} + 80.8 \cdot \frac{\lvert \exp(p_\mathrm{P}) - \exp(Z_\mathrm{eff}) \rvert}{\sqrt{\exp(V_\mathrm{mol})}} + 13.8$				
8	$\{\boldsymbol{p_\mathrm{P}}, \boldsymbol{Z_\mathrm{eff}}, \boldsymbol{V_\mathrm{mol}}, \boldsymbol{H_\mathrm{form}}\}$	0.93±0.03	18.3±3.6	0.94±0.01	18.2±2.1
	$K_\mathrm{VRH} = 38.6 \cdot \frac{p_\mathrm{P} H_\mathrm{form} \log(V_\mathrm{mol})}{\log(H_\mathrm{form}) V_\mathrm{mol}^2} - 0.00365 \cdot \frac{H_\mathrm{form}^2 \sqrt[3]{H_\mathrm{form}}}{p_\mathrm{P}^3 \exp(Z_\mathrm{eff})} + 81.4 \cdot \frac{\lvert \exp(p_\mathrm{P}) - \exp(Z_\mathrm{eff}) \rvert}{\sqrt{\exp(V_\mathrm{mol})}} + 13.5$				
9	$\{\boldsymbol{p_\mathrm{P}}, \boldsymbol{V_\mathrm{mol}}, \boldsymbol{H_\mathrm{form}}, \mathbf{H}\}$	0.93±0.03	19.4±3.8	0.93±0.03	19.6±3.8
	$K_\mathrm{VRH} = 1.92 \cdot \frac{p_\mathrm{P} H_\mathrm{form} - H_\mathrm{form}/p_\mathrm{P}}{\exp\left(\sqrt{V_\mathrm{mol}}\right)} - 29.7 \cdot \sqrt[3]{H} \sqrt{H_\mathrm{form} \frac{\log(p_\mathrm{P})}{V_\mathrm{mol}}}$ $+ 5.53 \cdot \frac{p_\mathrm{P}^3}{\exp(V_\mathrm{mol}) \cdot \left(\exp(H) - \exp(-V_\mathrm{mol})\right)} - 7.59$				
10	$\{\boldsymbol{p_\mathrm{P}}, \boldsymbol{V_\mathrm{mol}}, \boldsymbol{H_\mathrm{form}}, \boldsymbol{Y}\}$	0.93±0.03	19.5±3.5	0.94±0.01	18.0±2.1
	$K_\mathrm{VRH} = -182 \cdot \frac{H_\mathrm{form} - p_\mathrm{P} H_\mathrm{form}}{V_\mathrm{mol}^2 \sqrt[3]{H_\mathrm{form}}} + 918 \cdot \exp\left(-\frac{Y}{V_\mathrm{mol}}\right) \frac{Y^3}{\exp(V_\mathrm{mol})} - 2.15\mathrm{e}{+}04 \cdot \frac{(p_\mathrm{P}^3)^3}{\exp(V_\mathrm{mol}) p_\mathrm{P} H_\mathrm{form}}$ $+ 10.6$				

Tab. 5.17. The first 10 best-performing symbolic-regression models for the prediction of the bulk modulus K_VRH in the Voigt-Reuss-Hill notation [485]. Shown is the prediction performance of the statistical equivalent sub-optimal non-redundant feature subsets of the tolerance-based branch-and-bound algorithm (TB3) using the gradient-boosting decision trees (GBDT) algorithm [249–252, 267, 268] as the feature-selection criterion (Sections 3.1.2 and 4.1). Statistical equivalent feature subsets were sorted in increasing order of the SISSO [197] prediction errors (RMSE - ΔRMSE). Prediction errors of SISSO were approximated as described in Section 5.2.3. For comparison, 10-fold cross-validated prediction performances of SISSO and of GBDT (cf., Section 5.1) are reported in terms of the mean $\bar{x}$ and standard deviation Δx as $\bar{x} \pm \Delta x$ of the Pearson's coefficient of determination $x = R^2$ [101] and the root-mean-squared error (x = RMSE). The RMSE is in the unit of gigapascals (GPa).

#	**Features**	**GBDT**		**SISSO**	
	+ Model	R^2	RMSE [GPa]	R^2	RMSE [GPa]
1	$\{\boldsymbol{p_\mathrm{P}}, \boldsymbol{n}, \boldsymbol{V_\mathrm{mol}}, \boldsymbol{g_\mathrm{P}}, \boldsymbol{Y}, \boldsymbol{\Delta H_\mathrm{at}}\}$	0.82±0.08	18.7±6.4	0.82±0.04	18.6±1.9
	$G_\mathrm{VRH} = 2.72 \cdot \frac{(g_\mathrm{P}/n + \sqrt[3]{p_\mathrm{P}})(p_\mathrm{P} \Delta H_\mathrm{at})}{V_\mathrm{mol}^2} + 0.00201 \cdot \frac{Y^3}{n \Delta H_\mathrm{at}} \frac{1}{(n - g_\mathrm{P}) + (p_\mathrm{P}/g_\mathrm{P})}$ $- 189 \cdot \frac{\lvert n p_\mathrm{P} - (g_\mathrm{P} + p_\mathrm{P}) \rvert}{n^2} \frac{p_\mathrm{P}}{V_\mathrm{mol}} + 20.5$				

(continues on next page)

(continued from previous page)

2	$\{\boldsymbol{c}, \boldsymbol{n_d}, \boldsymbol{n_p}, \boldsymbol{V_{\mathrm{mol}}}, \boldsymbol{r_{\mathrm{cov}}}, \boldsymbol{Y}, \mathbf{MN}, \boldsymbol{\Delta H_{\mathrm{at}}}\}$	0.83±0.09	18.1±7.1	0.82±0.05	18.4±2.2
	$G_{\mathrm{VRH}} = -\,0.0398 \cdot \frac{(n_d - \mathrm{MN}) r_{\mathrm{cov}} \cdot \Delta H_{\mathrm{at}} \sqrt[3]{\Delta H_{\mathrm{at}}}}{V_{\mathrm{mol}}^2} - 0.00236 \cdot \frac{\exp\left(\sqrt[3]{Y}\right)}{n_p r_{\mathrm{cov}} - c/\mathrm{MN}} + 665 \cdot \frac{n_d \sqrt[3]{n_p} r_{\mathrm{cov}}^2}{V_{\mathrm{mol}} \exp(n_d)} + 15.5$				
3	$\{\boldsymbol{\bar{n}}, \boldsymbol{n_p}, \boldsymbol{V_{\mathrm{mol}}}, \boldsymbol{H_{\mathrm{form}}}, \boldsymbol{Y}, \boldsymbol{m_{\mathrm{atom}}}, \boldsymbol{\lambda}\}$	0.82±0.09	18.5±6.8	0.82±0.04	19.1±2.7
	$G_{\mathrm{VRH}} = 7.79 \cdot \frac{H_{\mathrm{form}} \log(m_{\mathrm{atom}})}{V_{\mathrm{mol}} (\log(V_{\mathrm{mol}}))^3} + 0.0892 \cdot \left(\frac{n_p}{V_{\mathrm{mol}}}\right)^2 \lambda \bar{n}^4 + 0.0057 \cdot \frac{Y^3 \exp(-n_p \cdot m_{\mathrm{atom}})}{H_{\mathrm{form}} V_{\mathrm{mol}}} + 12.8$				
4	$\{\boldsymbol{n}, \boldsymbol{V_{\mathrm{mol}}}, \boldsymbol{g_{\mathrm{P}}}, \boldsymbol{Y}, \boldsymbol{\rho}, \boldsymbol{\Delta H_{\mathrm{at}}}\}$	0.82±0.07	18.6±5.7	0.81±0.04	19.0±1.7
	$G_{\mathrm{VRH}} = 28.7 \cdot \frac{\Delta H_{\mathrm{at}} \sqrt[3]{\rho}}{V_{\mathrm{mol}}^2 \sqrt{n}} + 0.00431 \cdot \frac{\exp\left(\sqrt[3]{Y}\right)}{n^{-1} + (n - g_{\mathrm{P}})} + 43.8 \cdot \left(\frac{n - g_{\mathrm{P}}}{\exp(\rho)} + \frac{\sqrt{g_{\mathrm{P}}}}{\sqrt[3]{n}}\right) - 47.5$				
5	$\{\boldsymbol{\bar{n}}, \boldsymbol{n}, \boldsymbol{n_p}, \boldsymbol{V_{\mathrm{mol}}}, \boldsymbol{Y}, \boldsymbol{m_{\mathrm{atom}}}, \boldsymbol{\Delta H_{\mathrm{at}}}\}$	0.82±0.09	18.6±6.7	0.80±0.04	19.6±2.1
	$G_{\mathrm{VRH}} = 5.77 \cdot \frac{\Delta H_{\mathrm{at}} + (\Delta H_{\mathrm{at}}/n) \log(m_{\mathrm{atom}})}{V_{\mathrm{mol}}^2} + 918 \cdot \log\left(\frac{m_{\mathrm{atom}}}{V_{\mathrm{mol}}}\right) * \frac{\bar{n} \cdot n_p}{n \cdot m_{\mathrm{atom}}} + 0.032 \cdot \frac{Y^3}{\Delta H_{\mathrm{at}}^2} \frac{1}{n^{-1} - \sqrt[3]{n_p}} + 6.87$				
6	$\{\boldsymbol{\bar{n}}, \boldsymbol{\bar{n}_p}, \boldsymbol{\bar{n}_s}, \boldsymbol{p_{\mathrm{P}}}, \boldsymbol{n_d}, \boldsymbol{V_{\mathrm{mol}}}, \boldsymbol{H_{\mathrm{form}}}\}$	0.82±0.08	18.6±6.4	0.82±0.05	18.7±1.8
	$G_{\mathrm{VRH}} = 978 \cdot \left(\frac{p_{\mathrm{P}}}{V_{\mathrm{mol}}}\right)^3 \frac{\sqrt{H_{\mathrm{form}}}}{n_d + p_{\mathrm{P}}} - 6.18 \cdot \left\lvert \left(\bar{n}_s + (\bar{n}_p + \bar{n}_s)\right) - \left((\bar{n} - \bar{n}_s) - \lvert p_{\mathrm{P}} - \bar{n} \rvert\right) \right\rvert - 3.6\mathrm{e}{+}04 \cdot \frac{\bar{n} \cdot V_{\mathrm{mol}}}{\exp(\bar{n}_p)} \frac{\sqrt[3]{\bar{n}_p}}{\exp(V_{\mathrm{mol}})} + 31.8$				
7	$\{\boldsymbol{\bar{n}}, \boldsymbol{p_{\mathrm{P}}}, \boldsymbol{n}, \boldsymbol{n_p}, \boldsymbol{V_{\mathrm{mol}}}, \boldsymbol{Y}, \boldsymbol{\Delta H_{\mathrm{at}}}\}$	0.82±0.09	18.5±6.6	0.81±0.05	19.2±1.4
	$G_{\mathrm{VRH}} = -\,18.3 \cdot \frac{p_{\mathrm{P}} \Delta H_{\mathrm{at}}}{V_{\mathrm{mol}}^2} \left(\frac{n_p}{n} - \sqrt{n}\right) - 0.013 \cdot \exp\left(\sqrt[3]{Y}\right) \frac{\bar{n} - p_{\mathrm{P}}}{\exp(n_p)} + 1.45 \cdot \frac{(n + \bar{n}) + (\bar{n} \cdot n_p)}{\exp(p_{\mathrm{P}})/p_{\mathrm{P}}^3} - 9.51$				
8	$\{\boldsymbol{\bar{n}_p}, \boldsymbol{n}, \boldsymbol{V_{\mathrm{mol}}}, \boldsymbol{H_{\mathrm{form}}}, \boldsymbol{Y}, \boldsymbol{\rho}\}$	0.81±0.07	18.9±5.8	0.81±0.05	19.3±2.3
	$G_{\mathrm{VRH}} = 10.5 \cdot \frac{(n + \bar{n}_p) H_{\mathrm{form}} \log(\rho)}{n V_{\mathrm{mol}}^2} - 0.0139 \cdot \frac{Y^3}{V_{\mathrm{mol}}^2} \left(\bar{n}_p H_{\mathrm{form}} - \frac{H_{\mathrm{form}}}{n}\right) + 32.5 \cdot \frac{\sqrt[3]{H_{\mathrm{form}}}}{\sqrt{V_{\mathrm{mol}}}} \exp\left(\frac{V_{\mathrm{mol}}}{\rho}\right) - 16.4$				

(continues on next page)

(continued from previous page)

9	$\{\bar{\boldsymbol{n}}, \boldsymbol{n_d}, \boldsymbol{n_p}, \boldsymbol{V}_{\mathbf{mol}}, \boldsymbol{H}_{\mathbf{form}}, \boldsymbol{\rho}, \boldsymbol{\Delta H}_{\mathbf{fusion}}\}$	0.82±0.08	18.4±6.1	0.80±0.04	19.5±1.9
	$G_{\mathrm{VRH}} = 6.67 \cdot \frac{H_{\mathrm{form}}}{V_{\mathrm{mol}}^2} \log(\rho V_{\mathrm{mol}}) + 0.363 \cdot \left\lvert \frac{n_d \cdot \Delta H_{\mathrm{fusion}}}{\exp(n_p)} - \frac{H_{\mathrm{form}} + \Delta H_{\mathrm{fusion}}}{\log(H_{\mathrm{form}})} \right\rvert - 0.14 \cdot \left\lvert \bar{n}^2 - \frac{H_{\mathrm{form}}}{\Delta H_{\mathrm{fusion}}} \right\rvert \frac{\Delta H_{\mathrm{fusion}}}{V_{\mathrm{mol}}} \lvert n_d - n_p \rvert + 5.48$				
10	$\{\boldsymbol{p}_{\mathbf{P}}, \boldsymbol{V}_{\mathbf{atom}}, \boldsymbol{V}_{\mathbf{mol}}, \boldsymbol{g}_{\mathbf{P}}, \boldsymbol{Y}, \boldsymbol{\Delta H}_{\mathbf{vapor}}\}$	0.81±0.09	19.2±6.8	0.80±0.04	19.6±1.8
	$G_{\mathrm{VRH}} = 6.9 \cdot \sqrt{p_{\mathrm{P}} V_{\mathrm{atom}}} \frac{\Delta H_{\mathrm{vapor}} V_{\mathrm{atom}}}{V_{\mathrm{mol}}^3} + 5.42 \cdot \frac{Y \exp(p_{\mathrm{P}})}{\exp(g_{\mathrm{P}}) + \exp(V_{\mathrm{atom}})} + 3.02 \cdot \frac{g_{\mathrm{P}} \sqrt[3]{Y}}{p_{\mathrm{P}} \exp(Y / \Delta H_{\mathrm{vapor}})} - 5.77$				

Tab. 5.18. The first 10 best-performing symbolic-regression models for the prediction of the shear modulus G_{VRH} in the Voigt-Reuss-Hill notation [485]. Shown is the prediction performance of the statistical equivalent sub-optimal non-redundant feature subsets of the tolerance-based branch-and-bound algorithm (TB3) using the gradient-boosting decision trees (GBDT) algorithm [249–252, 267, 268] as the feature-selection criterion (Sections 3.1.2 and 4.1). Statistical equivalent feature subsets were sorted in increasing order of the SISSO [197] prediction errors (RMSE - ΔRMSE). Prediction errors of SISSO were approximated as described in Section 5.2.3. For comparison, 10-fold cross-validated prediction performances of SISSO and of GBDT (cf., Section 5.1) are reported in terms of the mean $\bar{x}$ and standard deviation Δx as $\bar{x} \pm \Delta x$ of the Pearson's coefficient of determination $x = R^2$ [101] and the root-mean-squared error (x = RMSE). The RMSE is in the unit of gigapascals (GPa).

as the average and the αth-percentile of the different model's predictions on the same set of data (Section 4.3).

Ensemble-based prediction intervals from resampling methods are smaller than conformal-based prediction intervals at the same confidence level α. However, they are as large as conformal-based prediction intervals at the same validity[28] $\hat{\alpha}$ (Tab. 5.19). In particular, the validity of ensemble-based methods is always less than that of the specified confidence level. A smaller validity than a specified confidence level results in uncertainty estimates being largely determined by the choice of training data (Section 4.3). Unlike ensemble-based methods, the validity of conformal prediction is expected to be at least as large as the confidence level on average (Section 4.3.1). Though, the validity of conformal prediction can be lower than the confidence level (e.g, at a confidence level of about $\alpha = 50\%$), conformal prediction is expected to produce more robust estimates than ensemble-based resampling methods, especially for the screening and application of the symbolic-regression models on new data.

The choice of conformal prediction comes at a cost (cf., [386]): for instance, at a confidence level of $\alpha = 0.95$, the prediction intervals are on average about three to four times larger than the root-mean-squared error of the generated symbolic-regression models. However, the prediction performances (Fig. 5.25) show that bulk and shear moduli are typically within the 50% confidence prediction bands and are about the same size as the root-mean-squared error of the generated symbolic-regression models.

[28] The probability that in about $\hat{\alpha}$% of the cases the actual values of the target property are within the prediction interval at a confidence level of α%.

↲ **Algorithm (TB3-...)**		**Ensemble prediction**			**Conformal prediction**		
	Confidence level (α)	Validity ($\hat{\alpha}$)	Mean (Δ)	Max (Δ)	Validity ($\hat{\alpha}$)	Mean (Δ)	Max (Δ)
a.) Bulk modulus (K_{VRH})							
GBDT	0.50	*0.27*	4.9	34.6	*0.50*	10.9	73.7
	0.80	*0.55*	11.1	83.3	*0.80*	23.3	152.4
	0.95	*0.76*	17.9	157.0	*0.95*	56.6	371.1
b.) Shear modulus (G_{VRH})							
GBDT	0.50	*0.24*	5.3	65.0	*0.49*	11.6	189.9
	0.80	*0.47*	10.2	143.6	*0.80*	27.7	449.0
	0.95	*0.67*	16.0	247.5	*0.96*	67.5	1111.7

Tab. 5.19. Validity of ensemble and conformal prediction of TB3-GBDT at three different confidence levels $\alpha = [50\%, 80\%, 95\%]$. The validity ($\hat{\alpha}$) specifies the probability ($\hat{\alpha}\% = 100\hat{\alpha}$) that the target property is within the prediction interval $x \in [\bar{x} - \bar{\Delta}, \bar{x} + \bar{\Delta}]$ of the statistical model (Eq. 4.22), where $\bar{x}$ and $\pm\bar{\Delta}$ ($\bar{\Delta} = \text{Mean}(\Delta) \leq \text{Max}(\Delta)$) denote the (point) estimation and variance of the ensemble or conformal prediction. The validity of the ensemble prediction and prediction intervals were computed as the αth-percentile of the ensemble machine-learning predictions. The validity and prediction intervals of conformal prediction were computed using the setup as described in Sections 4.3.2 and 5.1. Performance statistics of the bulk (K_{VRH}, a) and the shear modulus (G_{VRH}, b) were approximated with the approach as described in Section 5.2.3 using the sure-independence screening and sparsifying operator (SISSO) [197]. Units are in gigapascal (GPa).

Anomalous materials

One of the key advantages of conformal prediction is the identification of materials that are underrepresented or cannot be adequately estimated from the full set of data (so-called anomalous materials, Section 4.4). Although none of the inorganic crystalline compounds can be classified as anomalous (Fig. 5.25), there are some materials with very large prediction intervals. These include unary and binary oxide and nitride compounds such as Iridium (Ir), aluminum oxide (Al_2O_3), and platinum nitride (PtN_2), contributing most to the high uncertainty in the bulk and shear moduli in the range of 100 GPa–300 GPa of the predictions (depicted as squares in Fig. 5.25, not marked separately).

The largest prediction intervals occur for carbon[29] (C) and related compounds such as tungsten carbide (WC) all with bulk or shear moduli above 300 GPa. All of these compounds have a low molar volume and a high enthalpy of atomization and as such exhibit high bulk and shear moduli in line with the statistical trends as shown in the partial-dependence plots (Fig. 5.24) of the generated symbolic-regression models (Fig. 5.23). In total, there are less than 25 compounds with bulk and shear moduli above 300 GPa (Fig. 5.25). Because the mean predictions of most of these compounds are close to the actual values, but the prediction intervals are large, conformal prediction indicates that compounds with bulk and shear moduli above 300 GPa are underrepresented in the data set. As a result, the generated symbolic regression models should generally not be used to for the screening or estimation of ultrahigh bulk and shear modulus of inorganic crystalline compounds.

[29] Carbon in the diamond structure has the highest bulk and shear modulus in the data set ($K_{\text{VRH}} = 436$ GPa and $G_{\text{VRH}} = 523$ GPa).

↲ **Target property**	**Performance statistics**	
Model applied to Materials Project data	R^2	RMSE [GPa]
a.) Bulk modulus (K_{VRH})		
Elastic data (TB3-GBDT)	0.81 ± 0.09	31.6 ± 9.6
Elastic data (TB3-GBDT → SISSO)	0.79 ± 0.01	34.7 ± 0.6
Materials Project (TB3-GBDT)	0.84 ± 0.00	29.1 ± 0.3
b.) Shear modulus (G_{VRH})		
Elastic data (TB3-GBDT)	0.18 ± 0.03	73.1 ± 8.0
Elastic data (TB3-GBDT → SISSO)	0.17 ± 0.00	73.7 ± 0.4
Materials Project (TB3-GBDT)	0.23 ± 0.04	71.1 ± 2.1

Tab. 5.20. Comparison of machine-learning models applied to the Materials Project data set. The machine-learning models were constructed based on the 1,181 inorganic crystalline compounds [35, 412] data set using TB3-GBDT algorithm (cf., Tabs. 5.17 and 5.18), based on the same data set but using symbolic regression (SISSO [197]) with the most frequent identified feature subsets of TB3-GBDT (Fig. 5.23), and on the Materials Project data set [13, 15, 35] using TB3-GBDT. Prediction errors are in units of gigapascal (GPa).

Machine-learning predictions

The central objective of using machine learning in materials science is to build statistical models from easily accessible features and to apply these models to estimate otherwise difficult to measure or to calculate properties (cf., [55]). The quality and usefulness therefore crucially depends on how well these models perform on new data. In the following, the generalizability of the generated machine-learning models (Tabs. 5.15 and 5.16 and Fig. 5.23) is investigated on a much larger set of inorganic crystalline compounds than were available for statistical modeling, thus simulating a potential screening application in materials science.

A total of 13,172 inorganic crystalline compounds were downloaded from the Materials Project [13, 15, 35] with reported space group, volume, formula, and elastic properties. Of these, 2,566 compounds had to be removed because they were either already included in the curated data set (cf., Section 5.2.3) or had missing tabulated elemental properties for one of the compounds. The remaining 10,606 inorganic crystalline compounds were then used to test the generalizability of the generated machine-learning models from the elastic data set by comparing their prediction performance to machine-learning models constructed directly on the Materials Project data. Results are reported in Table 5.20.

Overall, it is observed that the machine-learning models applied to or constructed from the Materials Project data have larger prediction errors and higher variances (Tab. 5.20) as compared to the elastic data set (Tabs. 5.15 and 5.16 and Fig. 5.23) and that the prediction performance for the bulk modulus is still good given that the models were applied on a much larger and more diverse data set. However, the prediction performances of the elastic and Materials Project data are hardly comparable. First, because the prediction performances of the machine-learning models were estimated on materials other than those included in the elastic data set. And second, because the Materials Project

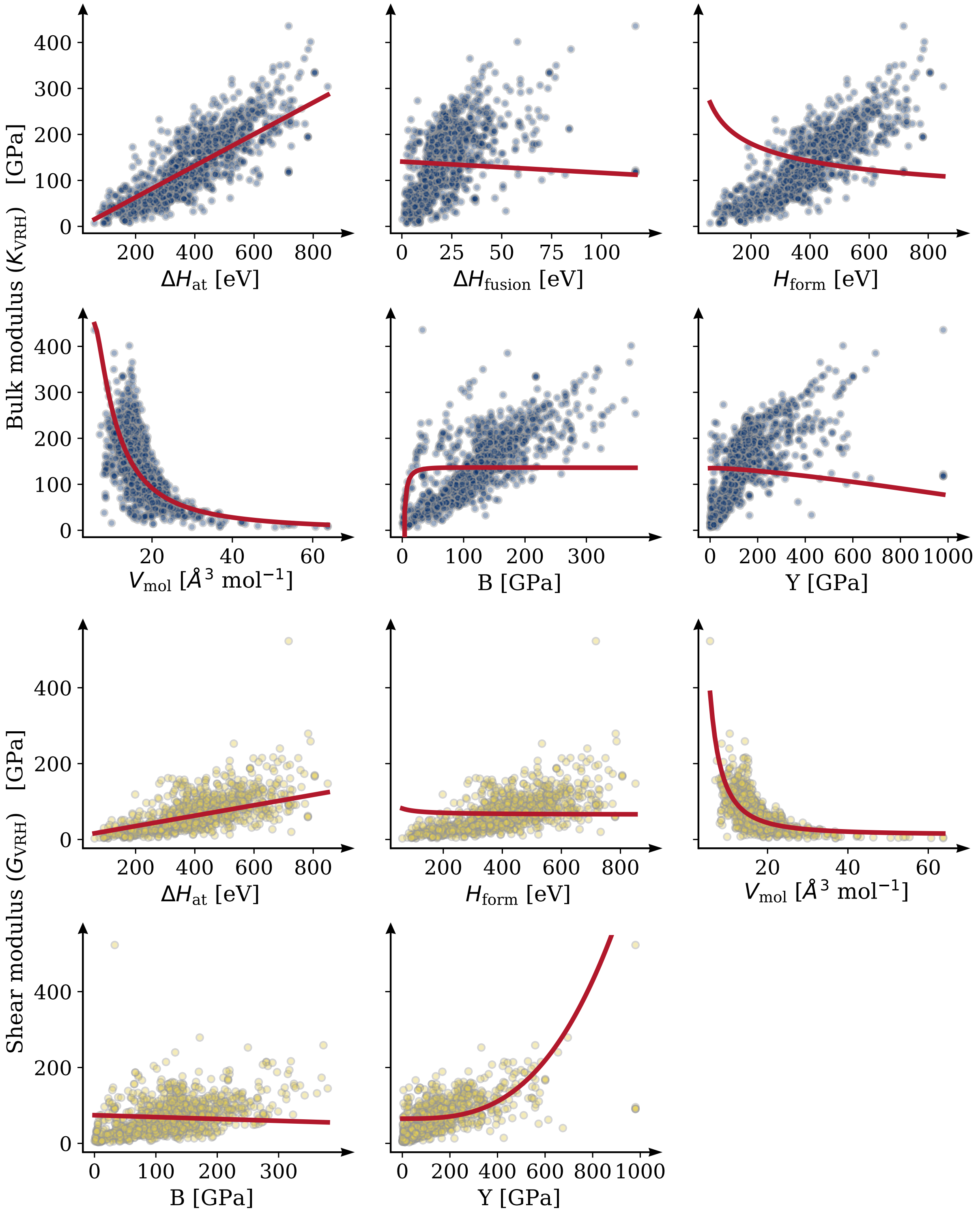


Fig. 5.24. Partial-dependence plots of the generated symbolic-regression models (Fig. 5.23) for the bulk (K_{VRH}) and the shear (G_{VRH}) modulus as a function of the atomization (ΔH_{at}), fusion enthalpy (ΔH_{fusion}), formation enthalpy (H_{form}), the molar volume (V_{mol}), and the atomic bulk (B) and Young's modulus (Y). Shown are the distribution of actual values of the inorganic crystalline compounds (points) and the statistical trend (line) obtained by marginalizing each feature (ΔH_{at}, ΔH_{fusion}, . . .) over the values of all other features of the data set [249].

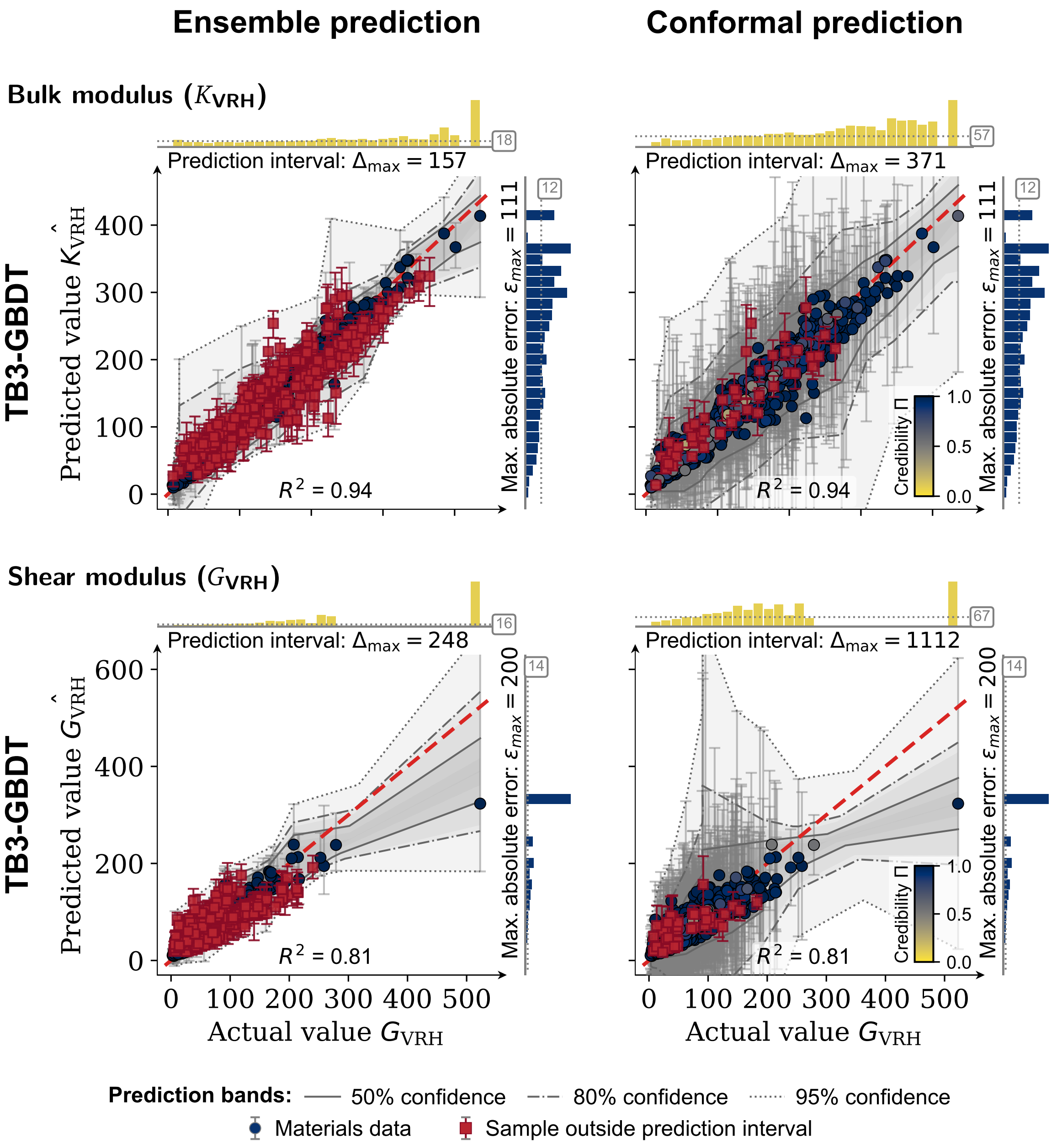


Fig. 5.25. Ensemble prediction performance of machine-learning models from TB3-GBDT using the sure-independence screening and sparsifying operator (SISSO) [197] to estimate the bulk (K_{VRH}) and shear (G_{VRH}) modulus in the Voigt-Reuss-Hill notation [485] from the set of statistically equivalent feature subsets. The prediction errors of SISSO were approximated using the setup as described in Section 5.2.3. Shown are the prediction bands (50th, 80th, 95th-percentiles) of the model's predictions, the credibility Π (Eq. 4.29) at a confidence level of $\alpha = 0.95$, the distribution of the size of prediction intervals (diagram above the parity plot, Δ), and the errors of the machine-learning model (diagram right of the parity plot, $\varepsilon = |a_0 - \hat{a}_0|$). The numbers in the boxes display the mean values ($\bar{\Delta}$, $\bar{\varepsilon}$), while the maximum errors are given in the texts below or to the left of the diagrams. Units are in gigapascal (GPa). Machine-learning estimates outside the prediction intervals are depicted as squares.

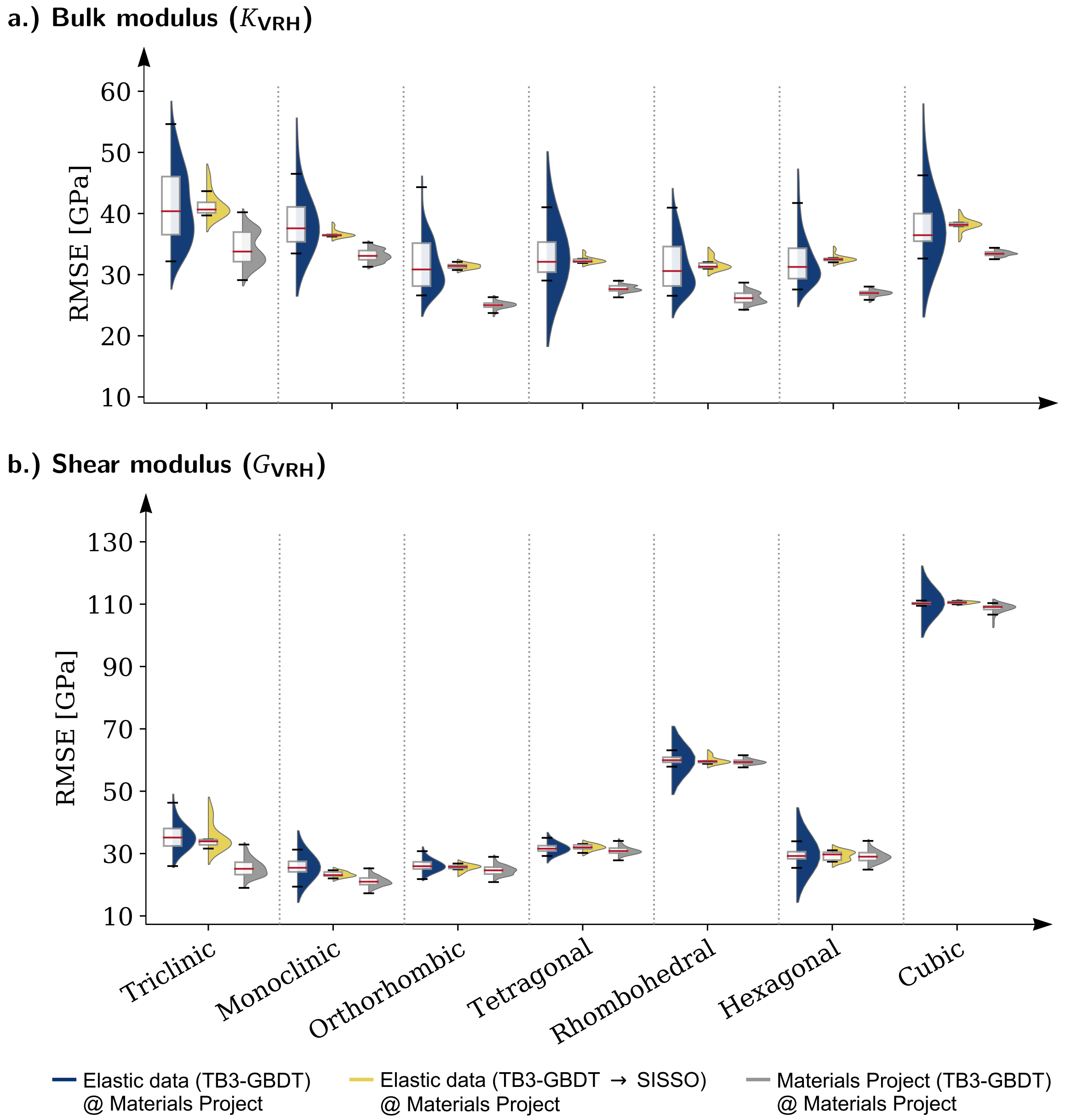


Fig. 5.26. Comparison of the root-mean-squared prediction errors (RMSE) of the bulk (a.) and shear (b.) moduli of three different machine-learning models across all seven crystal lattice systems. The machine-learning models were constructed based on the 1,181 inorganic crystalline compounds [35, 412] data set using TB3-GBDT algorithm (cf., Tabs. 5.17 and 5.18), based on the same data set but using symbolic regression (SISSO [197]) with the most frequent identified feature subsets of TB3-GBDT (Fig. 5.23), and on the Materials Project data set [13, 15, 35] using TB3-GBDT. All machine-learning models were applied to the Materials Project data set. Units are in gigapascal (GPa). Box plots visualize the interquartile range of the distribution, i.e., the 25th- and 75th-percentile of the prediction errors, the median (horizontal line), and the minimum and maximum values of the prediction errors (vertical lines). The distribution shape of the prediction errors are shown on the right of each box plots and were estimated using kernel density estimation with a Gaussian kernel and automatic bandwidth determination.

data set includes compounds of all 7 lattice systems[30] and covers a wider range of space groups (174 instead of 87) and elements (78 instead of 63).

Interestingly, the differences in the prediction errors are only less than 3 GPa and 6 GPa between the generated machine-learning models applied to the Materials Project data and the machine-learning models constructed directly on the Materials Project data (Tab. 5.20). This is surprising as the generated machine-learning models from the elastic data set were not re-generated (cf., Section 2.4) and there is an almost tenfold difference in the amount of data used to build these models. Yet, the elastic-data models either have low variances in the predictions and do not perform as good as a machine-learning model constructed on the full set of data (as in the case for the generated symbolic-regression models from the ensemble of elastic-data models (Fig. 5.23) due to their limited description of the material's behavior, cf., Section 5.2.3 and Fig. 5.24), or the elastic-data models have a large variance in the predictions, but are statistically equivalent to the Materials Project models. Still, the overall good performance of the elastic-data models emphasizes that both the bulk and shear moduli can in principle be described by a set of elemental properties and that their relationships can be represented as simple analytical expressions.

In terms of prediction performance, it is observed that the generated machine-learning models cannot estimate the bulk and shear modulus of the compounds equally well for all crystal structures (Fig. 5.26). In particular, compounds with low or high crystal symmetries (triclinic, monoclinic, and cubic) show larger prediction errors than compounds with orthorhombic, tetra-, and rhombohedral crystal structures independent of the applied machine-learning model. While these differences are only a few gigapascals for the bulk modulus, particularly large errors occur for the shear modulus of rhombohedral and cubic inorganic crystalline compounds. Both rhombohedral and cubic inorganic crystalline compounds have been shown to contribute most to the prediction errors of the machine learning models by having 2–3 times larger standard deviations in the shear modulus than the rest of the compounds in the Materials Project dataset. Furthermore, it has been found that rhombohedral and cubic compounds with shear moduli greater than 300 GPa have standard deviations up to 30 times larger than all other compounds of the same crystal structure. Given that both the elastic-data and Materials Project models (Figs. 5.27 and 5.28) under-estimate bulk and shear moduli greater than 300 GPa, none of the models can therefore reliably estimate ultrahigh bulk and shear modulus of inorganic crystalline compounds.

Among the compounds with the largest prediction intervals are oxide (AlNiO3, CuAgO2, EuGeO3) and nitride materials (Ca3AsN, TaCuN2), but also various other compounds of transition metals such as $BaMg_6Nb$, FeSn or $MgNiH_2$ (cf., Section 5.2.3). Conformal prediction further identifies a few anomalous materials with either relatively low (< 50 GPa) and high (> 200 GPa) bulk and shear moduli. Just as in the modeling of the symbolic-regression models (cf., Section 5.2.3), ensemble-based prediction intervals are smaller than conformal-prediction intervals at the the same confidence level α, but are as large as conformal-prediction intervals at the same validity $\hat{\alpha}$ (Tab. 5.21). Because the elastic-data set comprises a smaller materials space than the Materials Project data set (lattice systems, space groups, and elements), there is no guarantee that the validity of the elastic-data conformal-

[30] Cubic (4,200), hexagonal (1,698), rhombohedral (665), tetragonal (1,402), orthorhombic (566), monoclinic (1,967), and triclinic (108).

prediction intervals correspond to the confidence level on average (cf., Section 4.3.1). As such, in contrast to the Materials Project data set, the validity of elastic-data conformal prediction intervals is smaller than the actual confidence level, but is still substantially larger than estimates based on ensemble predictions. Although the difference between the validity and the confidence level of the conformal prediction lessens with higher confidence levels, the generated machine-learning models of the elastic-data set generally under-estimate the size of the prediction intervals in the screening of new inorganic crystalline compounds.

Due to the lack of materials data for statistical modeling (cf., Section 5.2.3), especially for compounds with bulk and shear moduli above 300 GPa[31], large errors occur in the estimation of elastic properties. As such, for an actual materials-science application or the screening of thousands or more compounds, the developed machine-learning models may be too inaccurate. Nevertheless, the elastic-data models show good prediction performances on the Materials Project data as compared to models constructed directly on the Materials Project data. In particular, the elastic-data models provide useful estimates for triclinic crystal structures, although no inorganic crystalline compounds with triclinic crystal structure were included in the elastic data set. Consequently, the developed feature-identification framework is able to identify a subset of relevant features in the description of the elastic properties of the inorganic crystalline compounds and is able to create generalizable statistical models from a smaller set of materials data with similar prediction performances as compared to statistical models constructed directly on larger sets of materials data.

Summary

The present analysis statistically identifies and characterizes compositional properties that are directly linked to the prediction performance of the generated machine-learning models such as the mean atomization enthalpy (ΔH_{at}) and the molar volume (V_{mol}). Although only averaged atomic features were used, the prediction performances of the generated machine-learning models for the bulk and shear modulus are comparable to previous studies (cf., [166, 412, 470, 473]). It should be emphasized that neither the developed feature-identification framework (Sec. 5.1) nor the model construction is limited to the bulk and shear modulus. In principle, the developed framework can be applied to other derived properties such as the elastic anisotropy or the isotropic Poisson's ratio.

In contrast to previous studies [166, 473], the features of the data set are based on a total of 54 fundamental, compositional, and structural properties that can be easily derived from tabulated elemental properties for a machine-learning model without having to compute n-wise combinations or to use experimentally demanding and computationally intensive first-principles calculations. Because some of the feature-identification methods were too computationally intensive to be applied to the full set of features (among them symbolic-regression algorithms), the main challenge of the data set was to reduce the large number of features prior to machine learning (cf., Section 5.1.2).

Feature-identification methods were characterized by a high variability in the identification of relevant features, while strong feature interactions complicated the identification of relevant features and the construction of machine-learning models. Results demonstrated that the feature-dependence

[31] The Materials Project data set contains only 154 and 39 inorganic crystalline compounds with bulk and shear moduli larger than 300 GPa, or less than 1.5% and 0.4% of all compounds in the data set.

↵ **Algorithm**		**Ensemble prediction**			**Conformal prediction**		
	Confidence level (α)	Validity ($\hat{\alpha}$)	Mean (Δ)	Max (Δ)	Validity ($\hat{\alpha}$)	Mean (Δ)	Max (Δ)
a.) Bulk modulus (K_{VRH})							
Elastic data (TB3-GBDT)	0.50	*0.19*	6.5	49.2	*0.33*	12.3	76.8
	0.80	*0.40*	14.0	161.0	*0.60*	27.2	171.9
	0.95	*0.61*	23.3	895.4	*0.84*	66.5	417.4
Elastic data (TB3-GBDT → SISSO)	0.50	*0.03*	1.3	19.2	*0.32*	13.0	69.0
	0.80	*0.08*	2.7	33.3	*0.58*	29.2	152.4
	0.95	*0.14*	5.3	80.8	*0.85*	78.8	414.3
Materials Project (TB3-GBDT)	0.50	*0.18*	4.4	44.0	*0.49*	15.7	267.1
	0.80	*0.32*	8.2	78.2	*0.80*	34.1	577.5
	0.95	*0.47*	12.5	108.1	*0.96*	85.3	1455.1
b.) Shear modulus (G_{VRH})							
Elastic data (TB3-GBDT)	0.50	*0.21*	5.4	66.5	*0.33*	9.4	62.5
	0.80	*0.40*	10.9	125.5	*0.64*	23.3	156.0
	0.95	*0.61*	18.1	309.9	*0.90*	61.0	412.4
Materials Project (TB3-GBDT)	0.50	*0.11*	3.3	45.8	*0.34*	10.4	60.0
	0.80	*0.18*	5.4	66.9	*0.68*	25.4	145.3
	0.95	*0.22*	6.5	73.9	*0.90*	56.8	325.4
Materials Project (TB3-GBDT)	0.50	*0.23*	5.7	56.4	*0.49*	14.0	494.0
	0.80	*0.44*	11.1	264.7	*0.80*	29.5	1037.3
	0.95	*0.64*	17.8	731.6	*0.96*	73.2	2581.2

Tab. 5.21. Validity of ensemble and conformal prediction of TB3-GBDT from three machine-learning models at three different confidence levels $\alpha = [50\%, 80\%, 95\%]$. The validity ($\hat{\alpha}$) specifies the probability ($\hat{\alpha}\% = 100\hat{\alpha}$) that the target property is within the prediction interval $x \in [\bar{x} - \bar{\Delta}, \bar{x} + \bar{\Delta}]$ of the statistical model (Eq. 4.22), with $\bar{x}$ and $\pm\bar{\Delta}$ ($\bar{\Delta} = \mathrm{Mean}(\Delta) \leq \mathrm{Max}(\Delta)$) being the (point) estimation and variance of the ensemble or conformal prediction. The validity of the ensemble prediction intervals were computed as the αth-percentile of the ensemble machine-learning predictions. The validity and prediction intervals of conformal prediction were estimated using the setup as described in Sections 4.3.2 and 5.1. Performance statistics of the bulk (K_{VRH}, a) and the shear modulus (G_{VRH}, b) notation [485] were approximated using the sure-independence screening and sparsifying operator (SISSO) [197] with the approach as described in Section 5.2.3. The machine-learning models were constructed based on the 1,181 inorganic crystalline compounds [35, 412] data set using TB3-GBDT algorithm (cf., Tabs. 5.17 and 5.18), based on the same data set but using symbolic regression (SISSO [197, 371]) with the most frequent identified feature subsets of TB3-GBDT (Fig. 5.23), and on the Materials Project data set [13, 15, 35] using TB3-GBDT. All machine-learning models were applied to the Materials Project data set. Units are in gigapascal (GPa).

maps exhibit a block-like structure of feature interactions (Fig. 5.22), both the bulk and shear modulus can in principle be described by the same set of features (Fig. 5.23), and the developed feature-identification framework (cf., Section 3.5) substantially simplifies the construction of machine-learning models as compared to TCMI, RFECV, and FS-GBDT (Tabs. 5.15 and 5.16 and Fig. 5.23). In addition, it was demonstrated that the developed feature-identification framework is able to relate the input features to the prediction performance of the generated machine-learning models, while providing a means to investigate the statistical relationships using symbolic-regression models (Fig. 5.23). Though the generated symbolic-regression models could not fully describe the material's behavior (Fig. 5.24), they have been shown to be capable of estimating the bulk and shear modulus for a wide range of k-nary compounds of different chemical compositions and structures (Fig. 5.23), even for compounds that were not included in the training data set (Fig. 5.26). It should be noted that such models (and machine-learning models in general) only capture an overall statistical trend in the data and therefore cannot be regarded as a physical law [55]. In particular, there is the risk of inaccurately modeling some of the statistical trends with the machine-learning models, which may significantly limit their applicability to new data. For example, the symbolic-regression models (Fig. 5.24) correctly captured the statistical trends of the bulk and shear modulus as a function of the marginal contributions of the mean atomization enthalpy (ΔH_{at}) and the molar volume (V_{mol}), but failed to reproduce the statistical trends of the mean fusion enthalpy ($\Delta H_{\mathrm{fusion}}$), heat of formation ($H_{\mathrm{form}}$), bulk ($B$) and Young's ($Y$) modulus.

Overall, constructed machine-learning models were affected by large errors in the prediction of oxide and nitride materials due to the lack of inorganic crystalline compounds in the data set for the estimation of ultrahigh elastic moduli ($\geq$ 300 GPa, Section 5.2.3). For example, at a confidence level of $\alpha = 0.95$ prediction intervals were on average about three to four times larger than the root-mean-squared error of the generated symbolic-regression models. While ensemble-based prediction intervals were always much lower than the specified confidence level, conformal-prediction intervals were only lower when the elastic-data models were applied to the Materials Project data set (Section 5.2.3). As a result, the generated machine-learning models have the tendency to under-estimate the size of the prediction intervals and the errors made in the machine-learning predictions. Moreover, it has been shown that machine-learning models are not reliable in estimating ultrahigh bulk and shear moduli of inorganic crystalline compounds and hence are not useful for screening materials with bulk or shear moduli greater than 300 GPa. It may be possible though to create a machine-learning model specifically for ultrahigh bulk or shear moduli by identifying anomalous materials (Sec. 4.4) or a domain of applicability [208], but the small number of materials may actually prevent a practical screening application using these models.

The present study can be considered as one of the largest studies to date on the extrapolation behavior of machine-learning models for estimating elastic properties of inorganic crystalline compounds. The efficiency of the developed framework was therefore of central importance for the identification of relevant features and the construction of predictive models. Despite the limitations of the generated machine-learning models, it was highlighted that the developed feature-identification framework is capable of (1) modeling the overall elastic behavior of the inorganic-crystalline compounds by representing the statistical relationships as simple analytical expressions and (2) building generalizable machine-learning models from a smaller set of materials data with similar prediction performance as models built directly on a larger materials data set.

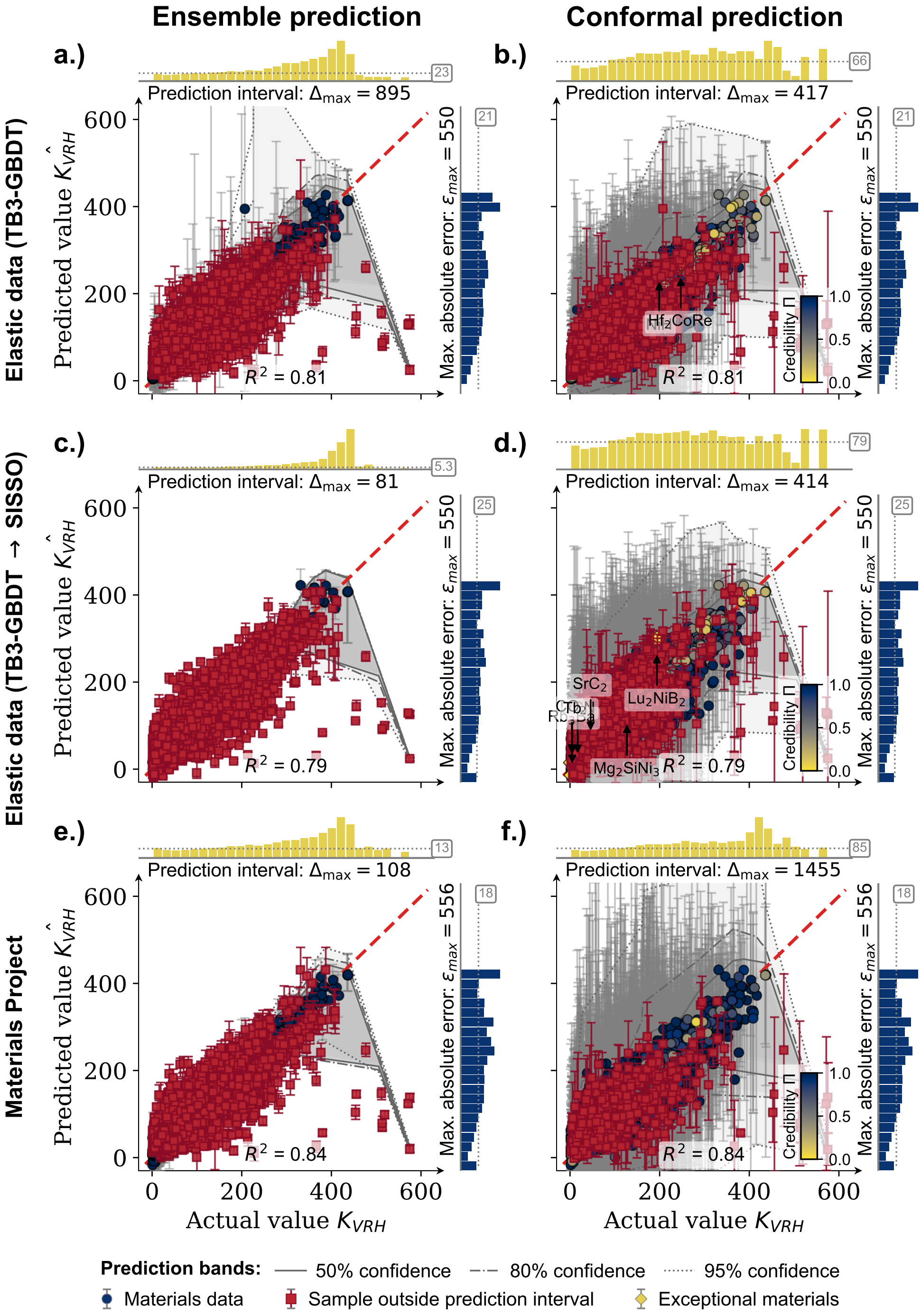


Fig. 5.27. Ensemble prediction performance of machine-learning models from TB3-GBDT using the sure-independence screening and sparsifying operator (SISSO) [197] to estimate the bulk (K_{VRH}) modulus in the Voigt-Reuss-Hill notation [485] from the set of statistically equivalent feature subsets. The prediction errors of SISSO were approximated using the setup as described in Section 5.2.3. Shown are the prediction bands (50th, 80th, 95th-percentiles) of the model's predictions, the credibility Π (Eq. 4.29) at a confidence level of $\alpha = 0.95$, the distribution of the size of prediction intervals (diagram above the parity plot, Δ), and the errors of the machine-learning model (diagram right of the parity plot, $\varepsilon = |a_0 - \hat{a_0}|$). The numbers in the boxes display the mean values ($\bar{\Delta}$, $\bar{\varepsilon}$), while the maximum errors are given in the texts below or to the left of the diagrams. The machine-learning models were constructed based on the 1,181 inorganic crystalline compounds [35, 412] data set using TB3-GBDT algorithm (cf., Tabs. 5.17 and 5.18), based on the same data set but using symbolic regression (SISSO [197]) with the most frequent identified feature subsets of TB3-GBDT (Fig. 5.23), and on the Materials Project data set [13, 15, 35] using TB3-GBDT. All machine-learning models were applied to the Materials Project data set. Units are in gigapascal (GPa). Machine-learning estimates outside the prediction intervals are depicted as squares and anomalous materials as diamond-shape symbols.

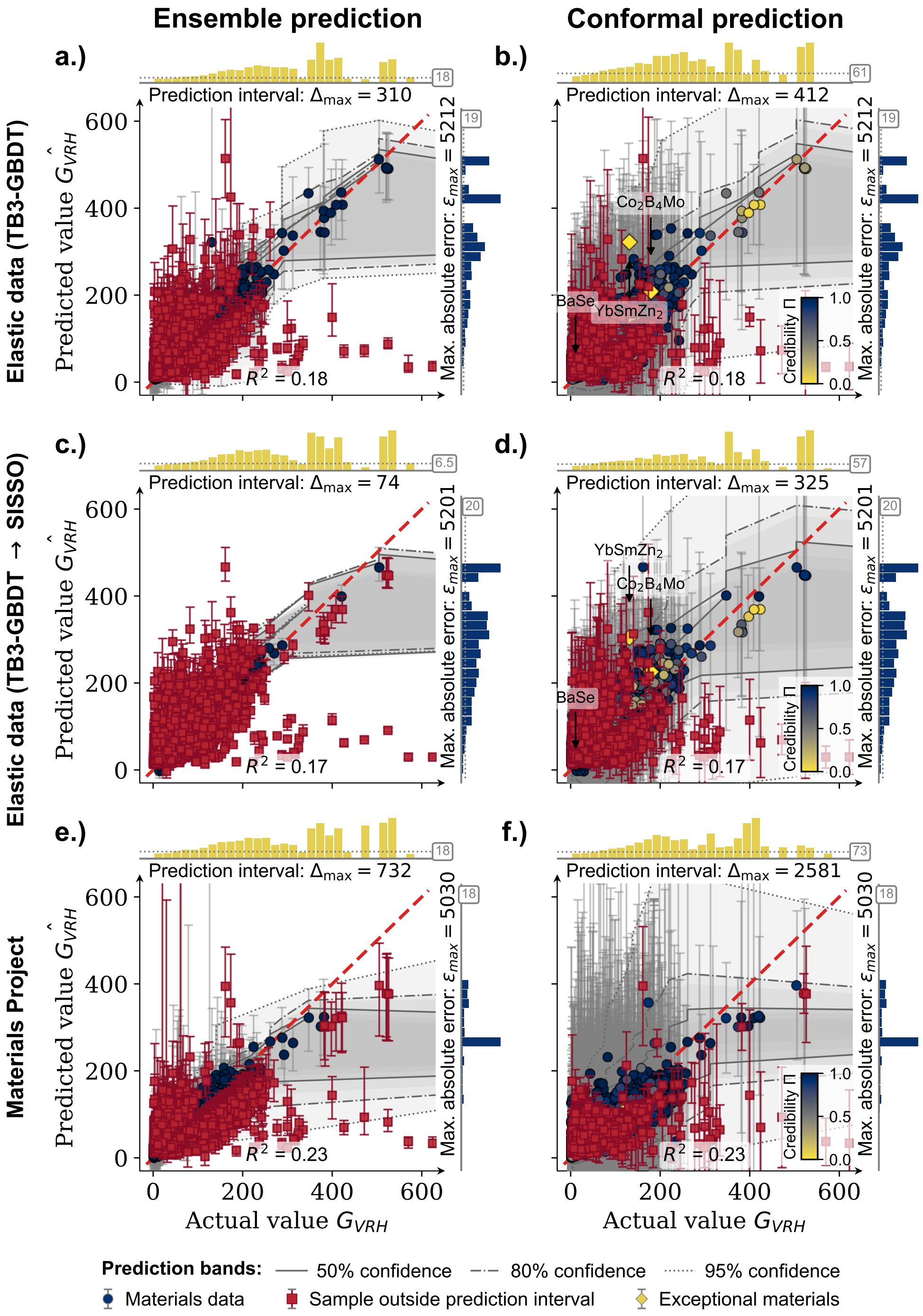


Fig. 5.28. Ensemble prediction performance of machine-learning models from TB3-GBDT using the sure-independence screening and sparsifying operator (SISSO) [197] to estimate the shear (G_{VRH}) modulus in the Voigt-Reuss-Hill notation [485] from the set of statistically equivalent feature subsets. The prediction errors of SISSO were approximated using the setup as described in Section 5.2.3. Shown are the prediction bands (50th, 80th, 95th-percentiles) of the model's predictions, the credibility Π (Eq. 4.29) at a confidence level of $\alpha = 0.95$, the distribution of the size of prediction intervals (diagram above the parity plot, Δ), and the errors of the machine-learning model (diagram right of the parity plot, $\varepsilon = |a_0 - \hat{a_0}|$). The numbers in the boxes display the mean values ($\bar{\Delta}$, $\bar{\varepsilon}$), while the maximum errors are given in the texts below or to the left of the diagrams. The machine-learning models were constructed based on the 1,181 inorganic crystalline compounds [35, 412] data set using TB3-GBDT algorithm (cf., Tabs. 5.17 and 5.18), based on the same data set but using symbolic regression (SISSO [197]) with the most frequent identified feature subsets of TB3-GBDT (Fig. 5.23), and on the Materials Project data set [13, 15, 35] using TB3-GBDT. All machine-learning models were applied to the Materials Project data set. Units are in gigapascal (GPa). Machine-learning estimates outside the prediction intervals are depicted as squares and anomalous materials as diamond-shape symbols.

5.3 Discussion

The three materials-science applications considered in this chapter were selected based on the availability of curated materials data. All three data sets proved challenging in terms of feature identification and were large and diverse enough to estimate the properties of interest with machine learning. In the first application, the energy difference between two crystal structures of octet-binary compound semiconductors was modeled. The most frequently identified features were the atomic radii of the *s*- and *p*-orbital, the Mulliken electronegativity, and the electron affinity. In the second application, the lattice constant and bulk modulus of perovskite oxides were studied. The atomic radii of the *s*- and *p*-orbital, the nuclear and ionic charges, as well as the electron affinity were the most frequently identified features. In the third application, the bulk and shear modulus of inorganic crystalline compounds were estimated, where the most frequently identified features were the mean atomization enthalpy and the molar volume.

Common to both the octet-binary compound semiconductors and the perovskite oxides data set is a strong feature interaction between features from each atomic species. Due to these feature interactions, many features were identified as relevant, with each feature subset containing at least one feature from each atomic species. In both data sets the materials are uniquely defined by their two constituent elements, i.e., confirming that at least one feature of each atomic species is needed to actually describe the property of interest. In contrast, the elastic data set contains materials with varying number of constituent elements. Therefore, properties of the constituent elements were averaged. Strong feature interactions were found between features of each type (fundamental, compositional, and structural properties). Due to the larger number of features requiring higher computational efforts as compared to the other data sets, not all machine-learning algorithms were applicable with the full set features. Feature identification prior to model construction reduced the computational complexity and allowed these machine-learning algorithms to be applied to the data set using the identified feature subsets.

In the octet-binary compound semiconductors data set, boron nitride and diamond were identified as anomalous materials. In the elastic data set, oxide and nitride materials were found to be anomalous. All these anomalous materials are characterized by a large prediction uncertainty due to lack of similar materials. They are underrepresented in the data set and hence cannot be reliably predicted. No anomalous materials were found in the perovskite oxides data set, as the values of the properties of interest are more evenly distributed across the materials of the data set.

The search for an accurate machine-learning model for predicting the property of interest includes a comparison of the prediction performance for each machine-learning model on the same set of features. As it turns out, all feature-identification methods from Section 3.5 identify more features as relevant than the TB3-algorithm with a larger variance in the prediction errors. Although FS-SISSO, RFECV, and FS-GBDT were performed with 10-fold cross-validation, the TB3-algorithm performs cross-validation during feature-subset search and as such provides more stable feature subsets and lower variances in the prediction errors. All machine-learning models based on the feature subsets identified by the TB3-algorithm are competitive with or are better than the models reported in the literature in terms of prediction performance and the required number of features [55, 161, 165, 166, 186, 410, 411, 470, 473, 474].

A model-independent feature selection with TCMI characterizes features as relevant only, if there is a sufficient number of materials in the data set showing a statistical relationship between the features and the property of interest. This is in contrast to machine learning, which improves as more data are available. By using an evaluation measure from a machine-learning model as a feature-selection criterion (e.g., the Pearson's determination coefficient R^2), the identification of relevant feature subsets can be directly linked to the prediction performance of the machine-learning models. Higher scores of the identified feature subsets are thus equivalent to higher prediction performances of the generated models.

Both GBDT and SISSO models were used for feature identification with the same feature-selection criterion (the Pearson's determination coefficient) and search strategy (the TB3-algorithm). Whereas a feature-subset search with the GBDT algorithm (TB3-GBDT) is based on constructing piecewise-constant models, a search with the SISSO algorithm (TB3-SISSO) is based on constructing functional relationships between a set of features and the property of interest. Generally, TB3-GBDT found feature subsets with a smaller cardinality, but with similar prediction performance as TB3-SISSO. However, in cases where the property of interest can be modeled more accurately with a functional interpolation than with piecewise-constant models, TB3-SISSO identified more compact feature subsets than TB3-GBDT (in the case of the shear modulus of the elastic data set) or achieved lower prediction errors (in the case of the octet-binary compound semiconductors data set).

The Jaccard similarity coefficient was used to compare feature subsets found with TB3-GBDT and TB3-SISSO. However, only a fraction of the identified feature subsets were found by both methods: first, because the search was not exhaustive. And second, the models required different number of features to achieve optimal performance. Nevertheless, a frequency analysis showed that many features were shared across the two different methods.

In order to compare the prediction performance of identified feature subsets, models were built with SISSO. Many different models based on different feature subsets achieved similar prediction performance. Each model can be considered as a single representation of the relationships between the features of the data set and the property of interest. Since many models are statistically equivalent, the physical interpretation of a single model is not meaningful.

In all materials-science applications considered in this chapter, elemental properties of the constituent elements were used to represent the materials as a function of their chemical composition and stoichiometry. Since features based on elemental properties do not require any information other than the elements and stoichiometry, they can be used to screen and estimate the target properties of materials whose structure are not known. However, machine-learning models based solely on the elemental properties neither account for atomic interactions nor for structure-related phenomena such as high-temperature superconductivity [167, 512, 513]. As such, these models are usually not based on any physical models of the target properties and therefore are expected to require either larger amounts of materials data or a combination of the features, if a highly predictive model can be generated at all. Moreover, elemental properties are strongly statistically related because both the features and the property of interest are determined by the Kohn-Sham equations with the atomic species, charges, and positions as the only physically relevant input variables. Therefore, statistical relationships between a subset of the relevant features and the properties of interest cannot be regarded as physical laws.

All machine-learning models were able to qualitatively capture the overall trend in the data, but were unreliable when estimations were made in regions of the materials space that were not part of the training data. The difference between the estimated and actual values of the property of interest were quite large. The multiplicity of the machine-learning models was therefore used to compute prediction intervals. Further, a heuristic measure, called credibility, was introduced to identify anomalous materials, which cannot be predicted well by the machine-learning models. Prediction intervals and the credibility measure can thus help to explore the materials space more efficiently and guide further investigations of materials with large prediction errors or low credibility.

Quantitatively, all investigated machine-learning algorithms had problems in accurately representing the statistical trends in the data: the applicability of the generated machine-learning models therefore remains limited and a problem yet to be solved in a potential high-throughput application for discovering new materials with targeted materials properties.

Two criteria give confidence that the developed framework is applicable to a wider range of materials-science applications: First, the analysis is neither limited to a particular machine-learning algorithm nor to the specific materials-science application. And second, the framework creates predictive machine-learning models with fewer features but with the same prediction performance as compared to a machine-learning model built on the full set of features. Nevertheless, the developed framework for feature-identification and model construction may not perform well for all materials-science applications, e.g., on imbalanced data sets or in cases where the features of the data set are completely unrelated to the properties of interests.

Chapter 6

Conclusion

With the increasing availability of materials data from first-principles electronic-structure codes and high-throughput experiments, machine learning is taking on an important role in materials science, offering new techniques for analyzing materials properties, screening materials spaces, and designing new materials. To date, many applications have been investigated [55, 161–168, 171, 172, 177, 179, 335], but the identification and characterization of relationships between one or more features and the creation of highly predictive machine-learning models are still challenging.

6.1 Challenges

The first challenge is related to the choice of relevant features: materials data sets can be represented in terms of hundreds to thousands of features. As a result, the generation of statistical models places high demands on computational resources. Sets of features are therefore constructed iteratively [55, 161], selected on the basis of predefined criteria or intuition [171, 172, 177, 179], or introduced without extensively and systematically analyzing their relevance [162, 163, 165–168, 335]. The second challenge is related to the multi-collinearity of features: different feature subsets may lead to the same prediction performance. As such, selecting only one feature subset can neither fully represent all the relationships present in the data nor ensure a robust prediction of the machine-learning models [358]. The third challenge is related to the inherent statistical modeling of feature-property relationships: Even though statistical models are optimized for the highest prediction performance, the prediction errors of the models are typically unknown, leading to potentially large uncertainties in the predictions when applied to new data.

6.2 Feature identification and model construction

With the framework and tools developed in this thesis, the complexity of machine-learning models can be reduced by identifying the relevant features for estimating the property of interest prior to machine learning. An extensive and systematic search of relevant features was carried out, identified feature subsets were related to the prediction performance of the machine-learning models, and (based on these subsets) an ensemble of machine-learning models was constructed to estimate the prediction uncertainty on new data. Combining all these steps, the developed framework deals comparatively well with small data sets (of about 50 data samples) and relatively large number of features (more than 50 features). In particular, it successfully identifies sets of features that lead to simpler predictive models without significantly decreasing the prediction performance of the generated machine-learning models.

6.2.1 Feature identification

In Chapter 3, sets of features related to the property of interest were identified using a score, which allows feature subsets to be ranked and ordered by relevance. There is a plethora of methods, all with different criteria for quantifying the relevance of features. One of these is feature selection with information theory. Information-theoretic feature-selection methods do not require the explicit modeling of the actual relationship and further have the advantage of being deterministic and providing a non-parametric quantification of relevance based on mutual dependence between a set of features and a property of interest. However, current information-theoretic methods involve probability densities which cannot be obtained directly from (real-valued) sample data, such as those found in materials science. Thus, a generalization called total cumulative mutual information (TCMI) was developed in this thesis, which is based on cumulative probability distributions. Cumulative probability distributions can be calculated directly from the data set without the need for discretization or additional parameters. The cumulative mutual dependence as obtained by TCMI was further corrected with respect to the baseline dependence when all features are completely independent prior to estimating the relevance of feature subsets.

The search for relevant feature subsets is a combinatorial optimization problem. As an exhaustive search is impractical for data sets with large number of features, the branch-and-bound algorithm was used. The branch-and-bound algorithm enumerates all combination of features and stops exploring further subsets whose feature-selection criterion (e.g., the score from TCMI) cannot be improved.

Finally, three examples with known as well as empirically identified feature-property relationships were discussed and TCMI was compared with existing methods for identifying relevant features. Overall, feature-selection methods based on machine learning are sensitive to multi-collinear features and are therefore not reliable for identifying the relevance of features. In contrast, TCMI is stable with increasing numbers of data samples, but requires more data to identify the same set of features than the other methods.

6.2.2 Conceptual framework

In Chapter 4, a framework for feature identification was developed that is applicable to any information-theoretic method or machine-learning algorithm. In the framework, the branch-and-bound algorithm was extended to work with non-monotonically increasing feature-selection criteria. This so-called tolerance-based branch-and-bound (TB3) algorithm identifies feature subsets that are close to the optimal subset within a specified tolerance. All identified feature subsets were then used to build machine-learning models. Because models based on different subsets led to similar prediction performances, a probabilistic threshold was introduced to determine which models are statistically equivalent. This ensemble of models was then used to estimate the prediction uncertainty by utilizing the ensemble mean and standard deviation to define a prediction interval for each prediction. In addition, conformal prediction was applied to estimate the prediction intervals based on the available materials data. Unlike ensemble-based prediction intervals, conformal prediction statistically ensures that the actual value of the targeted property lies within the prediction interval at a given confidence level. A comparison between these two approaches showed that ensemble-based prediction intervals under-estimate the error made in the prediction of the machine-learning models and uncertainty estimates from conformal prediction only hold on average. Therefore, a measure was developed (called credibility) to identify anomalous materials that cannot be predicted well by the machine-learning models. Credibility is calculated heuristically by comparing the prediction of a new material to known values of similar materials in the data set. In a simplistic example with known ground truth, the credibility measure identified all materials whose actual values could not be reliably estimated by the machine-learning model and were outside the prediction interval as estimated by conformal prediction within the specified probabilistic tolerance.

Like TCMI, the scaling behavior of the developed framework has a worst-case exponential computational time complexity in the number of features, but largely depends on the information-theoretic method or machine-learning algorithm in the scaling behavior on the number of data samples. However, by using branch and bound, which terminates the search for relevant subsets of features as early as possible, the TB3-algorithm can be used even when other feature-identification methods are no longer applicable (cf., Section 5.2.3).

6.2.3 Materials-science applications

Finally, in Chapter 5, TCMI, two machine-learning methods combined with the TB3-algorithm, and three further feature-identification methods were applied to the quantitative prediction of the crystal structure of octet-binary compound semiconductors, the prediction of structural properties of perovskites, and the prediction of elastic properties of inorganic crystalline compounds. The SISSO algorithm was used as the reference machine-learning algorithm for estimating the prediction performance of identified feature subsets. All methods resulted in similar prediction performance of the generated machine-learning models. However, unlike machine-learning methods combined with the TB3-algorithm, all other methods identified more features as dependent than were actually needed to generate predictive machine-learning models. In comparison, the TB3-algorithm achieved the best prediction performance in all applications with the least number of features.

Machine-learning models from the octet-binary compound semiconductors data set were characterized by large prediction errors and strong feature inter-correlations between the features of each atomic species: At least one feature from each atomic species was required to estimate the energy difference between the two crystal structures, rock salt and zinc blende. Boron nitride and diamond have been found to be anomalous materials: They exhibit a rare or otherwise unusual energy difference between rock salt and zinc blende crystal structures compared to the investigated data set of octet-binary compound semiconductors, consistent with the fact that they are the hardest naturally occurring materials on Earth.

Perovskites oxides are uniquely determined by the nuclear and ionic charges of the two non-oxygen constituents. These features have been consistently identified as relevant by all feature-identification methods. All generated machine-learning models performed well in predicting the lattice constant, but only moderately well in predicting the bulk modulus. The prediction of the bulk modulus became more accurate when the lattice constant was included as a feature. No anomalous materials were found in the data set.

Due to the varying number of constituent elements, averaged atomic properties were used to estimate the bulk and shear moduli of inorganic crystalline compounds. All feature-identification methods were characterized by a high variability in the identification of relevant features. Overall, constructed machine-learning models were affected by large errors in the prediction of oxide and nitride materials due to the lack of similar materials with ultrahigh elastic moduli in the data set. Both the bulk and shear modulus can in principle be described by the same set of features. Although the machine-learning models generated by SISSO could not fully capture the statistical relationship in the data, they were capable of estimating the bulk and shear modulus for a wide range of k-nary compounds of different chemical compositions and structures. In addition, their prediction performance was comparable to that of models constructed using all features, even when applied to a much larger data set of inorganic crystalline compounds than was available for statistical modeling.

6.2.4 Summary

The developed framework and additional diagnostic tools for feature identification and model construction provide some approaches to solving the above-mentioned challenges. To start with, feature identification was addressed by a systematic search using the (tolerance-based) branch-and-bound algorithm. Further, problems with multi-collinearity were addressed by exploring simple to more complex feature subsets. In particular, the inherent statistical nature of predictions from the statistical modeling was addressed by estimating the uncertainty in the model predictions based on an ensemble of models rather than a single model. In addition, a heuristic was developed to identify materials that cannot be predicted well.

A model-independent identification of feature subsets was aimed for, but failed for several reasons: first, the limited availability of materials data induced spurious relationships in the data. Second, a comparison of identified feature subsets across different machine-learning algorithms was hindered by the limited flexibility of the applied methods due to inherent assumptions made in the statistical modeling. Third, a model-independent feature-selection criterion such as TCMI identified fewer number of features as relevant as other feature-identification methods. As such, models built using these feature subsets did not achieve the same performance as models based on feature subsets identified using the other methods. Finally, a partial exploration of feature combinations in the search, as opposed to exhaustive search, prohibited the identification of all relevant (non-optimal) feature subsets in a data set.

Overall, the developed framework is effective in reducing the number of features in a data set prior to statistical modeling of relationships between a set of features and the property of interest. Using feature identification prior to machine learning has the advantage that, first, the number of features can be reduced independently of a machine-learning algorithm and, second, a more effective feature-selection criterion can be used to handle larger volumes of data. In addition, the framework goes beyond the pairwise identification of related features (that is used in essentially all feature-identification methods presented in the materials-science literature) by identifying the features that are multivariately and non-linearly related to a property of interest prior to statistical modeling. Thus, rather than expending time trying to intuit the relevant features with a feature-identification method specific to a machine-learning algorithm or from pairwise relationships, the developed framework can be utilized as a standardized procedure to automatically identify the relevant set of features in a data set as well as to generate predictive models including prediction intervals in the discovery and design of new materials.

6.3 Outlook

A data-driven extensive and systematic identification of features is pivotal for achieving a physical understanding of the existing relationships between a set of features and a property of interest. Currently, TCMI has exponential time complexity with increasing number of features. Thus, the faster feature identification based on machine learning seems to be the most viable approach in creating predictive models from data sets with a large number of features. However, TCMI can potentially be optimized by introducing an effective summation scheme to improve the computational efficiency of the cumulative joint mutual dependence (similar to the baseline adjustment approach of TCMI). There is also further potential for improvement in the construction of machine-learning models. For example, symbolic-regression models could be optimized by automatically adjusting the required complexity of the model. In addition, this thesis focused exclusively on available curated materials data sets. In practical application, it would be useful to interactively search the materials space for materials with desired properties. By using the tolerance-based branch-and-bound algorithm to first identify the features related to a property of interest and then constructing a statistical model, an active-learning workflow [102–104] could be designed that suggests new materials based on the developed credibility measure in this thesis for computation or synthesis that have a low credibility score. Last but not least, it would be desirable if machine-learning models could be optimized for their ability to represent the trend in the data rather than for prediction performance. Even though machine-learning models perform well, they have narrow applicability outside their training scope and therefore do not generalize well (cf., Section 5.2). Although the framework has been applied to only a few materials-science data sets in this thesis, it is promising to apply it to a wider range of applications and models. So far, the framework was only used to predict materials properties, but it would also be interesting to combine the framework with techniques for extracting fundamental equations from data [190, 514].

Appendix A

Machine-learning algorithms

A.1 The sure-independence screening and sparsifying operator (SISSO)

The sure-independence screening and sparsifying operator (SISSO) [197] is a compressed-sensing method [191] for generating explicit, analytic symbolic-regression models $\hat{f}$ based on the algebraic combination of physical quantities $\vec{X} = \{X_1, \ldots, X_d\}$ (the features of the data set). SISSO can be applied to millions and billions of feature-candidate combinations and is not affected by inter-correlated features [197, 371, 515].

Sure-independence screening (SIS [371, 515]) means that SISSO optimally estimates the property of interest from a set of initial or constructed candidate-feature combinations with a probability approaching one, the more candidate-feature combinations are screened for fitting the machine-learning model to the samples of the data set. Sparsifying means that SISSO uses a sparse-solution algorithm (a sparsifying operator, SO [516]) to build a machine-learning model of dimensionality *desc_dim*[1] based on a component-wise regression or equivalently correlation-learning technique to search the space and reduce the number of feature combinations (the so-called descriptors d_i) in the final model,

$$\hat{f}(\vec{X}) = \sum_{i=0}^{desc\_dim} c_i d_i(\vec{X}) \ , \quad d_0 \equiv 1 \ . \tag{A.1}$$

The coefficients c_i are obtained from a least-squares solution of fitting $\hat{f}(\vec{X})$ to the material's property of interest Y, with d_i as the i-th descriptors of the "*desc_dim*"[1]-dimensional model $\hat{f}$. The feature combinations are formed by recursively applying a set of functional/algebraic (unary, binary, etc.) operators

$$\begin{aligned} \Phi_k &= \Upsilon[\vec{\Phi}_{k-1}, \ldots, \vec{\Phi}_0] \ , \qquad\qquad k = 1, \ldots, rung^{1} \\ \Upsilon &= \left\{\cdot + \cdot, \cdot - \cdot, \cdot \times \cdot, \cdot/\cdot, |\cdot - \cdot|, \cdot^{-1}, \cdot^{2}, \cdot^{3}, \cdot^{1/3}, \exp(-\cdot), \exp\cdot, \ln\cdot, \sqrt{\cdot}\right\} \end{aligned} \tag{A.2}$$

[1] The name refer to the setting defined in the Fortran code of the SISSO paper [197].

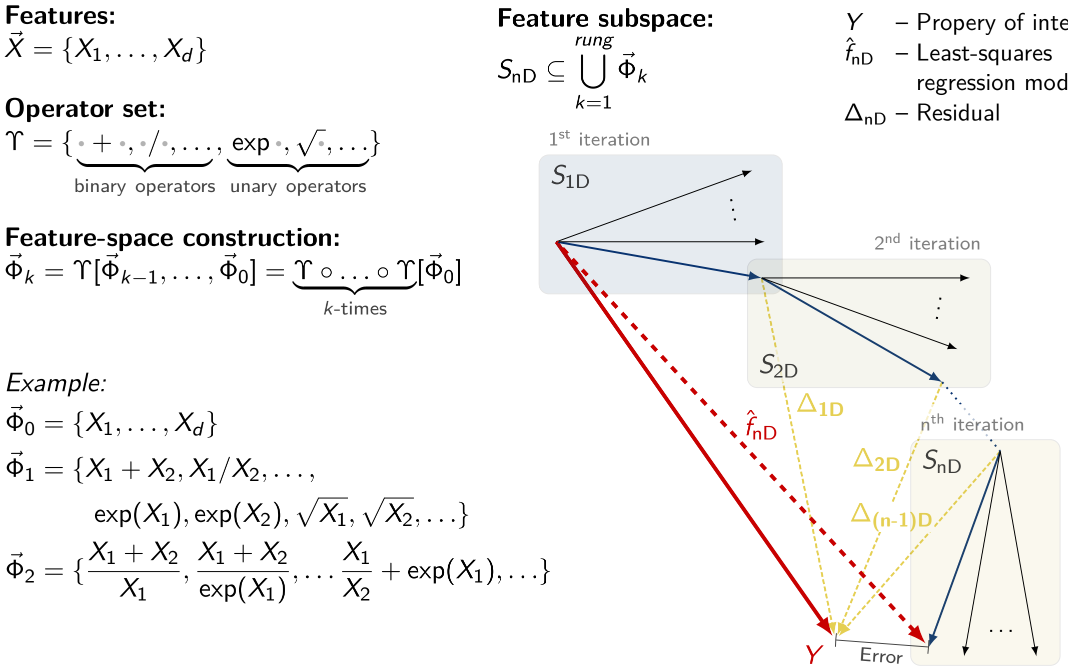

Features:
$\vec{X} = \{X_1, \dots, X_d\}$
Operator set:
$\Upsilon = \{\underbrace{\cdot + \cdot, \cdot/\cdot, \dots}_{\text{binary operators}}, \underbrace{\exp\cdot, \sqrt{\cdot}, \dots}_{\text{unary operators}}\}$
Feature-space construction:
$\vec{\Phi}_k = \Upsilon[\vec{\Phi}_{k-1}, \dots, \vec{\Phi}_0] = \underbrace{\Upsilon \circ \dots \circ \Upsilon}_{k\text{-times}}[\vec{\Phi}_0]$
Example:
$\vec{\Phi}_0 = \{X_1, \dots, X_d\}$
$\vec{\Phi}_1 = \{X_1 + X_2, X_1/X_2, \dots, \exp(X_1), \exp(X_2), \sqrt{X_1}, \sqrt{X_2}, \dots\}$
$\vec{\Phi}_2 = \{\frac{X_1+X_2}{X_1}, \frac{X_1+X_2}{\exp(X_1)}, \dots \frac{X_1}{X_2} + \exp(X_1), \dots\}$
Feature subspace:
$S_{\mathrm{nD}} \subseteq \bigcup_{k=1}^{rung} \vec{\Phi}_k$
Y – Propery of interest
$\hat{f}_{\mathrm{nD}}$ – Least-squares regression model
Δ_{nD} – Residual
1st iteration
S_{1D}
2nd iteration
S_{2D}
Δ_{1D}
$\hat{f}_{nD}$
nth iteration
Δ_{2D}
S_{nD}
$\Delta_{(n-1)D}$
Y
Error

Fig. A.1. Feature combinations are generated by recursively applying a set of functional/algebraic (unary, binary, etc.) operators Υ to the features $\vec{X}$ of a data set. The resulting candidate feature-combinations S_{nD} are then scored with a metric (correlation magnitude, i.e., the absolute of inner product between the feature combination and the property of interest Y or the residual Δ_{nD} in subsequent iterations). Next, a least-squares symbolic regression model is built to select the next subspace of candidate-feature combinations and to reduce the residual error Δ_{nD} to the property of interest in each step. This procedure is repeated iteratively until the error is within the expected error tolerance.

to a subset of (constructed) features, e.g., $\vec{\Phi}_0 \subseteq \{X_1, \dots, X_d\}$, where "$\cdot$" is a placeholder for any constructed feature combination in $\vec{\Phi}_{k-1}$. New feature combinations are then formed by recombining previous feature combinations. As the number of constructed feature-combinations scales exponentially with the number of recursive feature combinations (*rung*[1]), the SIS-step of SISSO scores each candidate-feature combination (standardized) with a metric (correlation magnitude $\langle\cdot,\cdot\rangle$, i.e., the absolute of inner product between the residual and the candidate-feature combination) and keeps only the top-ranked *subs_sis*[1] feature combinations S_{nD} with less than or equal *maxcomplexity*[1] features of the data set (Fig. A.1). Generally, the larger the set, the higher the probability it contains the optimal descriptor [371, 515].

SISSO then iteratively selects a set of candidate-feature combinations S_{nD} in a beam-search-like approach, effectively reducing the residual error Δ_{nD} to the property of interest in each step. The residual error of an n-dimensional model is defined as $\Delta_{\mathrm{nD}} = Y - \sum_{i=0}^{n} c_i d_i(\vec{X})$ with c_i as the coefficients from the solution of fitting the descriptors $\vec{d} = \{d_1, \dots, d_n\}$ to the property of interest Y. In the first iteration, the first-ranked feature of

$$S_{\mathrm{1D}} = \left\{ Z | Z \in \bigcup_{k=1}^{rung^1} \Phi_k, \text{sorted in desc. order of } \langle Z, Y\rangle \right\} \tag{A.3}$$

is the best 1D descriptor for a one-dimensional model. Next, the residual error between a least-squares fit of the first-ranked feature to the property of interest is computed, $\Delta_{1D} = Y - \sum_{i=0}^{1} c_i d_i(\vec{X})$, and the residual error is used to select a new set of *subs_sis*[1] features S_{2D} with the highest correlation to the residual error Δ_{1D}. A 2D descriptor is then constructed by performing a least-squares regression among all possible pairs contained in the union of the sets $\mathcal{S} = S_{\text{2D}} \cup S_{\text{1D}}$ selected in the first and second iteration. The procedure can be repeated to create higher dimensional models by testing progressively larger number of feature-combinations. For an n-dimensional model, SIS selects a subspace

$$S_{\text{nD}} = \left\{ Z | Z \in \bigcup_{k=1}^{rung^1} \Phi_k, \text{sorted in desc. order of } \langle Z, \Delta_{\text{(n-1)D}} \rangle \right\} \tag{A.4}$$

of feature combinations from the residual error of the model $\Delta_{\text{(n-1)D}}$, i.e., those with the largest correlation to the residual error of the previous $(n-1)$-dimensional model. The best n-dimensional descriptor $\vec{d}$ (an n-tuple of feature-combinations) is then determined by constructing the n-dimensional model from the union all previously selected subspaces $\mathcal{S} = S_{\text{nD}} \cup S_{\text{(n-1)D}} \cup \ldots \cup S_{\text{1D}}$ until the residual error of the generated n-dimensional symbolic-regression model reaches a desired level of accuracy, that most likely models the underlying trend in the data.

A.2 Gradient-boosting decision trees (GBDT)

The gradient-boosting decision tree (GBDT) [249–252, 267, 268] is a machine-learning method for generating piecewise-constant models $f(\vec{X})$ by creating an ensemble of tree-like models $h_m(\vec{X})$ from the features $\vec{X} = \{X_1, \ldots, X_d\}$ of a data set (Fig. A.2). An ensemble is a combination of individual models,

$$\hat{f}(\vec{X}) = \sum_{m=1}^{\text{n\_estimators}} \alpha_m h_m(\vec{X}) + \text{const.}\ , \qquad \alpha_m \in \mathbb{R}\ , \tag{A.5}$$

that is expected to perform better than models individually fitted to the data alone. Gradient boosting means that decision trees $h_m(\vec{X})$ are constructed sequentially to minimize the errors $L(Y, \hat{f}) = Y - \hat{f}_{(m-1)}(\vec{X})$ of the previous models $\hat{f}_{(m-1)}$ in a forward stage-wise approach[2],

$$\hat{f}_m(\vec{X}) = \hat{f}_{(m-1)}(\vec{X}) + \gamma \cdot \alpha_m h_m(\vec{X})\ , \qquad 0 < \gamma \leq 1\ , \quad \alpha_m \in \mathbb{R}\ . \tag{A.6}$$

As with each tree-like model (decision tree) the overall prediction error (i.e., the loss function $L(Y, \hat{f}_m)$) is continuously reduced,

$$\alpha_m = \arg\min_{\alpha} \sum_{i=1}^{n} L(y_i, \hat{f}_{(m-1)}(\vec{x}_i) + \alpha h_m(\vec{x}_i))\ , \quad \vec{x}_i \in \vec{X}\ , \tag{A.7}$$

GBDT is known to create highly efficient and accurate models. Decision trees are constructed iteratively (Fig. A.2), consisting of nodes (features) and leaves (predictions). A splitting criterion is used to find the best feature to separate the samples of the data set with respect to the property of interest Y. This

[2] Parallel constructed decision trees are used in bagging methods such as the random-forest algorithm [266]).

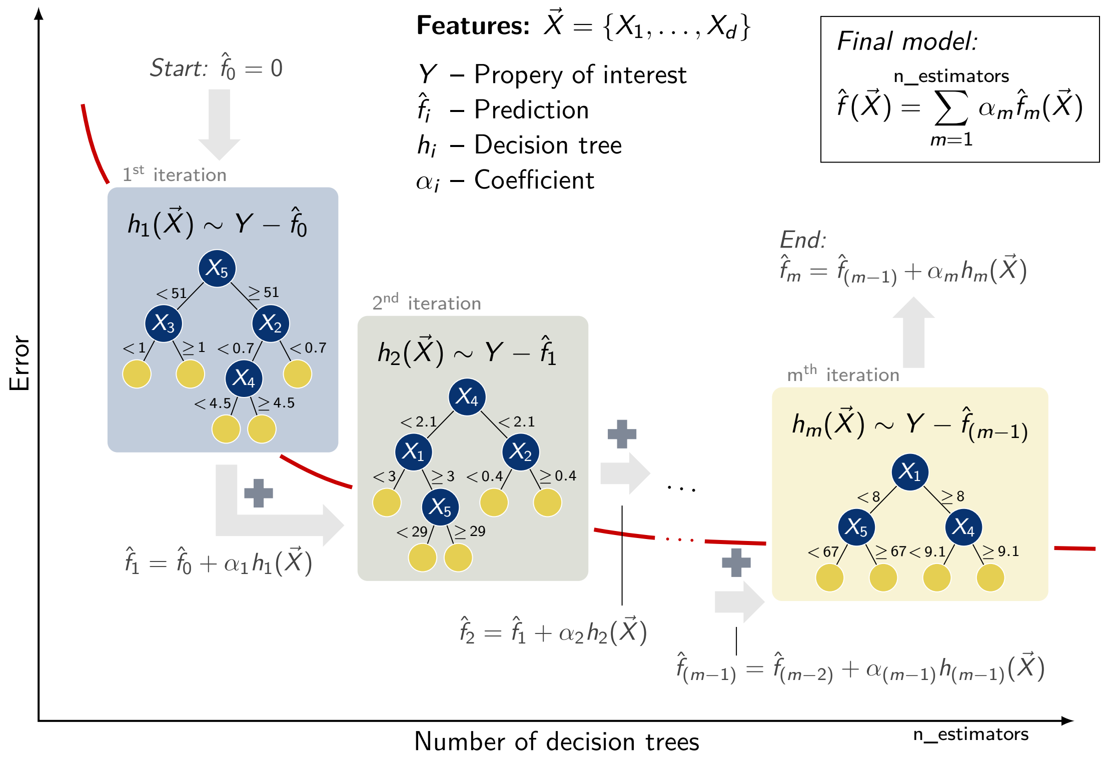

Features: $\vec{X} = \{X_1, \dots, X_d\}$
Y – Propery of interest
$\hat{f}_i$ – Prediction
h_i – Decision tree
α_i – Coefficient
Final model:
$\hat{f}(\vec{X}) = \sum_{m=1}^{\text{n\_estimators}} \alpha_m \hat{f}_m(\vec{X})$
Start: $\hat{f}_0 = 0$
1st iteration
$h_1(\vec{X}) \sim Y - \hat{f}_0$
X_5
< 51
≥ 51
X_3
X_2
< 1
≥ 1
< 0.7
< 0.7
X_4
< 4.5
≥ 4.5
$\hat{f}_1 = \hat{f}_0 + \alpha_1 h_1(\vec{X})$
2nd iteration
$h_2(\vec{X}) \sim Y - \hat{f}_1$
X_4
< 2.1
< 2.1
X_1
X_2
< 3
≥ 3
< 0.4
≥ 0.4
X_5
< 29
≥ 29
$\hat{f}_2 = \hat{f}_1 + \alpha_2 h_2(\vec{X})$
. . .
. . .
End:
$\hat{f}_m = \hat{f}_{(m-1)} + \alpha_m h_m(\vec{X})$
mth iteration
$h_m(\vec{X}) \sim Y - \hat{f}_{(m-1)}$
X_1
< 8
≥ 8
X_5
X_4
< 67
≥ 67
< 9.1
≥ 9.1
$\hat{f}_{(m-1)} = \hat{f}_{(m-2)} + \alpha_{(m-1)} h_{(m-1)}(\vec{X})$
Error
Number of decision trees
n_estimators

Fig. A.2. Gradient boosting continuously reduces the overall prediction error $|Y - \hat{f}|$ by sequentially minimizing the residual errors of decision trees $h_m(\vec{X})$ subject to a property of interest Y. At each step, gradient boosting iteratively constructs the decision trees $h_i(\vec{X})$ as tree-like models with nodes (features) and leaves (predictions) based on a splitting criterion to find the feature $X \in \vec{X} = \{X_1, \dots, X_d\}$ that best separates the samples of the data. This procedure is repeated until the requirements are within the expected error of tolerance.

procedure is repeated until the requirements are within the expected tolerance of error, i.e., when the overall prediction error has converged, there are no more than *min_child_samples*[3] samples left to split, or the specified maximum tree depth has been reached (*max_depth*[3]).

The efficiency and scalability of the piecewise-constant machine-learning model is largely determined by the shrinkage factor γ (*learning_rate*[3], cf., Eq. A.6) and the number of decision trees (*n_estimators*[3]). In general, high shrinkage factors and large number of decision trees increase the accuracy of the overall model, whereas lower shrinkage factors and smaller number of decision trees lead to more robust and generalizable models. Learning rate and number of decision trees can be adjusted by a hyper-parameter optimization and monitored during model creation (*eval_metric*[3]) by stopping the addition of new decision trees as soon as a stopping criterion (e.g., the ℓ_1- or the ℓ_2-norm [52, 249]) has not improved in the last *early_stopping_rounds*[3] iterations.

[3] All names in the parentheses refer to the parameters in the documentation of the LightGBM package (https://lightgbm.readthedocs.io/) [252, 267, 268].

References

[1] Gielen, D.; Boshell, F. *et al.*: "Climate and energy challenges for materials science". Nature Materials **15**, 117 (2016). DOI: 10.1038/nmat4545.

[2] Huggins, R.A.: *Advanced Batteries: Materials Science Aspects*. Springer US, Boston, MA (2009). ISBN: 978-0-387-76424-5. DOI: 10.1007/978-0-387-76424-5_20.

[3] Drozdov, A.P.; Kong, P.P. *et al.*: "Superconductivity at 250 K in lanthanum hydride under high pressures". Nature **569**, 7757, 528–531 (2019). DOI: 10.1038/s41586-019-1201-8.

[4] de Pablo, J.J.; Jackson, N.E. *et al.*: "New frontiers for the materials genome initiative". npj Computational Materials **5**, 1, 41 (2019). DOI: 10.1038/s41524-019-0173-4.

[5] Hey, T.; Tansley, S. *et al.*: *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research (2009). ISBN: 978-0-9825442-0-4. URL: https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/.

[6] Agrawal, A. & Choudhary, A.: "Perspective: Materials informatics and big data: Realization of the 'fourth paradigm' of science in materials science". APL Materials **4**, 5, 053208 (2016). DOI: 10.1063/1.4946894.

[7] Rajan, K.: "Materials informatics". Materials Today **8**, 10, 38–45 (2005). DOI: 10.1016/S1369-7021(05)71123-8.

[8] Rickman, J.; Lookman, T. *et al.*: "Materials informatics: From the atomic-level to the continuum". Acta Materialia **168**, 473–510 (2019). DOI: 10.1016/j.actamat.2019.01.051.

[9] Saal, J.E.; Oliynyk, A.O. *et al.*: "Machine Learning in Materials Discovery: Confirmed Predictions and Their Underlying Approaches". Annual Review of Materials Research **50**, 1, 49–69 (2020). DOI: 10.1146/annurev-matsci-090319-010954.

[10] Draxl, C. & Scheffler, M.: "NOMAD: The FAIR concept for big data-driven materials science". MRS Bulletin **43**, 9, 676–682 (2018). DOI: 10.1557/mrs.2018.208.

[11] Wilkinson, M.D.; Dumontier, M. *et al.*: "The FAIR Guiding Principles for scientific data management and stewardship". Scientific Data **3**, 160018 (2016). DOI: 10.1038/sdata.2016.18. Comment.

[12] Kim, E.; Huang, K. *et al.*: "Virtual screening of inorganic materials synthesis parameters with deep learning". npj Computational Materials **3**, 1, 53 (2017). DOI: 10.1038/s41524-017-0055-6.

[13] Jain, A.; Ong, S.P. *et al.*: "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation". APL Materials **1**, 1, 011002 (2013). DOI: 10.1063/1.4812323.

[14] Jain, A.; Ong, S.P. *et al.*: “FireWorks: a dynamic workflow system designed for high-throughput applications”. Concurrency and Computation: Practice and Experience **27**, 17, 5037–5059 (2015). DOI: 10.1002/cpe.3505.

[15] Jain, A.; Montoya, J. *et al.*: *The Materials Project: Accelerating Materials Design Through Theory-Driven Data and Tools*, 1–34. Springer International Publishing, Cham (2018). ISBN: 978-3-319-42913-7. DOI: 10.1007/978-3-319-42913-7_60-1.

[16] Curtarolo, S.; Setyawan, W. *et al.*: “AFLOW: An automatic framework for high-throughput materials discovery”. Computational Materials Science **58**, 218–226 (2012). DOI: 10.1016/j.commatsci.2012.02.005.

[17] Curtarolo, S.; Setyawan, W. *et al.*: “AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations”. Computational Materials Science **58**, 227 – 235 (2012). DOI: 10.1016/j.commatsci.2012.02.002.

[18] Saal, J.E.; Kirklin, S. *et al.*: “Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD)”. JOM **65**, 11, 1501–1509 (2013). DOI: 10.1007/s11837-013-0755-4.

[19] Kirklin, S.; Saal, J.E. *et al.*: “The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies”. npj Computational Materials **1**, 1, 15010 (2015). DOI: 10.1038/npjcompumats.2015.10.

[20] Larsen, A.H.; Mortensen, J.J. *et al.*: “The atomic simulation environment—a Python library for working with atoms”. Journal of Physics: Condensed Matter **29**, 27, 273002 (2017). DOI: 10.1088/1361-648x/aa680e.

[21] Pizzi, G.; Cepellotti, A. *et al.*: “AiiDA: automated interactive infrastructure and database for computational science”. Computational Materials Science **111**, 218–230 (2016). DOI: 10.1016/j.commatsci.2015.09.013.

[22] Huber, S.P.; Zoupanos, S. *et al.*: “AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance”. Scientific Data **7**, 1, 300 (2020). DOI: 10.1038/s41597-020-00638-4.

[23] Uhrin, M.; Huber, S.P. *et al.*: “Workflows in AiiDA: Engineering a high-throughput, event-based engine for robust and modular computational workflows”. Computational Materials Science **187**, 110086 (2021). DOI: 10.1016/j.commatsci.2020.110086.

[24] Mathew, K.; Montoya, J.H. *et al.*: “Atomate: A high-level interface to generate, execute, and analyze computational materials science workflows”. Computational Materials Science **139**, 140–152 (2017). DOI: 10.1016/j.commatsci.2017.07.030.

[25] Hohenberg, P. & Kohn, W.: “Inhomogeneous Electron Gas”. Physical Review **136**, B864–B871 (1964). DOI: 10.1103/PhysRev.136.B864.

[26] Kohn, W. & Sham, L.J.: “Self-Consistent Equations Including Exchange and Correlation Effects”. Physical Review **140**, A1133–A1138 (1965). DOI: 10.1103/PhysRev.140.A1133.

[27] Becke, A.D.: “Perspective: Fifty years of density-functional theory in chemical physics”. The Journal of Chemical Physics **140**, 18, 18A301 (2014). DOI: 10.1063/1.4869598.

[28] Jones, R.O.: “Density functional theory: Its origins, rise to prominence, and future”. Reviews of Modern Physics **87**, 897–923 (2015). DOI: 10.1103/RevModPhys.87.897.

[29] Lejaeghere, K.; Bihlmayer, G. *et al.*: "Reproducibility in density functional theory calculations of solids". Science **351**, 6280 (2016). DOI: 10.1126/science.aad3000.

[30] COD: "Crystallography Open Database" (2003). URL: https://crystallography.net. (accessed: September 2021).

[31] Springer: "SpringerMaterials" (2010). URL: https://materials.springer.com. (accessed: September 2021).

[32] CMR: "Computational Materials Repository" (2012). URL: https://cmr.fysik.dtu.dk. (accessed: September 2021).

[33] AFlow: "Automatic FLOW for Materials Discovery" (2012). URL: http://aflowlib.org. (accessed: September 2021).

[34] NREL: "Computational materials database for renewable energy applications" (2012). URL: https://materials.nrel.gov. (accessed: September 2021).

[35] MP: "Materials Project" (2013). URL: https://materialsproject.org. (accessed: September 2021).

[36] OQMD: "Open Quantum Materials Database" (2013). URL: https://oqmd.org. (accessed: September 2021).

[37] NOMAD: "The Novel Materials Discovery (NOMAD) Laboratory" (2016). URL: https://www.nomad-coe.eu/. (accessed: September 2021).

[38] Walsh, A.: "The quest for new functionality". Nature Chemistry **7**, 274–275 (2015). DOI: 10.1038/nchem.2213.

[39] Oganov, A.R.; Pickard, C.J. *et al.*: "Structure prediction drives materials discovery". Nature Reviews Materials **4**, 5, 331–348 (2019). DOI: 10.1038/s41578-019-0101-8.

[40] Donoho, D.: "50 Years of Data Science". Journal of Computational and Graphical Statistics **26**, 4, 745–766 (2017). DOI: 10.1080/10618600.2017.1384734.

[41] Schleder, G.R.; Padilha, A.C.M. *et al.*: "From DFT to machine learning: recent approaches to materials science–a review". Journal of Physics: Materials **2**, 3, 032001 (2019). DOI: 10.1088/2515-7639/ab084b.

[42] Vasudevan, R.K.; Choudhary, K. *et al.*: "Materials science in the artificial intelligence age: high-throughput library generation, machine learning, and a pathway from correlations to the underpinning physics". MRS Communications **9**, 3, 821–838 (2019). DOI: 10.1557/mrc.2019.95.

[43] Schütt, K.T.; Chmiela, S. *et al.*: *Machine Learning Meets Quantum Physics*. Springer, Cham (2020). ISBN: 978-3-030-40245-7. DOI: 10.1007/978-3-030-40245-7.

[44] Pedregosa, F.; Varoquaux, G. *et al.*: "Scikit-learn: Machine Learning in Python". Journal of Machine Learning Research **12**, 2825–2830 (2011). URL: http://www.jmlr.org/papers/v12/pedregosa11a.

[45] Chollet, F. & Others: "Keras" (2015). URL: https://keras.io.

[46] Abadi, M.; Agarwal, A. *et al.*: "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems" (2015). URL: https://www.tensorflow.org/.

[47] Zunger, A.: "Inverse design in search of materials with target functionalities". Nature Reviews Chemistry **2**, 4, 0121 (2018). DOI: 10.1038/s41570-018-0121.

[48] Eagar, T.W.: “Bringing New Materials to Market”. Technology Review **98**, 2, 42–49 (1995). URL: http://dl.acm.org/citation.cfm?id=204215.204220.

[49] Nilsson, N.J.: *Artificial Intelligence: A New Synthesis*. Morgan Kaufmann, San Francisco, CA, USA (1998). ISBN: 978-1-55860-467-4. DOI: 10.1016/C2009-0-27773-7.

[50] Poole, D.; Mackworth, A. *et al.*: *Computational Intelligence: A Logical Approach*. Oxford University Press, Ney York, Oxford (1998). ISBN: 0-19-510270-3.

[51] Hastie, T.; Tibshirani, R. *et al.*: *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, 2nd ed. (2009). DOI: 10.1007/978-0-387-84858-7.

[52] James, G.; Witten, D. *et al.*: *An Introduction to Statistical Learning*, vol. 103 of *Springer Texts in Statistics*. Springer, New York (2013). ISBN: 978-1-4614-7137-0. DOI: 10.1007/978-1-4614-7138-7.

[53] Samuel, A.L.: “Some Studies in Machine Learning Using the Game of Checkers”. IBM Journal of Research and Development **3**, 3, 210–229 (1959). DOI: 10.1147/rd.33.0210.

[54] Mitchell, T.M.: *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1st ed. (1997). ISBN: 0070428077, 9780070428072. URL: http://profsite.um.ac.ir/monsefi/machine-learning/pdf/Machine-Learning-Tom-Mitchell.pdf.

[55] Ghiringhelli, L.M.; Vybiral, J. *et al.*: “Big Data of Materials Science: Critical Role of the Descriptor”. Physical Review Letters **114**, 105503 (2015). DOI: 10.1103/PhysRevLett.114.105503.

[56] Ward, L.; Agrawal, A. *et al.*: “A general-purpose machine learning framework for predicting properties of inorganic materials”. npj Computational Materials **2**, 1, 16028 (2016). DOI: 10.1038/npjcompumats.2016.28.

[57] Yamada, H.; Liu, C. *et al.*: “Predicting Materials Properties with Little Data Using Shotgun Transfer Learning”. ACS Central Science **5**, 10, 1717–1730 (2019). DOI: 10.1021/acscentsci.9b00804.

[58] Kim, C.; Chandrasekaran, A. *et al.*: “Active-learning and materials design: the example of high glass transition temperature polymers”. MRS Communications **9**, 3, 860–866 (2019). DOI: 10.1557/mrc.2019.78.

[59] Zhang, Y. & Ling, C.: “A strategy to apply machine learning to small datasets in materials science”. npj Computational Materials **4**, 1, 25 (2018). DOI: 10.1038/s41524-018-0081-z.

[60] Saunders, C.; Gammerman, A. *et al.*: “Transduction with confidence and credibility”. In: T. Dean (ed.) “Proceedings of the 16th International Joint Conference on Artificial Intelligence”, vol. 2 of *IJCAI’99*, 722–726. Morgan Kaufmann (1999). URL: https://www.ijcai.org/Proceedings/1999-2/.

[61] Vovk, V.; Gammerman, A. *et al.*: “Machine-Learning Applications of Algorithmic Randomness”. In: “Proceedings of the Sixteenth International Conference on Machine Learning”, ICML’99, 444–453. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1999). ISBN: 1558606122.

[62] Vovk, V.; Gammerman, A. *et al.*: *Algorithmic Learning in a Random World*. Springer, Boston, MA (2005). ISBN: 978-0-387-25061-8. DOI: 10.1007/b106715.

[63] Shafer, G. & Vovk, V.: “A Tutorial on Conformal Prediction”. The Journal of Machine Learning Research **9**, 371–421 (2008). URL: http://www.jmlr.org/papers/v9/shafer08a.html.

[64] Balasubramanian, V.N.; Ho, S.S. *et al.*: *Conformal Prediction for Reliable Machine Learning*. Morgan Kaufmann, Boston (2014). ISBN: 978-0-12-398537-8. DOI: 10.1016/C2012-0-00234-7.

[65] Jain, A.; Hautier, G. *et al.*: "New opportunities for materials informatics: Resources and data mining techniques for uncovering hidden relationships". Journal of Materials Research **31**, 8, 977–994 (2016). DOI: 10.1557/jmr.2016.80.

[66] Pilania, G.; Gubernatis, J. *et al.*: "Multi-fidelity machine learning models for accurate bandgap predictions of solids". Computational Materials Science **129**, 156 – 163 (2017). DOI: 10.1016/j.commatsci.2016.12.004.

[67] Lipton, Z.C.: "The Mythos of Model Interpretability". arXiv e-prints (2016). ARXIV: 1606.03490.

[68] Murdoch, W.J.; Singh, C. *et al.*: "Interpretable machine learning: definitions, methods, and applications". arXiv e-prints (2019). ARXIV: 1901.04592.

[69] Schrödinger, E.: "An Undulatory Theory of the Mechanics of Atoms and Molecules". Physical Review **28**, 1049–1070 (1926). DOI: 10.1103/PhysRev.28.1049.

[70] Gordon, D.F. & Desjardins, M.: "Evaluation and Selection of Biases in Machine Learning". Machine Learning **20**, 1, 5–22 (1995). DOI: 10.1023/A:1022630017346.

[71] Blum, A.L. & Langley, P.: "Selection of relevant features and examples in machine learning". Artificial Intelligence **97**, 1, 245–271 (1997). DOI: 10.1016/S0004-3702(97)00063-5.

[72] Kohavi, R. & John, G.H.: "Wrappers for feature subset selection". Artificial Intelligence **97**, 1, 273–324 (1997). DOI: 10.1016/S0004-3702(97)00043-X.

[73] Pearson, K.: "On lines and planes of closest fit to systems of points in space". The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **2**, 11, 559–572 (1901). DOI: 10.1080/14786440109462720.

[74] Hotelling, H.: "Analysis of a complex of statistical variables into principal components". Journal of Educational Psychology **24**, 6, 417–441 (1933). DOI: 10.1037/h0071325.

[75] Hotelling, H.: "Relations Between Two Sets of Variates". Biometrika **28**, 3/4, 321–377 (1936). URL: http://www.jstor.org/stable/2333955.

[76] Borg, I. & Groenen, P.J.F.: *Modern Multidimensional Scaling: Theory and Applications*. Springer Series in Statistics. Springer, New York, NY, 2nd ed. (2005). ISBN: 978-0-387-28981-6. DOI: 10.1007/0-387-28981-X.

[77] Zhang, Z.y. & Zha, H.y.: "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment". Journal of Shanghai University (English Edition) **8**, 4, 406–424 (2004). DOI: 10.1007/s11741-004-0051-1.

[78] Tenenbaum, J.B.; Silva, V.d. *et al.*: "A Global Geometric Framework for Nonlinear Dimensionality Reduction". Science **290**, 5500, 2319–2323 (2000). DOI: 10.1126/science.290.5500.2319.

[79] Donoho, D.L. & Grimes, C.: "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data". Proceedings of the National Academy of Sciences **100**, 10, 5591–5596 (2003). DOI: 10.1073/pnas.1031596100.

[80] Van der Maaten, L. & Hinton, G.: "Visualizing High-Dimensional Data Using t-SNE". Journal of Machine Learning Research **9**, 2579–2605 (2008). URL: http://www.jmlr.org/papers/v9/vandermaaten08a.html.

[81] Wattenberg, M.; Viégas, F. *et al.*: "How to Use t-SNE Effectively". Distill (2016). DOI: 10.23915/distill.00002.

[82] McInnes, L.; Healy, J. *et al.*: "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". arXiv e-prints (2018). ARXIV: 1802.03426. Source code: https://github.com/lmcinnes/umap.

[83] Shannon, C.E. & Weaver, W.: *The Mathematical Theory of Communication*, vol. III. Illinois Press (1949).

[84] Cover, T.M. & Thomas, J.A.: *Elements of Information Theory, Second Edition*. John Wiley & Sons, Ltd, 2nd ed. (2005). ISBN: 9780471241959. DOI: 10.1002/047174882X.

[85] Kullback, S. & Leibler, R.A.: "On Information and Sufficiency". The Annals of Mathematical Statistics **22**, 1, 79–86 (1951). URL: http://www.jstor.org/stable/2236703.

[86] Wolpert, D.H.: "The Lack of A Priori Distinctions Between Learning Algorithms". Neural Computation **8**, 7, 1341–1390 (1996). DOI: 10.1162/neco.1996.8.7.1341.

[87] Wolpert, D.H.: "The Existence of A Priori Distinctions Between Learning Algorithms". Neural Computation **8**, 7, 1391–1420 (1996). DOI: 10.1162/neco.1996.8.7.1391.

[88] Wolpert, D.H. & Macready, W.G.: "No free lunch theorems for search". Technical Report SFI-TR-95-02-010 10, Santa Fe Institute (1995). URL: http://jmvidal.cse.sc.edu/library/wolpert95a.pdf.

[89] Wolpert, D.H. & Macready, W.G.: "No free lunch theorems for optimization". IEEE Transactions on Evolutionary Computation **1**, 1, 67–82 (1997). DOI: 10.1109/4235.585893.

[90] Regler, B.; Scheffler, M. *et al.*: "TCMI: a non-parametric mutual-dependence estimator for multivariate continuous distributions". arXiv e-prints (2020). ARXIV: arXiv:2001.11212.

[91] Land, A.H. & Doig, A.G.: "An Automatic Method of Solving Discrete Programming Problems". Econometrica **28**, 3, 497–520 (1960). DOI: 10.2307%2F1910129.

[92] Narendra & Fukunaga: "A Branch and Bound Algorithm for Feature Subset Selection". IEEE Transactions on Computers **C-26**, 9, 917–922 (1977). DOI: 10.1109/TC.1977.1674939.

[93] Yu, B. & Yuan, B.: "A more efficient branch and bound algorithm for feature selection". Pattern Recognition **26**, 6, 883–889 (1993). DOI: 10.1016/0031-3203(93)90054-Z.

[94] Clausen, J.: "Branch and Bound Algorithms – Principles And Examples". Tech. rep., Department of Computer Science, University of Copenhagen, Universitetsparken 1, DK2100 Copenhagen, Denmark (1999). URL: https://pdfs.semanticscholar.org/fffd/ebefecd6a60a841acc8aafb7f0e89a76f996.pdf.

[95] Pudil, P.; Novovičová, J. *et al.*: *Recent Feature Selection Methods in Statistical Pattern Recognition*, 565–615. Springer US, Boston, MA (2002). ISBN: 978-1-4613-0231-5. DOI: 10.1007/978-1-4613-0231-5_23.

[96] Morrison, D.R.; Jacobson, S.H. *et al.*: "Branch-and-bound algorithms: A survey of recent advances in searching, branching, and pruning". Discrete Optimization **19**, 79–102 (2016). DOI: 10.1016/j.disopt.2016.01.005.

[97] Zheng, Y. & Kwoh, C.K.: "A Feature Subset Selection Method Based On High-Dimensional Mutual Information". Entropy **13**, 4, 860–901 (2011). DOI: 10.3390/e13040860.

[98] Mandros, P.; Boley, M. *et al.*: "Discovering Reliable Approximate Functional Dependencies". In: "Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining", KDD '17, 355–363. ACM, New York, NY, USA (2017). ISBN: 978-1-4503-4887-4. DOI: 10.1145/3097983.3098062.

[99] Dormann, C.F.; Elith, J. *et al.*: "Collinearity: a review of methods to deal with it and a simulation study evaluating their performance". Ecography **36**, 1, 27–46 (2013). DOI: 10.1111/j.1600-0587.2012.07348.x.

[100] Parr, T. & Turgutlu, K.: "Feature importances for scikit-learn machine learning models" (2018). URL: https://github.com/parrt/random-forest-importances. Also read "Beware Default Random Forest Importances" (https://explained.ai/rf-importance/index.html) for a deeper discussion of the issues surrounding feature importances in random forests.

[101] Pearson, K.: "Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia." Philosophical Transactions of the Royal Society of London Series A **187**, 253–318 (1896). DOI: 10.1098/rsta.1896.0007.

[102] Settles, B.: "Active Learning Literature Survey". Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009). URL: http://burrsettles.com/pub/settles.activelearning.pdf.

[103] Lookman, T.; Balachandran, P.V. *et al.*: "Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design". npj Computational Materials **5**, 1, 21 (2019). DOI: 10.1038/s41524-019-0153-8.

[104] Kim, Y.; Kim, E. *et al.*: "Machine-learned metrics for predicting the likelihood of success in materials discovery". arXiv e-prints (2019). ARXIV: 1911.11201.

[105] Schmidhuber, J.: "Deep learning in neural networks: An overview". Neural Networks **61**, 85–117 (2015). DOI: 10.1016/j.neunet.2014.09.003.

[106] LeCun, Y.; Bengio, Y. *et al.*: "Deep learning". Nature **521**, 7553, 436–444 (2015). DOI: 10.1038/nature14539.

[107] Gal, Y.: "Uncertainty in Deep Learning" (2016). URL: https://mlg.eng.cam.ac.uk/yarin/thesis/thesis.pdf. Blog post: http://mlg.eng.cam.ac.uk/yarin/blog_2248.html.

[108] Rasmussen, C.E.: *Gaussian Processes in Machine Learning*, 63–71. Springer Berlin Heidelberg, Berlin, Heidelberg (2004). ISBN: 978-3-540-28650-9. DOI: 10.1007/978-3-540-28650-9_4.

[109] Williams, C.K.I. & Rasmussen, C.E.: "Gaussian Processes for Regression". In: "Proceedings of the 8th International Conference on Neural Information Processing Systems", NIPS'95, 514–520. MIT Press, Cambridge, MA, USA (1995). URL: http://dl.acm.org/citation.cfm?id=2998828.2998901.

[110] Kohn, W.: "Nobel Lecture: Electronic structure of matter—wave functions and density functionals". Rev. Mod. Phys. **71**, 1253–1266 (1999). DOI: 10.1103/RevModPhys.71.1253.

[111] Jain, A.; Shin, Y. *et al.*: "Computational predictions of energy materials using density functional theory". Nature Reviews Materials **1**, 15004 (2016). URL: 10.1038/natrevmats.2015.4. Review Article.

[112] Ramprasad, R.; Batra, R. *et al.*: “Machine learning in materials informatics: recent applications and prospects”. npj Computational Materials **3**, 1, 54 (2017). DOI: 10.1038/s41524-017-0056-5.

[113] Pickard, C.J. & Needs, R.J.: “Ab initiorandom structure searching”. Journal of Physics: Condensed Matter **23**, 5, 053201 (2011). DOI: 10.1088/0953-8984/23/5/053201.

[114] Wang, Y.; Lv, J. *et al.*: “Crystal structure prediction via particle-swarm optimization”. Physical Review B **82**, 094116 (2010). DOI: 10.1103/PhysRevB.82.094116.

[115] Wang, Y.; Lv, J. *et al.*: “CALYPSO: A method for crystal structure prediction”. Computer Physics Communications **183**, 10, 2063–2070 (2012). DOI: 10.1016/j.cpc.2012.05.008.

[116] Li, Z. & Scheraga, H.A.: “Monte Carlo-minimization approach to the multiple-minima problem in protein folding”. Proceedings of the National Academy of Sciences **84**, 19, 6611–6615 (1987). DOI: 10.1073/pnas.84.19.6611.

[117] Wales, D.J. & Doye, J.P.K.: “Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms”. The Journal of Physical Chemistry A **101**, 28, 5111–5116 (1997). DOI: 10.1021/jp970984n.

[118] Kirkpatrick, S.; Gelatt, C.D. *et al.*: “Optimization by Simulated Annealing”. Science **220**, 4598, 671–680 (1983). DOI: 10.1126/science.220.4598.671.

[119] Černý, V.: “Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm”. Journal of Optimization Theory and Applications **45**, 1, 41–51 (1985). DOI: 10.1007/BF00940812.

[120] Ceperley, D. & Alder, B.: “Quantum Monte Carlo”. Science **231**, 4738, 555–560 (1986). DOI: 10.1126/science.231.4738.555.

[121] van Laarhoven, P.J.M. & Aarts, E.H.L.: *Simulated annealing*. Springer Netherlands, Dordrecht (1987). ISBN: 978-94-015-7744-1. DOI: 10.1007/978-94-015-7744-1.

[122] Goedecker, S.: “Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems”. The Journal of Chemical Physics **120**, 21, 9911–9917 (2004). DOI: 10.1063/1.1724816.

[123] Amsler, M. & Goedecker, S.: “Crystal structure prediction using the minima hopping method”. The Journal of Chemical Physics **133**, 22, 224104 (2010). DOI: 10.1063/1.3512900.

[124] Stillinger, F.H.: “Exponential multiplicity of inherent structures”. Physical Review E **59**, 48–51 (1999). DOI: 10.1103/PhysRevE.59.48.

[125] Franceschetti, A. & Zunger, A.: “The inverse band-structure problem of finding an atomic configuration with given electronic properties”. Nature **402**, 6757, 60–63 (1999). DOI: 10.1038/46995.

[126] Jain, A.; Bollinger, J.A. *et al.*: “Inverse methods for material design”. AIChE Journal **60**, 8, 2732–2740 (2014). DOI: 10.1002/aic.14491.

[127] Born, M. & Oppenheimer, R.: “Zur Quantentheorie der Molekeln”. Annalen der Physik **389**, 20, 457–484 (1927). DOI: 10.1002/andp.19273892002.

[128] Hartree, D.R.: “The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods”. Mathematical Proceedings of the Cambridge Philosophical Society **24**, 1, 89–110 (1928). DOI: 10.1017/S0305004100011919.

[129] Hartree, D.R.: "The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part II. Some Results and Discussion". Mathematical Proceedings of the Cambridge Philosophical Society **24**, 1, 111–132 (1928). DOI: 10.1017/S0305004100011920.

[130] Fock, V.: "Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems". Zeitschrift für Physik **61**, 1, 126–148 (1930). DOI: 10.1007/BF01340294.

[131] Slater, J.C.: "Note on Hartree's Method". Physical Review **35**, 210–211 (1930). DOI: 10.1103/PhysRev.35.210.2.

[132] Slater, J.C.: "A Simplification of the Hartree-Fock Method". Physical Review **81**, 385–390 (1951). DOI: 10.1103/PhysRev.81.385.

[133] Møller, C. & Plesset, M.S.: "Note on an Approximation Treatment for Many-Electron Systems". Physical Review **46**, 618–622 (1934). DOI: 10.1103/PhysRev.46.618.

[134] Bartlett, R.J.: "Coupled-cluster approach to molecular structure and spectra: a step toward predictive quantum chemistry". The Journal of Physical Chemistry **93**, 5, 1697–1708 (1989). DOI: 10.1021/j100342a008.

[135] Mattsson, A.E.; Schultz, P.A. *et al.*: "Designing meaningful density functional theory calculations in materials science—a primer". Modelling and Simulation in Materials Science and Engineering **13**, 1, R1–R31 (2004). DOI: 10.1088/0965-0393/13/1/r01.

[136] Perdew, J.P. & Schmidt, K.: "Jacob's ladder of density functional approximations for the exchange-correlation energy". AIP Conference Proceedings **577**, 1, 1–20 (2001). DOI: 10.1063/1.1390175.

[137] Pyzer-Knapp, E.O.; Suh, C. *et al.*: "What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery". Annual Review of Materials Research **45**, 1, 195–216 (2015). DOI: 10.1146/annurev-matsci-070214-020823.

[138] Castelli, I.E.; Landis, D.D. *et al.*: "New cubic perovskites for one- and two-photon water splitting using the computational materials repository". Energy Environmental Science **5**, 9034–9043 (2012). DOI: 10.1039/C2EE22341D.

[139] Landis, D.D.; Hummelshøj, J.S. *et al.*: "The Computational Materials Repository". Computing in Science Engineering **14**, 6, 51–57 (2012). DOI: 10.1109/MCSE.2012.16.

[140] NOMAD: "NOMAD Repository" (2020). URL: https://encyclopedia.nomad-coe.eu/gui/. (accessed: September 2021).

[141] Stevanović, V.; Lany, S. *et al.*: "Correcting density functional theory for accurate predictions of compound enthalpies of formation: Fitted elemental-phase reference energies". Physical Review B **85**, 115104 (2012). DOI: 10.1103/PhysRevB.85.115104.

[142] Lany, S.: "Band-structure calculations for the 3d transition metal oxides in GW". Physical Review B **87**, 085112 (2013). DOI: 10.1103/PhysRevB.87.085112.

[143] Lany, S.: "Semiconducting transition metal oxides". Journal of Physics: Condensed Matter **27**, 28, 283203 (2015). DOI: 10.1088/0953-8984/27/28/283203.

[144] O'Mara, J.; Meredig, B. *et al.*: "Materials Data Infrastructure: A Case Study of the Citrination Platform to Examine Data Import, Storage, and Access". JOM **68**, 8, 2031–2034 (2016). DOI: 10.1007/s11837-016-1984-0.

[145] Informatics, C.: “Citrination” (2016). URL: https://citrination.com. (accessed: September 2021).

[146] Gražulis, S.; Chateigner, D. *et al.*: “Crystallography Open Database – an open-access collection of crystal structures”. Journal of Applied Crystallography **42**, 4, 726–729 (2009). DOI: 10.1107/S0021889809016690.

[147] Gražulis, S.; Daškevič, A. *et al.*: “Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration”. Nucleic Acids Research **40**, D1, D420–D427 (2011). DOI: 10.1093/nar/gkr900.

[148] ICSD: “Inorganic Crystal Structure Database” (1978). URL: https://icsd.fiz-karlsruhe.de. (accessed: September 2021).

[149] Hellenbrandt, M.: “The Inorganic Crystal Structure Database (ICSD) - Present and Future”. Crystallography Reviews **10**, 1, 17–22 (2004). DOI: 10.1080/08893110410001664882.

[150] Himanen, L.; Geurts, A. *et al.*: “Data-Driven Materials Science: Status, Challenges, and Perspectives”. Advanced Science **6**, 21, 1900808 (2019). DOI: 10.1002/advs.201900808. Erratum: 10.1002/advs.201903667.

[151] Ong, S.P.; Richards, W.D. *et al.*: “Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis”. Computational Materials Science **68**, 314–319 (2013). DOI: 10.1016/j.commatsci.2012.10.028.

[152] Baldi, P. & Brunak, S.: *Bioinformatics: The Machine Learning Approach*. MIT Press, Cambridge, London, 2nd ed. (2001). ISBN: 9780262025065. URL: https://mitpress.mit.edu/books/bioinformatics-second-edition.

[153] Butler, K.T.; Davies, D.W. *et al.*: “Machine learning for molecular and materials science”. Nature **559**, 7715, 547–555 (2018). DOI: 10.1038/s41586-018-0337-2.

[154] LeSar, R.: “Materials informatics: An emerging technology for materials development”. Statistical Analysis and Data Mining: The ASA Data Science Journal **1**, 6, 372–374 (2009). DOI: 10.1002/sam.10034.

[155] Pilania, G.; Wang, C. *et al.*: “Accelerating materials property predictions using machine learning”. Scientific Reports **3**, 1, 2810 (2013). DOI: 10.1038/srep02810.

[156] Bartok, A.P.; Gillan, M.J. *et al.*: “Machine learning for predictive condensed-phase Simulation”. arXiv e-prints (2013). ARXIV: 1302.5680.

[157] Behler, J.: “Perspective: Machine learning potentials for atomistic simulations”. The Journal of Chemical Physics **145**, 17, 170901 (2016). DOI: 10.1063/1.4966192.

[158] Podryabinkin, E.V. & Shapeev, A.V.: “Active learning of linearly parametrized interatomic potentials”. Computational Materials Science **140**, 171–180 (2017). DOI: 10.1016/j.commatsci.2017.08.031.

[159] Bisbo, M.K. & Hammer, B.: “Efficient global structure optimization with a machine learned surrogate model”. arXiv e-prints (2019). ARXIV: 1907.05741.

[160] Schmidt, J.; Marques, M.R.G. *et al.*: “Recent advances and applications of machine learning in solid-state materials science”. npj Computational Materials **5**, 1, 83 (2019). DOI: 10.1038/s41524-019-0221-0.

[161] Ghiringhelli, L.M.; Vybiral, J. *et al.*: "Learning physical descriptors for materials science by compressed sensing". New Journal of Physics **19**, 2, 023017 (2017). DOI: 10.1088/1367-2630/aa57bf.

[162] Meredig, B.; Agrawal, A. *et al.*: "Combinatorial screening for new materials in unconstrained composition space with machine learning". Physical Review B **89**, 094104 (2014). DOI: 10.1103/PhysRevB.89.094104.

[163] Lee, J.; Seko, A. *et al.*: "Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques". Physical Review B **93**, 115104 (2016). DOI: 10.1103/PhysRevB.93.115104.

[164] Sutton, C.; Ghiringhelli, L.M. *et al.*: "Crowd-sourcing materials-science challenges with the NOMAD 2018 Kaggle competition". npj Computational Materials **5**, 1, 111 (2019). DOI: 10.1038/s41524-019-0239-3.

[165] de Jong, M.; Chen, W. *et al.*: "A Statistical Learning Framework for Materials Science: Application to Elastic Moduli of k-nary Inorganic Polycrystalline Compounds". Scientific Reports **6**, 1, 34256 (2016). DOI: 10.1038/srep34256.

[166] Isayev, O.; Oses, C. *et al.*: "Universal fragment descriptors for predicting properties of inorganic crystals". Nature Communications **8**, 15679 (2017). URL: 10.1038/ncomms15679.

[167] Stanev, V.; Oses, C. *et al.*: "Machine learning modeling of superconducting critical temperature". npj Computational Materials **4**, 1, 29 (2018). DOI: 10.1038/s41524-018-0085-8.

[168] Hamidieh, K.: "A data-driven statistical model for predicting the critical temperature of a superconductor". Computational Materials Science **154**, 346–354 (2018). DOI: 10.1016/j.commatsci.2018.07.052.

[169] Gu, G.H.; Noh, J. *et al.*: "Machine learning for renewable energy materials". J. Mater. Chem. A **7**, 17096–17117 (2019). DOI: 10.1039/C9TA02356A.

[170] Oses, C.; Gossett, E. *et al.*: "AFLOW-CHULL: Cloud-Oriented Platform for Autonomous Phase Stability Analysis". Journal of Chemical Information and Modeling **58**, 12, 2477–2490 (2018). DOI: 10.1021/acs.jcim.8b00393.

[171] Li, W.; Jacobs, R. *et al.*: "Predicting the thermodynamic stability of perovskite oxides using machine learning models". Computational Materials Science **150**, 454–463 (2018). DOI: 10.1016/j.commatsci.2018.04.033. Source code: https://github.com/uw-cmg/perovskite-oxide-stability-prediction.

[172] Iwasaki, Y.; Sawada, R. *et al.*: "Identification of advanced spin-driven thermoelectric materials via interpretable machine learning". npj Computational Materials **5**, 1, 103 (2019). DOI: 10.1038/s41524-019-0241-9.

[173] Newell, A. & Simon, H.: "The logic theory machine – A complex information processing system". IRE Transactions on Information Theory **2**, 3, 61–79 (1956). DOI: 10.1109/TIT.1956.1056797.

[174] Stefferud, E.: *The Logic Theory Machine: A Model Heuristic Program*. RM-3731-CC. RAND Corporation, Santa Monica, USA (1963). URL: https://www.rand.org/pubs/research_memoranda/RM3731.html.

[175] Isayev, O.; Fourches, D. *et al.*: “Materials Cartography: Representing and Mining Materials Space Using Structural and Electronic Fingerprints”. Chemistry of Materials **27**, 3, 735–743 (2015). DOI: 10.1021/cm503507h.

[176] Ouyang, R.; Ahmetcik, E. *et al.*: “Simultaneous learning of several materials properties from incomplete databases with multi-task SISSO”. Journal of Physics: Materials **2**, 2, 024002 (2019). DOI: 10.1088/2515-7639/ab077b.

[177] Seko, A.; Togo, A. *et al.*: *Descriptors for Machine Learning of Materials Data*, chap. 1, 3–23. Springer Singapore, Singapore (2018). ISBN: 978-981-10-7617-6. DOI: 10.1007/978-981-10-7617-6_1.

[178] Kim, C.; Pilania, G. *et al.*: “From Organized High-Throughput Data to Phenomenological Theory using Machine Learning: The Example of Dielectric Breakdown”. Chemistry of Materials **28**, 5, 1304–1311 (2016). DOI: 10.1021/acs.chemmater.5b04109.

[179] Kim, C.; Pilania, G. *et al.*: “Machine Learning Assisted Predictions of Intrinsic Dielectric Breakdown Strength of ABX3 Perovskites”. The Journal of Physical Chemistry C **120**, 27, 14575–14580 (2016). DOI: 10.1021/acs.jpcc.6b05068.

[180] Toyao, T.; Suzuki, K. *et al.*: “Toward Effective Utilization of Methane: Machine Learning Prediction of Adsorption Energies on Metal Alloys”. The Journal of Physical Chemistry C **122**, 15, 8315–8326 (2018). DOI: 10.1021/acs.jpcc.7b12670.

[181] De Breuck, P.P.; Hautier, G. *et al.*: “Materials property prediction for limited datasets enabled by feature selection and joint learning with MODNet”. npj Computational Materials **7**, 1, 83 (2021). DOI: 10.1038/s41524-021-00552-2.

[182] Rupp, M.; Tkatchenko, A. *et al.*: “Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning”. Physical Review Lett. **108**, 058301 (2012). DOI: 10.1103/PhysRevLett.108.058301.

[183] Bartók, A.P.; Kondor, R. *et al.*: “On representing chemical environments”. Physical Review B **87**, 184115 (2013). DOI: 10.1103/PhysRevB.87.184115. Erratum: 10.1103/PhysRevB.87.184115.

[184] Schütt, K.T.; Glawe, H. *et al.*: “How to represent crystal structures for machine learning: Towards fast prediction of electronic properties”. Physical Review B **89**, 205118 (2014). DOI: 10.1103/PhysRevB.89.205118.

[185] Huo, H. & Rupp, M.: “Unified Representation of Molecules and Crystals for Machine Learning”. arXiv e-prints (2017). ARXIV: 1704.06439.

[186] Chen, C.; Ye, W. *et al.*: “Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals”. Chemistry of Materials **31**, 9, 3564–3572 (2019). DOI: 10.1021/acs.chemmater.9b01294.

[187] De Breuck, P.P. & Hautier, Geoffroy andRignanese, G.M.: “Machine learning materials properties for small datasets”. arXiv e-prints (2020). ARXIV: 2004.14766. Source code: https://github.com/ppdebreuck/modnet.

[188] Ward, L. & Wolverton, C.: “Magpie: A Materials-Agnostic Platform for Informatics and Exploration” (2015). URL: https://bitbucket.org/wolverton/magpie/.

[189] Ward, L.; Dunn, A. *et al.*: “Matminer: An open source toolkit for materials data mining”. Computational Materials Science **152**, 60–69 (2018). DOI: 10.1016/j.commatsci.2018.05.018.

[190] Schmidt, M. & Lipson, H.: “Distilling Free-Form Natural Laws from Experimental Data”. Science **324**, 5923, 81–85 (2009). DOI: 10.1126/science.1165893.

[191] Donoho, D.L.: “Compressed sensing”. IEEE Transactions on Information Theory **52**, 4, 1289–1306 (2006). DOI: 10.1109/TIT.2006.871582.

[192] Lookman, T.; Alexander, F.J. *et al.*: *Information Science for Materials Discovery and Design*. Springer, Cham, 1st ed. (2015). ISBN: 3319238701. DOI: 10.1007/978-3-319-23871-5.

[193] Cristianini, N. & Shawe-Taylor, J.: *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press (2000). DOI: 10.1017/CBO9780511801389.

[194] Saunders, C.; Gammerman, A. *et al.*: “Ridge Regression Learning Algorithm in Dual Variables”. In: “Proceedings of the Fifteenth International Conference on Machine Learning”, ICML ’98, 515–521. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1998). ISBN: 1-55860-556-8. URL: http://dl.acm.org/citation.cfm?id=645527.657464.

[195] Vapnik, V.N.: *Statistical learning theory*. Wiley, New York (1998). URL: https://www.wiley.com/en-us/Statistical+Learning+Theory-p-9780471030034.

[196] Vapnik, V.N.: *The Nature of Statistical Learning Theory*. Information Science and Statistics. Springer, New York (2000). ISBN: 978-1-4757-3264-1. DOI: 10.1007/978-1-4757-3264-1.

[197] Ouyang, R.; Curtarolo, S. *et al.*: “SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates”. Physical Review Materials **2**, 8, 083802 (11) (2018). DOI: 10.1103/PhysRevMaterials.2.083802. Source code: https://github.com/rouyang2017/SISSO.

[198] Kotsiantis, S.B.: “Decision trees: a recent overview”. Artificial Intelligence Review **39**, 4, 261–283 (2013). DOI: 10.1007/s10462-011-9272-4.

[199] Rokach, L. & Maimon, O.: *Data Mining With Decision Trees: Theory and Applications*. World Scientific, River Edge, NJ, USA (2007). ISBN: 9789814590075, 981459007X. DOI: 10.1142/6604.

[200] Scalia, G.; Grambow, C.A. *et al.*: “Evaluating Scalable Uncertainty Estimation Methods for Deep Learning-Based Molecular Property Prediction”. Journal of Chemical Information and Modeling **60**, 6, 2697–2717 (2020). DOI: 10.1021/acs.jcim.9b00975.

[201] Kohavi, R.: “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection”. In: “Proceedings of the 14th International Joint Conference on Artificial Intelligence”, vol. 2 of *IJCAI’95*, 1137–1143. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1995). ISBN: 1558603638. URL: https://dl.acm.org/doi/abs/10.5555/1643031.1643047.

[202] Evans, G.H. & Stanley, W.J.: *The Working Man’s Advocate*, vol. 4 Oct,1/1. New York (1834).

[203] Efron, B.: “Bootstrap Methods: Another Look at the Jackknife”. The Annals of Statistics **7**, 1, 1–26 (1979). URL: http://www.jstor.org/stable/2958830.

[204] Efron, B. & Tibshirani, R.: *An Introduction to the Bootstrap*. Chapman and Hall/CRC, Boca Raton, London, New York (1993). ISBN: 9780429246593. URL: https://www.taylorfrancis.com/books/9780429246593.

[205] Efron, B. & Tibshirani, R.: "Improvements on Cross-Validation: The .632+ Bootstrap Method". Journal of the American Statistical Association **92**, 438, 548–560 (1997). URL: http://www.jstor.org/stable/2965703.

[206] Krstajic, D.; Buturovic, L.J. *et al.*: "Cross-validation pitfalls when selecting and assessing regression and classification models". Journal of Cheminformatics **6**, 1, 10 (2014). DOI: 10.1186/1758-2946-6-10.

[207] Raccuglia, P.; Elbert, K.C. *et al.*: "Machine-learning-assisted materials discovery using failed experiments". Nature **533**, 7601, 73–76 (2016). DOI: 10.1038/nature17439.

[208] Sutton, C.; Boley, M. *et al.*: "Identifying domains of applicability of machine learning models for materials science". Nature Communications **11**, 1, 4428 (2020). DOI: 10.1038/s41467-020-17112-9. Preprint: ChemRxiv:9778670.

[209] Brochu, E.; Cora, V.M. *et al.*: "A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning". arXiv e-prints (2010). ARXIV: 1012.2599.

[210] Todorovic, M.; Gutmann, M.U. *et al.*: "Bayesian inference of atomistic structure in functional materials". npj Computational Materials **5**, 1, 35 (2019). DOI: 10.1038/s41524-019-0175-2.

[211] Bellman, R.E.: *Adaptive Control Processes: A Guided Tour*. Princeton University Press (1961). ISBN: 9780691079011. URL: http://www.jstor.org/stable/j.ctt183ph6v.

[212] Toher, C.; Oses, C. *et al.*: *The AFLOW Fleet for Materials Discovery*, 1–28. Springer International Publishing, Cham (2018). ISBN: 978-3-319-42913-7. DOI: 10.1007/978-3-319-42913-7_63-1.

[213] Hacking Materials Research Group: "Automatminer". Lawrence Berkeley National Laboratory (2020). URL: https://github.com/hackingmaterials/automatminer.

[214] Allahyari, Z. & Oganov, A.R.: "Coevolutionary search for optimal materials in the space of all possible compounds". arXiv e-prints (2018). ARXIV: 1807.00854.

[215] Tom, R.; Rose, T. *et al.*: "Genarris 2.0: A Random Structure Generator for Molecular Crystals". arXiv e-prints (2019). ARXIV: 1909.10629.

[216] Breiman, L.; Friedman, J. *et al.*: *Classification and regression trees*. Chapman and Hall/CRC, Boca Raton, FL (1984).

[217] Fischer, C.C.; Tibbetts, K.J. *et al.*: "Predicting crystal structure by merging data mining with quantum mechanics". Nature Materials **5**, 8, 641–646 (2006). DOI: 10.1038/nmat1691.

[218] Rajan, K.: "Combinatorial Materials Sciences: Experimental Strategies for Accelerated Knowledge Discovery". Annual Review of Materials Research **38**, 1, 299–322 (2008). DOI: 10.1146/annurev.matsci.38.060407.130217.

[219] Nelson, L.J.; Hart, G.L.W. *et al.*: "Compressive sensing as a paradigm for building physics models". Physical Review B **87**, 035125 (2013). DOI: 10.1103/PhysRevB.87.035125.

[220] Mueller, T.; Johlin, E. *et al.*: "Origins of hole traps in hydrogenated nanocrystalline and amorphous silicon revealed through machine learning". Physical Review B **89**, 115202 (2014). DOI: 10.1103/PhysRevB.89.115202.

[221] Candes, E.J.; Romberg, J. *et al.*: "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information". IEEE Transactions on Information Theory **52**, 2, 489–509 (2006). DOI: 10.1109/TIT.2005.862083.

[222] Candes, E.J. & Wakin, M.B.: "An Introduction To Compressive Sampling". IEEE Signal Processing Magazine **25**, 2, 21–30 (2008). DOI: 10.1109/MSP.2007.914731.

[223] Davenport, M.A.; Duarte, M.F. *et al.*: *Introduction to compressed sensing*. Cambridge University Press, Cambridge (2012). ISBN: 9781107005587. URL: https://www.cambridge.org/de/academic/subjects/engineering/communications-and-signal-processing/compressed-sensing-theory-and-applications?format=HB&isbn=9781107005587.

[224] Boche, H.; Calderbank, R. *et al.*: *Compressed Sensing and its Applications: MATHEON Workshop 2013*. Applied and Numerical Harmonic Analysis. Springer, Cham (2015). ISBN: 978-3-319-16042-9. DOI: 10.1007/978-3-319-16042-9.

[225] Tibshirani, R.: "Regression Shrinkage and Selection via the Lasso". Journal of the Royal Statistical Society. Series B (Methodological) **58**, 1, 267–288 (1996). URL: http://www.jstor.org/stable/2346178.

[226] Liu, H. & Motoda, H.: *Feature Extraction, Construction and Selection*. The Springer International Series in Engineering and Computer Science. Springer, Boston, MA (1998). ISBN: 978-1-4615-5725-8. DOI: 10.1007/978-1-4615-5725-8.

[227] Lee, J.A. & Verleysen, M.: *Nonlinear Dimensionality Reduction*. Information Science and Statistics. Springer, New York, NY (2007). ISBN: 978-0-387-39351-3. DOI: 10.1007/978-0-387-39351-3.

[228] Kira, K. & Rendell, L.A.: "A Practical Approach to Feature Selection". In: D. Sleeman & P. Edwards (eds.) "Machine Learning Proceedings 1992", 249–256. Morgan Kaufmann, San Francisco (CA) (1992). ISBN: 978-1-55860-247-2. DOI: 10.1016/B978-1-55860-247-2.50037-1.

[229] Pudil, P.; Novovičová, J. *et al.*: "Floating search methods in feature selection". Pattern Recognition Letters **15**, 11, 1119–1125 (1994). DOI: 10.1016/0167-8655(94)90127-9.

[230] Peng, H.; Long, F. *et al.*: "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy". IEEE Transactions on Pattern Analysis and Machine Intelligence **27**, 8, 1226–1238 (2005). DOI: 10.1109/TPAMI.2005.159.

[231] Basseville, M.: "Distance measures for signal processing and pattern recognition". Signal Processing **18**, 4, 349 – 369 (1989). DOI: 10.1016/0165-1684(89)90079-0.

[232] Almuallim, H. & Dietterich, T.G.: "Learning Boolean concepts in the presence of many irrelevant features". Artificial Intelligence **69**, 1, 279 – 305 (1994). DOI: 10.1016/0004-3702(94)90084-1.

[233] Modrzejewski, M.: "Feature selection using rough sets theory". In: P.B. Brazdil (ed.) "Machine Learning: ECML-93", 213–226. Springer Berlin Heidelberg, Berlin, Heidelberg (1993). ISBN: 978-3-540-47597-2. DOI: 10.1007/3-540-56602-3_138.

[234] Arauzo-Azofra, A.; Benitez, J.M. *et al.*: "Consistency measures for feature selection". Journal of Intelligent Information Systems **30**, 3, 273–292 (2008). DOI: 10.1007/s10844-007-0037-0.

[235] Vergara, J.R. & Estévez, P.A.: "A review of feature selection methods based on mutual information". Neural Computing and Applications **24**, 1, 175–186 (2014). DOI: 10.1007/s00521-013-1368-0.

[236] Guyon, I. & Elisseeff, A.: "An Introduction to Variable and Feature Selection". Journal of Machine Learning Research **3**, 1157–1182 (2003). URL: http://www.jmlr.org/papers/v3/guyon03a.html.

[237] Siedlecki, W. & Sklansky, J.: *On automatic feature selection*, 63–87. World Scientific (1993). DOI: 10.1142/9789814343138_0004.

[238] Eberhart, R. & Kennedy, J.: "A new optimizer using particle swarm theory". In: "MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science", 39–43 (1995). DOI: 10.1109/MHS.1995.494215.

[239] Michalewicz, Z. & Fogel, D.B.: *How to Solve It: Modern Heuristics*. Springer Berlin Heidelberg, Berlin, Heidelberg (2004). ISBN: 978-3-662-07807-5. DOI: 10.1007/978-3-662-07807-5.

[240] Whitney, A.W.: "A Direct Method of Nonparametric Measurement Selection". IEEE Transactions on Computers **C-20**, 9, 1100–1103 (1971). DOI: 10.1109/T-C.1971.223410.

[241] Marill, T. & Green, D.: "On the effectiveness of receptors in recognition systems". IEEE Transactions on Information Theory **9**, 1, 11–17 (1963). DOI: 10.1109/TIT.1963.1057810.

[242] Forsati, R.; Moayedikia, A. *et al.*: "Heuristic Approach to Solve Feature Selection Problem". In: H. Cherifi; J.M. Zain; & E. El-Qawasmeh (eds.) "Digital Information and Communication Technology and Its Applications", 707–717. Springer Berlin Heidelberg, Berlin, Heidelberg (2011). ISBN: 978-3-642-22027-2.

[243] Reunanen, J.: *Search Strategies*, 119–136. Springer Berlin Heidelberg, Berlin, Heidelberg (2006). ISBN: 978-3-540-35488-8. DOI: 10.1007/978-3-540-35488-8_5.

[244] Granata, D. & Carnevale, V.: "Accurate Estimation of the Intrinsic Dimension Using Graph Distances: Unraveling the Geometric Complexity of Datasets". Scientific Reports **6**, 1, 31377 (2016). DOI: 10.1038/srep31377.

[245] Facco, E.; d'Errico, M. *et al.*: "Estimating the intrinsic dimension of datasets by a minimal neighborhood information". Scientific Reports **7**, 1, 12140 (2017). DOI: 10.1038/s41598-017-11873-y.

[246] Allegra, M.; Facco, E. *et al.*: "Data segmentation based on the local intrinsic dimension". Scientific Reports **10**, 1, 16449 (2020). DOI: 10.1038/s41598-020-72222-0.

[247] Van Der Maaten, L.; Postma, E. *et al.*: "Dimensionality reduction: a comparative review". Journal of Machine Learning Research **10**, 1–41, 66–71 (2009). URL: https://lvdmaaten.github.io/publications/#2009.

[248] Bengio, Y.; Courville, A. *et al.*: "Representation Learning: A Review and New Perspectives". IEEE Transactions on Pattern Analysis and Machine Intelligence **35**, 8, 1798–1828 (2013). DOI: 10.1109/TPAMI.2013.50.

[249] Friedman, J.H.: "Greedy function approximation: A gradient boosting machine." The Annals of Statistics **29**, 5, 1189–1232 (2001). DOI: 10.1214/aos/1013203451.

[250] Friedman, J.H.: "Stochastic gradient boosting". Computational Statistics & Data Analysis **38**, 4, 367–378 (2002). DOI: 10.1016/S0167-9473(01)00065-2.

[251] Natekin, A. & Knoll, A.: "Gradient boosting machines, a tutorial". Frontiers in neurorobotics **7**, 21–21 (2013). DOI: 10.3389/fnbot.2013.00021.

[252] Ke, G.; Meng, Q. *et al.*: "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In: I. Guyon; U.V. Luxburg; S. Bengio; H. Wallach; R. Fergus; S. Vishwanathan; & R. Garnett (eds.) "Advances in Neural Information Processing Systems 30", 3146–3154. Curran Associates, Inc. (2017). URL: http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf.

[253] Koller, D. & Sahami, M.: "Toward Optimal Feature Selection". Technical Report 1996-77, Stanford InfoLab, Bari, Italy (1996). URL: http://ilpubs.stanford.edu:8090/208/.

[254] Yu, L. & Liu, H.: "Efficient Feature Selection via Analysis of Relevance and Redundancy". Journal of Machine Learning Research **5**, 1205–1224 (2004). URL: http://www.jmlr.org/papers/v5/yu04a.html.

[255] Pearl, J.: *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann (1988). ISBN: 0-934613-73-7. DOI: 10.1016/C2009-0-27609-4.

[256] Mertens, S.: "Computational Complexity for Physicists". Computing in Science & Engineering **4**, 3, 31–47 (2002). DOI: 10.1109/5992.998639.

[257] Arora, S. & Barak, B.: *Computational Complexity: A Modern Approach*. Cambridge University Press, New York, 1st ed. (2009). ISBN: 978-0-521-42426-4. URL: https://theory.cs.princeton.edu/complexity/.

[258] Maier, W.F.; Stöwe, K. *et al.*: "Combinatorial and High-Throughput Materials Science". Angewandte Chemie International Edition **46**, 32, 6016–6067 (2007). DOI: 10.1002/anie.200603675.

[259] Jović, A.; Brkić, K. *et al.*: "A review of feature selection methods with applications". In: "International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)", 38, 1200–1205 (2015). DOI: 10.1109/MIPRO.2015.7160458.

[260] Bolón-Canedo, V.; Sánchez-Maro no, N. *et al.*: "A review of feature selection methods on synthetic data". Knowledge and Information Systems **34**, 3, 483–519 (2013). DOI: 10.1007/s10115-012-0487-8.

[261] Santosa, F. & Symes, W.W.: "Linear Inversion of Band-Limited Reflection Seismograms". SIAM Journal on Scientific and Statistical Computing **7**, 4, 1307–1330 (1986). DOI: 10.1137/0907087.

[262] Pearson, K.: "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **50**, 302, 157–175 (1900). DOI: 10.1080/14786440009463897.

[263] Guyon, I.; Weston, J. *et al.*: "Gene Selection for Cancer Classification using Support Vector Machines". Machine Learning **46**, 1, 389–422 (2002). DOI: 10.1023/A:1012487302797.

[264] Icke, I. & Bongard, J.C.: "Improving genetic programming based symbolic regression using deterministic machine learning". In: "2013 IEEE Congress on Evolutionary Computation", 1763–1770 (2013). DOI: 10.1109/CEC.2013.6557774.

[265] Lal, T.N.; Chapelle, O. *et al.*: *Embedded Methods*, 137–165. Springer Berlin Heidelberg, Berlin, Heidelberg (2006). ISBN: 978-3-540-35488-8. DOI: 10.1007/978-3-540-35488-8_6.

[266] Breiman, L.: “Random Forests”. Machine Learning **45**, 1, 5–32 (2001). DOI: 10.1023/A:1010933404324.

[267] Meng, Q.; Ke, G. *et al.*: “A Communication-Efficient Parallel Algorithm for Decision Tree”. In: D.D. Lee; M. Sugiyama; U.V. Luxburg; I. Guyon; & R. Garnett (eds.) “Advances in Neural Information Processing Systems 29”, 1279–1287. Curran Associates, Inc. (2016). URL: http://papers.nips.cc/paper/6380-a-communication-efficient-parallel-algorithm-for-decision-tree.

[268] Zhang, H.; Si, S. *et al.*: “GPU-acceleration for Large-scale Tree Boosting”. arXiv e-prints (2017). ARXIV: 1706.08359.

[269] Spearman, C.: “The Proof and Measurement of Association between Two Things”. The American Journal of Psychology **15**, 1, 72–101 (1904). DOI: 10.2307/1412159.

[270] Székely, G.J.; Rizzo, M.L. *et al.*: “Measuring and testing dependence by correlation of distances”. The Annals of Statistics **35**, 6, 2769–2794 (2007). DOI: 10.1214/009053607000000505.

[271] Székely, G.J. & Rizzo, M.L.: “Brownian distance covariance”. The Annals of Applied Statistics **3**, 4, 1236–1265 (2009). DOI: 10.1214/09-AOAS312.

[272] Székely, G.J. & Rizzo, M.L.: “Partial distance correlation with methods for dissimilarities”. The Annals of Statistics **42**, 6, 2382–2412 (2014). DOI: 10.1214/14-AOS1255.

[273] Scott, D.W.: *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, New York (1982). DOI: 10.1002/9780470316849.

[274] Silverman, B.W.: *Density Estimation for Statistics and Data Analysis*, vol. 1. Chapman and Hall/CRC, New York (1986). URL: https://www.crcpress.com/Density-Estimation-for-Statistics-and-Data-Analysis/Silverman/p/book/9780412246203.

[275] Kozachenko, L.F. & Leonenko, N.N.: “Sample Estimate of the Entropy of a Random Vector”. Problems of Information Transmission **23**, 2, 95–101 (1987). URL: http://mi.mathnet.ru/eng/ppi797.

[276] Wilde, M.M.: *Quantum Information Theory*. Cambridge University Press, Cambride, UK, 2nd ed. (2017). ISBN: 9781316809976. DOI: 10.1017/9781316809976. ARXIV: 1106.1445.

[277] Braunstein, S.L. & Pati, A.K.: “Quantum Information Cannot Be Completely Hidden in Correlations: Implications for the Black-Hole Information Paradox”. Physical Review Letters **98**, 080502 (2007). DOI: 10.1103/PhysRevLett.98.080502.

[278] Yockey, H.P.: *Information theory, evolution, and the origin of life*. Cambridge University Press (2005). ISBN: 9780521169585. URL: https://www.cambridge.org/de/academic/subjects/life-sciences/evolutionary-biology/information-theory-evolution-and-origin-life.

[279] Nalewajski, R.F.: *Information Theory of Molecular Systems*. Elsevier Science, Amsterdam (2006). ISBN: 978-0-444-51966-5. DOI: 10.1016/B978-0-444-51966-5.X5063-4.

[280] Tishby, N.; Pereira, F.C. *et al.*: “The information bottleneck method”. physics/0004057 (2000). DOI: physics/0004057.

[281] Tishby, N. & Zaslavsky, N.: “Deep learning and the information bottleneck principle”. In: “2015 IEEE Information Theory Workshop (ITW)”, 1–5 (2015). DOI: 10.1109/ITW.2015.7133169.

[282] Nguyen, H.V.; Müller, E. *et al.*: "Multivariate Maximal Correlation Analysis". In: T. Jebara & E.P. Xing (eds.) "Proceedings of the 31st International Conference on Machine Learning (ICML-14)", vol. 32, 775–783. JMLR Workshop and Conference Proceedings, Beijing, China (2014). URL: http://www.jmlr.org/proceedings/papers/v32/nguyenc14.html.

[283] Reshef, D.N.; Reshef, Y.A. *et al.*: "Detecting Novel Associations in Large Data Sets". Science **334**, 6062, 1518–1524 (2011). DOI: 10.1126/science.1205438.

[284] Nguyen, H.V.; Mandros, P. *et al.*: *Universal Dependency Analysis*, 792–800. Proceedings. Society for Industrial and Applied Mathematics (2016). DOI: 10.1137/1.9781611974348.89.

[285] Wang, Y.; Romano, S. *et al.*: "Unbiased Multivariate Correlation Analysis". In: "Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)", (2017). URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14214.

[286] Kwak, N. & Chong-Ho Choi: "Input feature selection by mutual information based on Parzen window". IEEE Transactions on Pattern Analysis and Machine Intelligence **24**, 12, 1667–1671 (2002). DOI: 10.1109/TPAMI.2002.1114861.

[287] Chow, T.W.S. & Huang, D.: "Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information". IEEE Transactions on Neural Networks **16**, 1, 213–224 (2005). DOI: 10.1109/TNN.2004.841414.

[288] Estevez, P.A.; Tesmer, M. *et al.*: "Normalized Mutual Information Feature Selection". IEEE Transactions on Neural Networks **20**, 2, 189–201 (2009). DOI: 10.1109/TNN.2008.2005601.

[289] Hu, Q.; Zhang, L. *et al.*: "Measuring relevance between discrete and continuous features based on neighborhood mutual information". Expert Systems with Applications **38**, 9, 10737–10750 (2011). DOI: 10.1016/j.eswa.2011.01.023.

[290] Bennasar, M.; Hicks, Y. *et al.*: "Feature selection using Joint Mutual Information Maximisation". Expert Systems with Applications **42**, 22, 8520–8532 (2015). DOI: 10.1016/j.eswa.2015.07.007.

[291] Nguyen, H.V.; Müller, E. *et al.*: *CMI: An Information-Theoretic Contrast Measure for Enhancing Subspace Cluster and Outlier Detection*, chap. 21, 198–206. Proc. SIAM International Conference on Data Mining (2013). DOI: 10.1137/1.9781611972832.22.

[292] Foucart, S. & Rauhut, H.: *A Mathematical Introduction to Compressive Sensing*. Springer, New York, NY (2013). ISBN: 978-0-8176-4948-7. DOI: 10.1007/978-0-8176-4948-7.

[293] Dunleavy, A.J.; Wiesner, K. *et al.*: "Mutual information reveals multiple structural relaxation mechanisms in a model glass former". Nature Communications **6**, 1, 6089 (2015). DOI: 10.1038/ncomms7089.

[294] Mandros, P.; Boley, M. *et al.*: "Discovering dependencies with reliable mutual information". Knowledge and Information Systems **62**, 11, 4223–4253 (2020). DOI: 10.1007/s10115-020-01494-9.

[295] Jaynes, E.T.: "Information Theory and Statistical Mechanics". Physical Review **106**, 620–630 (1957). DOI: 10.1103/PhysRev.106.620.

[296] Jaynes, E.T.: "Information Theory and Statistical Mechanics. II". Physical Review **108**, 171–190 (1957). DOI: 10.1103/PhysRev.108.171.

[297] Kullback, S.: *Information Theory and Statistics*. John Wiley and Sons, New York (1959). URL: https://store.doverpublications.com/0486696847.html.

[298] Reimherr, M. & Nicolae, D.L.: "On Quantifying Dependence: A Framework for Developing Interpretable Measures". Statistical Science **28**, 1, 116–130 (2013). DOI: 10.1214/12-STS405.

[299] Carrara, N. & Ernst, J.: "On the Estimation of Mutual Information". Proceedings **33**, 1 (2020). DOI: 10.3390/proceedings2019033031.

[300] Venn, J.: "I. On the diagrammatic and mechanical representation of propositions and reasonings". The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **10**, 59, 1–18 (1880). DOI: 10.1080/14786448008626877.

[301] Venn, J.: "On the employment of geometrical diagrams for the sensible representations of logical propositions". In: "Proceedings of the Cambridge Philosophical Society", vol. 4, 47–59. Cambridge (1880). URL: https://archive.org/stream/proceedingsofcam4188083camb#page/47/mode/1up.

[302] Shannon, C.E.: "A mathematical theory of communication". The Bell System Technical Journal **27**, 3, 379–423 (1948). DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[303] Pfitzner, D.; Leibbrandt, R. *et al.*: "Characterization and evaluation of similarity measures for pairs of clusterings". Knowledge and Information Systems **19**, 3, 361 (2008). DOI: 10.1007/s10115-008-0150-6.

[304] Vinh, N.X.; Epps, J. *et al.*: "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance". Journal of Machine Learning Research **11**, 2837–2854 (2010). URL: http://www.jmlr.org/papers/v11/vinh10a.html.

[305] Coombs, C.; Dawes, R. *et al.*: *Mathematical Psychology: An Elementary Introduction*. Prentice-Hall, Englewood Cliffs, NJ (1970). URL: https://psycnet.apa.org/record/1970-35004-000.

[306] Press, W.H.; Flannery, B.P. *et al.*: *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge (1988). DOI: 10.1137/1031025.

[307] White, J.V.; Steingold, S. *et al.*: "Performance Metrics for Group-Detection Algorithms". In: Y.H. Said; D.J. Marchette; & J.L. Solka (eds.) "Computing Science and Statistics: Computational Biology and Informatics - Proceedings of the 36th Symposium on the Interface", vol. 36. Baltimore, Maryland (2004). URL: https://www.interfacesymposia.org/I04/I2004Proceedings/WhiteJim/WhiteJim.paper.pdf.

[308] Xu, D. & Tian, Y.: "A Comprehensive Survey of Clustering Algorithms". Annals of Data Science **2**, 2, 165–193 (2015). DOI: 10.1007/s40745-015-0040-1.

[309] Fayyad, U. & Irani, K.: "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning". In: "Proceedings of the 13th Int. Joint Conference on Artificial Intelligence", 1022–1027. Morgan Kaufmann (1993). URL: http://web.donga.ac.kr/kjunwoo/files/Multi%20interval%20discretization%20of%20continuous%20valued%20attributes%20for%20classification%20learning.pdf.

[310] Dougherty, J.; Kohavi, R. *et al.*: "Supervised and Unsupervised Discretization of Continuous Features". In: A. Prieditis & S.J. Russell (eds.) "Machine Learning: Proceedings of the Twelfth International Conference", 194–202. Morgan Kaufmann (1995). URL: http://robotics.stanford.edu/users/sahami/papers-dir/disc.pdf.

[311] Nguyen, H.V.; Müller, E. *et al.*: "Unsupervised interaction-preserving discretization of multivariate data". Data Mining and Knowledge Discovery **28**, 5, 1366–1397 (2014). DOI: 10.1007/s10618-014-0350-5.

[312] Parzen, E.: "On Estimation of a Probability Density Function and Mode". The Annals of Mathematical Statistics **33**, 3, 1065–1076 (1962). DOI: 10.1214/aoms/1177704472.

[313] Garcia, D.: "Robust smoothing of gridded data in one and higher dimensions with missing values". Computational Statistics & Data Analysis **54**, 4, 1167–1178 (2010). DOI: 10.1016/j.csda.2009.09.020.

[314] Bernacchia, A. & Pigolotti, S.: "Self-consistent method for density estimation". Journal of the Royal Statistical Society: Series B (Statistical Methodology) **73**, 3, 407–422 (2011). DOI: 10.1111/j.1467-9868.2011.00772.x.

[315] Keller, F.; Muller, E. *et al.*: "HiCS: High Contrast Subspaces for Density-Based Outlier Ranking". In: "28th International Conference on Data Engineering", 1037–1048. IEEE (2012). DOI: 10.1109/ICDE.2012.88.

[316] O'Brien, T.A.; Collins, W.D. *et al.*: "Reducing the computational cost of the ECF using a nuFFT: A fast and objective probability density estimation method". Computational Statistics & Data Analysis **79**, 222–234 (2014). DOI: 10.1016/j.csda.2014.06.002.

[317] O'Brien, T.A.; Kashinath, K. *et al.*: "A fast and objective multidimensional kernel density estimation method: fastKDE". Computational Statistics & Data Analysis **101**, 148–160 (2016). DOI: 10.1016/j.csda.2016.02.014.

[318] Crescenzo, A.D. & Longobardi, M.: "On Cumulative Entropies and Lifetime Estimations". In: J. Mira; J.M. Ferrández; J.R. Álvarez; F. de la Paz; & F.J. Toledo (eds.) "Methods and Models in Artificial and Natural Computation. A Homage to Professor Mira's Scientific Legacy: Third International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2009, Santiago de Compostela, Spain, June 22-26, 2009, Proceedings, Part I", 132–141. Springer Berlin Heidelberg, Berlin, Heidelberg (2009). ISBN: 978-3-642-02264-7. DOI: 10.1007/978-3-642-02264-7_15.

[319] Crescenzo, A.D. & Longobardi, M.: "On cumulative entropies". Journal of Statistical Planning and Inference **139**, 12, 4072–4087 (2009). DOI: 10.1016/j.jspi.2009.05.038.

[320] Mira, C.: "Noninvertible maps". Scholarpedia **2**, 9, 2328 (2007). DOI: 10.4249/scholarpedia.2328. Revision #153207.

[321] Wang, F.; Vemuri, B.C. *et al.*: *A New & Robust Information Theoretic Measure and Its Application to Image Alignment*, 388–400. Springer Berlin Heidelberg, Berlin, Heidelberg (2003). ISBN: 978-3-540-45087-0. DOI: 10.1007/978-3-540-45087-0_33.

[322] Rao, M.; Chen, Y. *et al.*: "Cumulative residual entropy: a new measure of information". IEEE Transactions on Information Theory **50**, 6, 1220–1228 (2004). DOI: 10.1109/TIT.2004.828057.

[323] Rao, M.: "More on a New Concept of Entropy and Information". Journal of Theoretical Probability **18**, 4, 967–981 (2005). DOI: 10.1007/s10959-005-7541-3.

[324] Beirlant, J.; Dudewicz, E. *et al.*: "Nonparametric entropy estimation. An overview". International Journal Of Mathematical And Statistical Sciences **6**, 1, 17–39 (1997). URL: http://eprints.sztaki.hu/1417/.

[325] Paninski, L.: "Estimation of Entropy and Mutual Information". Neural Computation **15**, 6, 1191–1253 (2003). DOI: 10.1162/089976603321780272.

[326] Dutta, M.: "On Maximum (Information-Theoretic) Entropy Estimation". Sankhyā: The Indian Journal of Statistics, Series A (1961-2002) **28**, 4, 319–328 (1966). URL: http://www.jstor.org/stable/25049432.

[327] Rossi, R.J.: *Mathematical Statistics: An Introduction to Likelihood Based Inference* (2018). ISBN: 978-1-118-77104-4. URL: https://www.wiley.com/en-us/MathematicalStatistics:AnIntroductiontoLikelihoodBasedInference-p-9781118771044.

[328] Glivenko, V.: "Sulla determinazione empirica delle leggi di probabilita". Gion. Ist. Ital. Attauri. **4**, 92–99 (1933). URL: https://ci.nii.ac.jp/naid/10026792179/en/.

[329] Cantelli, F.P.: "Sulla determinazione empirica delle leggi di probabilita". Giorn. Ist. Ital. Attuari **4**, 421–424 (1933).

[330] Vinh, N.X.; Epps, J. *et al.*: "Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary?" In: "Proceedings of the 26th Annual International Conference on Machine Learning", ICML '09, 1073–1080. ACM, New York, NY, USA (2009). ISBN: 978-1-60558-516-1. DOI: 10.1145/1553374.1553511.

[331] Fouché, E. & Böhm, K.: "Monte Carlo Dependency Estimation". In: "Proceedings of the 31st International Conference on Scientific and Statistical Database Management", SSDBM '19, 13–24. ACM, New York, NY, USA (2019). ISBN: 978-1-4503-6216-0. DOI: 10.1145/3335783.3335795.

[332] Romano, S.; Vinh, N.X. *et al.*: "A Framework to Adjust Dependency Measure Estimates for Chance". In: "Proceedings of the 2016 SIAM International Conference on Data Mining", 423–431 (2016). DOI: 10.1137/1.9781611974348.48.

[333] Lancaster, H.O.: *The chi-squared distribution*. Wiley series in probability and mathematical statistics. Wiley & Sons, Inc., New York, NY (1969). ISBN: 0471512303.

[334] Romano, S.; Bailey, J. *et al.*: "Standardized Mutual Information for Clustering Comparisons: One Step Further in Adjustment for Chance". In: T. Jebara & E.P. Xing (eds.) "Proceedings of the 31st International Conference on Machine Learning (ICML-14)", vol. 32, 1143–1151. JMLR Workshop and Conference Proceedings, Beijing, China (2014). URL: http://www.jmlr.org/proceedings/papers/v32/romano14.html.

[335] Li, C.; Hao, H. *et al.*: "A progressive learning method for predicting the band gap of ABO3 perovskites using an instrumental variable". J. Mater. Chem. C **8**, 3127–3136 (2020). DOI: 10.1039/C9TC06632B.

[336] Ring, M. & Eskofier, B.M.: "Optimal feature selection for nonlinear data using branch-and-bound in kernel space". Pattern Recognition Letters **68**, 56–62 (2015). DOI: 10.1016/j.patrec.2015.08.007.

[337] Gregorutti, B.; Michel, B. *et al.*: "Correlation and variable importance in random forests". Statistics and Computing **27**, 3, 659–678 (2017). DOI: 10.1007/s11222-016-9646-1.

[338] Altmann, A.; Toloşi, L. *et al.*: "Permutation importance: a corrected feature importance measure". Bioinformatics **26**, 10, 1340–1347 (2010). DOI: 10.1093/bioinformatics/btq134.

[339] Team Hyperion Gray: "ELI5: A library for debugging/inspecting machine learning classifiers and explaining their predictions" (2016). URL: https://github.com/TeamHG-Memex/eli5.

[340] Cao, G.; Ouyang, R. *et al.*: "Artificial intelligence for high-throughput discovery of topological insulators: The example of alloyed tetradymites". Physical Review Materials **4**, 034204 (2020). DOI: 10.1103/PhysRevMaterials.4.034204.

[341] Dutta, A.; Vreeken, J. *et al.*: "Data-driven equation for drug-membrane permeability across drugs and membranes". The Journal of Chemical Physics **154**, 24, 244114 (2021). DOI: 10.1063/5.0053931. Correction: J. Chem. Phys. 155, 039901 (2021).

[342] Han, Z.K.; Sarker, D. *et al.*: "Single-atom alloy catalysts designed by first-principles calculations and artificial intelligence". Nature Communications **12**, 1, 1833 (2021). DOI: 10.1038/s41467-021-22048-9.

[343] He, N.; Ouyang, R. *et al.*: "Learning interpretable descriptors for the fatigue strength of steels". AIP Advances **11**, 3, 035018 (2021). DOI: 10.1063/5.0045561.

[344] Mao, Y.; Yang, H. *et al.*: "Prediction and Classification of Formation Energies of Binary Compounds by Machine Learning: An Approach without Crystal Structure Information". ACS Omega **6**, 22, 14533–14541 (2021). DOI: 10.1021/acsomega.1c01517.

[345] Alcalá-Fdez, J.; Sánchez, L. *et al.*: "KEEL: a software tool to assess evolutionary algorithms for data mining problems". Soft Computing **13**, 3, 307–318 (2009). DOI: 10.1007/s00500-008-0323-y.

[346] Alcalá-Fdez, J.; Fernández, A. *et al.*: "KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework". Journal of Multiple-Valued Logic and Soft Computing **17**, 2–3, 255–287 (2011). URL: http://www.oldcitypublishing.com/journals/mvlsc-home/mvlsc-issue-contents/mvlsc-volume-17-number-2-3-2011/mvlsc-17-2-3-p-99-100/.

[347] Friedman, J.H.: "Multivariate Adaptive Regression Splines". The Annals of Statistics **19**, 1, 1–67 (1991). DOI: 10.1214/aos/1176347963. The data set can be downloaded from the KEEL repository [345, 346]: https://sci2s.ugr.es/keel/dataset.php?cod=81.

[348] Feng, D.C.; Liu, Z.T. *et al.*: "Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach". Construction and Building Materials **230**, 117000 (2020). DOI: 10.1016/j.conbuildmat.2019.117000.

[349] Dua, D. & Graff, C.: "UCI Machine Learning Repository" (2017). URL: http://archive.ics.uci.edu/ml.

[350] Yeh, I.C.: "Modeling of strength of high-performance concrete using artificial neural networks". Cement and Concrete Research **28**, 12, 1797–1808 (1998). DOI: 10.1016/S0008-8846(98)00165-3. The data set can be downloaded from the UCI Machine Learning Repository [349]: http://archive.ics.uci.edu/ml/datasets/concrete+compressive+strength.

[351] Yeh, I.C.: "Modeling slump of concrete with fly ash and superplasticizer". Computers and Concrete **5**, 6, 559–572 (2008). DOI: 10.12989/cac.2008.5.6.559.

[352] Nagaraj, T. & Banu, Z.: "Generalization of Abrams' law". Cement and Concrete Research **26**, 6, 933–942 (1996). DOI: 10.1016/0008-8846(96)00065-8.

[353] Wrobel, S.: "An algorithm for multi-relational discovery of subgroups". In: J. Komorowski & J. Zytkow (eds.) "Principles of Data Mining and Knowledge Discovery", 78–87. Springer Berlin Heidelberg, Berlin, Heidelberg (1997). ISBN: 978-3-540-69236-2. DOI: 10.1007/3-540-63223-9_108.

[354] Friedman, J.H. & Fisher, N.I.: "Bump hunting in high-dimensional data". Statistics and Computing **9**, 2, 123–143 (1999). DOI: 10.1023/A:1008894516817.

[355] Herrera, F.; Carmona, C.J. *et al.*: "An overview on subgroup discovery: foundations and applications". Knowledge and Information Systems **29**, 3, 495–525 (2011). DOI: 10.1007/s10115-010-0356-2.

[356] Atzmueller, M.: "Subgroup Discovery". Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **5**, 1, 35–49 (2015). DOI: 10.1002/widm.1144.

[357] Mehta, P.; Bukov, M. *et al.*: "A high-bias, low-variance introduction to Machine Learning for physicists". Physics Reports **810**, 1–124 (2019). DOI: 10.1016/j.physrep.2019.03.001. A high-bias, low-variance introduction to Machine Learning for physicists.

[358] Breiman, L.: "Statistical Modeling: The Two Cultures". Statistical Science **16**, 3, 199–215 (2001). DOI: 10.1214/ss/1009213726.

[359] Reunanen, J.: "A Pitfall in Determining the Optimal Feature Subset Size". In: "Proceedings of the 4th International Workshop on Pattern Recognition in Information Systems - Volume 1: PRIS, (ICEIS 2004)", 176–185 (2004). ISBN: 972-8865-01-5. DOI: 10.5220/0002650001760185.

[360] Pes, B.: "Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains". Neural Computing and Applications **32**, 10, 5951–5973 (2020). DOI: 10.1007/s00521-019-04082-3.

[361] Foroutan, I. & Sklansky, J.: "Feature Selection for Automatic Classification of Non-Gaussian Data". IEEE Transactions on Systems, Man, and Cybernetics **17**, 2, 187–198 (1987). DOI: 10.1109/TSMC.1987.4309029.

[362] Lu, F. & Petkova, E.: "A comparative study of variable selection methods in the context of developing psychiatric screening instruments". Statistics in Medicine **33**, 3, 401–421 (2014). DOI: 10.1002/sim.5937.

[363] Peterson, A.A.; Christensen, R. *et al.*: "Addressing uncertainty in atomistic machine learning". Physical Chemistry Chemical Physics **19**, 10978–10985 (2017). DOI: 10.1039/C7CP00375G.

[364] Honarmandi, P. & Arróyave, R.: "Uncertainty Quantification and Propagation in Computational Materials Science and Simulation-Assisted Materials Design". Integrating Materials and Manufacturing Innovation **9**, 1, 103–143 (2020). DOI: 10.1007/s40192-020-00168-2.

[365] Musil, F.; Willatt, M.J. *et al.*: "Fast and Accurate Uncertainty Estimation in Chemical Machine Learning". Journal of Chemical Theory and Computation **15**, 2, 906–915 (2019). DOI: 10.1021/acs.jctc.8b00959.

[366] Papadopoulos, H.: "Inductive Conformal Prediction: Theory and Application to Neural Networks". In: P. Fritzsche (ed.) "Tools in Artificial Intelligence", chap. 18. IntechOpen, Rijeka (2008). DOI: 10.5772/6078.

[367] Bassman, L.; Rajak, P. *et al.*: "Active learning for accelerated design of layered materials". npj Computational Materials **4**, 1, 74 (2018). DOI: 10.1038/s41524-018-0129-0.

[368] Gopakumar, A.M.; Balachandran, P.V. *et al.*: "Multi-objective Optimization for Materials Discovery via Adaptive Design". Scientific Reports **8**, 1, 3738 (2018). DOI: 10.1038/s41598-018-21936-3.

[369] Fisher, R.A.: "Applications of "Student's" distribution". Metron **5**, 90–104 (1925). URL: http://hdl.handle.net/2440/15187.

[370] Domingos, P.: "The Role of Occam's Razor in Knowledge Discovery". Data Mining and Knowledge Discovery **3**, 4, 409–425 (1999). DOI: 10.1023/A:1009868929893.

[371] Fan, J. & Lv, J.: "Sure independence screening for ultrahigh dimensional feature space". Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70**, 5, 849–911 (2008). DOI: 10.1111/j.1467-9868.2008.00674.x.

[372] Hochachka, W.M.; Caruana, R. *et al.*: "Data-Mining Discovery of Pattern and Process in Ecological Systems". The Journal of Wildlife Management **71**, 7, 2427–2437 (2007). DOI: 10.2193/2006-503.

[373] Li, F. & Yang, Y.: "Analysis of Recursive Feature Elimination Methods". In: "Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval", SIGIR '05, 633–634. Association for Computing Machinery, New York, NY, USA (2005). ISBN: 1595930345. DOI: 10.1145/1076034.1076164.

[374] Chen, X. & Jeong, J.C.: "Enhanced recursive feature elimination". In: "Sixth International Conference on Machine Learning and Applications (ICMLA 2007)", 429–435 (2007). DOI: 10.1109/ICMLA.2007.35.

[375] Goldstein, A.; Kapelner, A. *et al.*: "Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation". Journal of Computational and Graphical Statistics **24**, 1, 44–65 (2015). DOI: 10.1080/10618600.2014.907095.

[376] Ribeiro, M.T.; Singh, S. *et al.*: ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: "Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining", KDD '16, 1135–1144. Association for Computing Machinery, New York, NY, USA (2016). ISBN: 9781450342322. DOI: 10.1145/2939672.2939778.

[377] Lundberg, S.M. & Lee, S.I.: "A Unified Approach to Interpreting Model Predictions". In: I. Guyon; U.V. Luxburg; S. Bengio; H. Wallach; R. Fergus; S. Vishwanathan; & R. Garnett (eds.) "Advances in Neural Information Processing Systems 30", 4765–4774. Curran Associates, Inc. (2017). URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predicti.

[378] Mentch, L. & Hooker, G.: "Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests". The Journal of Machine Learning Research **17**, 26, 1–41 (2016). URL: http://jmlr.org/papers/v17/14-168.html.

[379] Ghahramani, Z.: "Probabilistic machine learning and artificial intelligence". Nature **521**, 7553, 452–459 (2015). DOI: 10.1038/nature14541.

[380] Botu, V.; Batra, R. *et al.*: "Machine Learning Force Fields: Construction, Validation, and Outlook". The Journal of Physical Chemistry C **121**, 1, 511–522 (2017). DOI: 10.1021/acs.jpcc.6b10908.

[381] Balachandran, P.V.; Xue, D. *et al.*: "Adaptive Strategies for Materials Design using Uncertainties". Scientific Reports **6**, 1, 19660 (2016). DOI: 10.1038/srep19660.

[382] Terayama, K.; Tamura, R. *et al.*: "Efficient construction method for phase diagrams using uncertainty sampling". Physical Review Materials **3**, 033802 (2019). DOI: 10.1103/PhysRevMaterials.3.033802.

[383] Ling, J.; Hutchinson, M. *et al.*: "High-Dimensional Materials and Process Optimization Using Data-Driven Experimental Design with Well-Calibrated Uncertainty Estimates". Integrating Materials and Manufacturing Innovation **6**, 3, 207–217 (2017). DOI: 10.1007/s40192-017-0098-z.

[384] *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*, vol. 48 of *ICML'16*. JMLR.org (2016). URL: https://dl.acm.org/doi/10.5555/3045390.3045502.

[385] Alexander, F. & Lookman, T.: "Chapter 3 - Novel Approaches to Statistical Learning in Materials Science". In: K. Rajan (ed.) "Informatics for Materials Science and Engineering", 37–51. Butterworth-Heinemann, Oxford (2013). ISBN: 978-0-12-394399-6. DOI: 10.1016/B978-0-12-394399-6.00003-5.

[386] Norinder, U.; Carlsson, L. *et al.*: "Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination". Journal of Chemical Information and Modeling **54**, 6, 1596–1603 (2014). DOI: 10.1021/ci5001168.

[387] Boström, H.; Linusson, H. *et al.*: "Accelerating difficulty estimation for conformal regression forests". Annals of Mathematics and Artificial Intelligence **81**, 1, 125–144 (2017). DOI: 10.1007/s10472-017-9539-9.

[388] Lei, J.; G'Sell, M. *et al.*: "Distribution-Free Predictive Inference for Regression". Journal of the American Statistical Association **113**, 523, 1094–1111 (2018). DOI: 10.1080/01621459.2017.1307116.

[389] Beyer, K.; Goldstein, J. *et al.*: "When Is "Nearest Neighbor" Meaningful?" In: C. Beeri & P. Buneman (eds.) "Database Theory — ICDT'99", 217–235. Springer Berlin Heidelberg, Berlin, Heidelberg (1999). ISBN: 978-3-540-49257-3. DOI: 10.1007/3-540-49257-7_15.

[390] Durrant, R.J. & Kabán, A.: "When is 'nearest neighbour' meaningful: A converse theorem and implications". Journal of Complexity **25**, 4, 385–397 (2009). DOI: 10.1016/j.jco.2009.02.011.

[391] Papadopoulos, H.; Proedrou, K. *et al.*: "Inductive Confidence Machines for Regression". In: T. Elomaa; H. Mannila; & H. Toivonen (eds.) "Machine Learning: ECML 2002", 345–356. Springer, Berlin, Heidelberg (2002). ISBN: 978-3-540-36755-0. DOI: 10.1007/3-540-36755-1_29.

[392] Papadopoulos, H.; Vovk, V. *et al.*: "Qualified predictions for large data sets in the case of pattern recognition". In: M. Wani; H. Arabnia; K. Cios; K. Hafeez; & G. Kendall (eds.) "Proceedings of the International Conference on Machine Learning and Applications (ICMLA'02)", 159–163. CSREA Press, Las Vegas, NV (2002).

[393] Carlsson, L.; Eklund, M. *et al.*: "Aggregated Conformal Prediction". In: L. Iliadis; I. Maglogiannis; H. Papadopoulos; S. Sioutas; & C. Makris (eds.) "Artificial Intelligence Applications and Innovations", 231–240. Springer Berlin Heidelberg, Berlin, Heidelberg (2014). ISBN: 978-3-662-44722-2. DOI: 10.1007/978-3-662-44722-2_25.

[394] Linusson, H.; Norinder, U. *et al.*: "On the Calibration of Aggregated Conformal Predictors". In: A. Gammerman; V. Vovk; Z. Luo; & H. Papadopoulos (eds.) "Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications", vol. 60 of *Proceedings of Machine Learning Research*, 154–173. PMLR, Stockholm, Sweden (2017). URL: http://proceedings.mlr.press/v60/linusson17a.html.

[395] Vovk, V.: "Cross-conformal predictors". Annals of Mathematics and Artificial Intelligence **74**, 1, 9–28 (2015). DOI: 10.1007/s10472-013-9368-4.

[396] Linusson, H.; Johansson, U. *et al.*: "Efficiency Comparison of Unstable Transductive and Inductive Conformal Classifiers". In: L. Iliadis; I. Maglogiannis; H. Papadopoulos; S. Sioutas; & C. Makris (eds.) "Artificial Intelligence Applications and Innovations", 261–270. Springer Berlin Heidelberg, Berlin, Heidelberg (2014). ISBN: 978-3-662-44722-2. DOI: 10.1007/978-3-662-44722-2_28.

[397] Johansson, U.; Ahlberg, E. *et al.*: "Handling Small Calibration Sets in Mondrian Inductive Conformal Regressors". In: A. Gammerman; V. Vovk; & H. Papadopoulos (eds.) "Statistical Learning and Data Sciences", 271–280. Springer International Publishing, Cham (2015). ISBN: 978-3-319-17091-6. DOI: 10.1007/978-3-319-17091-6_22.

[398] Linusson, H.: "Nonconformist: Python implementation of the conformal prediction framework." (2015). URL: https://github.com/donlnz/nonconformist.

[399] Cortés-Ciriano, I. & Bender, A.: "Reliable Prediction Errors for Deep Neural Networks Using Test-Time Dropout". Journal of Chemical Information and Modeling **59**, 7, 3330–3339 (2019). DOI: 10.1021/acs.jcim.9b00297.

[400] Laxhammar, R. & Falkman, G.: "Sequential Conformal Anomaly Detection in trajectories based on Hausdorff distance". In: "14th International Conference on Information Fusion", 1–8 (2011). URL: https://ieeexplore.ieee.org/document/5977571.

[401] Laxhammar, R. & Falkman, G.: "Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories". Annals of Mathematics and Artificial Intelligence **74**, 1, 67–94 (2015). DOI: 10.1007/s10472-013-9381-7.

[402] Hawkins, D.M.: *Identification of Outliers*. Springer Netherlands, Dordrecht (1980). ISBN: 978-94-015-3994-4. DOI: 10.1007/978-94-015-3994-4.

[403] Rousseeuw, P.J. & Driessen, K.V.: "A Fast Algorithm for the Minimum Covariance Determinant Estimator". Technometrics **41**, 3, 212–223 (1999). DOI: 10.1080/00401706.1999.10485670.

[404] Liu, F.T.; Ting, K.M. *et al.*: "Isolation Forest". In: "2008 Eighth IEEE International Conference on Data Mining", 413–422 (2008). DOI: 10.1109/ICDM.2008.17.

[405] MacQueen, J.: "Some methods for classification and analysis of multivariate observations". In: "Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics", 281–297. University of California Press, Berkeley, California (1967). URL: https://projecteuclid.org/euclid.bsmsp/1200512992.

[406] Schölkopf, B.; Platt, J.C. *et al.*: "Estimating the Support of a High-Dimensional Distribution". Neural Computation **13**, 7, 1443–1471 (2001). DOI: 10.1162/089976601750264965.

[407] Ester, M.; Kriegel, H.P. *et al.*: "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: "KDD", 226–231. AAAI Press, Portland, OR (1996). URL: http://www.aaai.org/Library/KDD/1996/kdd96-037.php.

[408] Breunig, M.M.; Kriegel, H.P. *et al.*: "LOF: Identifying Density-Based Local Outliers". In: "Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data", SIGMOD '00, 93–104. Association for Computing Machinery, New York, NY, USA (2000). ISBN: 1581132174. DOI: 10.1145/342009.335388.

[409] Campello, R.J.G.B.; Moulavi, D. *et al.*: "Density-Based Clustering Based on Hierarchical Density Estimates". In: J. Pei; V.S. Tseng; L. Cao; H. Motoda; & G. Xu (eds.) "Advances in Knowledge Discovery and Data Mining", 160–172. Springer Berlin Heidelberg, Berlin, Heidelberg (2013). ISBN: 978-3-642-37456-2. DOI: 10.1007/978-3-642-37456-2_14.

[410] Verma, A. & Kumar, A.: "Bulk modulus of cubic perovskites". Journal of Alloys and Compounds **541**, 210–214 (2012). DOI: 10.1016/j.jallcom.2012.07.027.

[411] Foppa, L.; Scheffler, M. *et al.*: "Materials-genome (descriptor) identification by the hierarchical SISSO approach". In: "Conference on a FAIR Data Infrastructure for Materials Genomics", FAIR-DI e.V, Virtual Meeting (2020). URL: https://th.fhi-berlin.mpg.de/meetings/fairdi2020/index.php?n=Meeting.PosterDetails&poster_id=41. Private Communication.

[412] de Jong, M.; Chen, W. *et al.*: "Charting the complete elastic properties of inorganic crystalline compounds". Scientific Data **2**, 1, 150009 (2015). DOI: 10.1038/sdata.2015.9. Dryad Digital Repository: 10.5061/dryad.h505v.

[413] Jaccard, P.: "Lois de distribution florale dans la zone alpine". Bulletin de la Société Vaudoise des Sciences Naturelles **38**, 144, 69–130 (1902). DOI: 10.1111/j.1469-8137.1912.tb05611.x. English version: 10.1111/j.1469-8137.1912.tb05611.x.

[414] Tanimoto, T.T.: *An elementary mathematical theory of classification and prediction*. International Business Machines Corporation, New York (1958). URL: http://dalkescientific.com/tanimoto.pdf.

[415] Dietterich, T.G.: "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms". Neural Computation **10**, 7, 1895–1923 (1998). DOI: 10.1162/089976698300017197.

[416] Polikar, R.: *Ensemble Learning*, chap. 1, 1–34. Springer US, Boston, MA (2012). ISBN: 978-1-4419-9326-7. DOI: 10.1007/978-1-4419-9326-7_1.

[417] Phillips, J.C.: "Ionicity of the Chemical Bond in Crystals". Reviews of Modern Physics **42**, 317–356 (1970). DOI: 10.1103/RevModPhys.42.317.

[418] John, J. & Bloch, A.N.: "Quantum-Defect Electronegativity Scale for Nontransition Elements". Physical Review Lett. **33**, 1095–1098 (1974). DOI: 10.1103/PhysRevLett.33.1095.

[419] Zunger, A.: "Systematization of the stable crystal structure of all AB-type binary compounds: A pseudopotential orbital-radii approach". Physical Review B **22**, 5839–5872 (1980). DOI: 10.1103/PhysRevB.22.5839.

[420] Chelikowsky, J.R.: "Diagrammatic separation scheme for transition-metal binary compounds". Physical Review B **26**, 3433–3435 (1982). DOI: 10.1103/PhysRevB.26.3433.

[421] Pettifor, D.: "A chemical scale for crystal-structure maps". Solid State Communications **51**, 1, 31–34 (1984). DOI: 10.1016/0038-1098(84)90765-8.

[422] Villars, P.: "A semiempirical approach to the prediction of compound formation for 3486 binary alloy systems". Journal of the Less Common Metals **109**, 1, 93–115 (1985). DOI: 10.1016/0022-5088(85)90110-9.

[423] Andreoni, W.; Galli, G. *et al.*: "Structural classification of AB_2 molecules and A_3 clusters from valence electron orbital radii". Physical Review Lett. **55**, 1734–1737 (1985). DOI: 10.1103/PhysRevLett.55.1734.

[424] Andreoni, W. & Galli, G.: "Structural classification of polyatomic molecules based on valence electron orbital radii: AB_3 and A_2B_2 compounds". Physical Review Lett. **58**, 2742–2745 (1987). DOI: 10.1103/PhysRevLett.58.2742.

[425] Saad, Y.; Gao, D. *et al.*: "Data mining for materials: Computational experiments with *AB* compounds". Physical Review B **85**, 104104 (2012). DOI: 10.1103/PhysRevB.85.104104.

[426] Pilania, G.; Gubernatis, J.E. *et al.*: "Classification of octet AB-type binary compounds using dynamical charges: A materials informatics perspective". Scientific Reports **5**, 1, 17504 (2015). DOI: 10.1038/srep17504.

[427] Van Vechten, J.A.: "Quantum Dielectric Theory of Electronegativity in Covalent Systems. I. Electronic Dielectric Constant". Physical Review **182**, 891–905 (1969). DOI: 10.1103/PhysRev.182.891.

[428] Ceperley, D.M. & Alder, B.J.: "Ground State of the Electron Gas by a Stochastic Method". Physical Review Letters **45**, 566–569 (1980). DOI: 10.1103/PhysRevLett.45.566.

[429] Aggarwal, R. & Ranganathan, P.: "Common pitfalls in statistical analysis: The use of correlation techniques". Perspectives in clinical research **7**, 4, 187–190 (2016). DOI: 10.4103/2229-3485.192046.

[430] Simons, G.: "New Model Potential for Pseudopotential Calculations". The Journal of Chemical Physics **55**, 2, 756–761 (1971). DOI: 10.1063/1.1676142.

[431] Jain, A. & Zongker, D.: "Feature selection: evaluation, application, and small sample performance". IEEE Transactions on Pattern Analysis and Machine Intelligence **19**, 2, 153–158 (1997). DOI: 10.1109/34.574797.

[432] Veprek, S.: "The search for novel, superhard materials". Journal of Vacuum Science & Technology A **17**, 5, 2401–2420 (1999). DOI: 10.1116/1.581977.

[433] Léger, J.M. & Haines, J.: "The search for superhard materials". Endeavour **21**, 3, 121–124 (1997). DOI: 10.1016/S0160-9327(97)80221-9.

[434] Teter, D.M.: "Computational Alchemy: The Search for New Superhard Materials". MRS Bulletin **23**, 1, 22–27 (1998). DOI: 10.1557/S0883769400031420.

[435] Brazhkin, V.V. & Solozhenko, V.L.: "Myths about new ultrahard phases: Why materials that are significantly superior to diamond in elastic moduli and hardness are impossible". Journal of Applied Physics **125**, 13, 130901 (2019). DOI: 10.1063/1.5082739.

[436] Katz, E.A.: "Perovskite: Name Puzzle and German-Russian Odyssey of Discovery". Helvetica Chimica Acta **103**, 6, e2000061 (2020). DOI: 10.1002/hlca.202000061.

[437] Voorhoeve, R.J.H.; Johnson, D.W. *et al.*: "Perovskite Oxides: Materials Science in Catalysis". Science **195**, 4281, 827–833 (1977). DOI: 10.1126/science.195.4281.827.

[438] Tanaka, H. & Misono, M.: "Advances in designing perovskite catalysts". Current Opinion in Solid State and Materials Science **5**, 5, 381–387 (2001). DOI: 10.1016/S1359-0286(01)00035-3.

[439] Bednorz, J.G. & Müller, K.A.: "Perovskite-type oxides — The new approach to high-T_c superconductivity". Rev. Mod. Phys. **60**, 585–600 (1988). DOI: 10.1103/RevModPhys.60.585.

[440] Rao, C.N.R.: "Perovskite oxides and high-temperature superconductivity". Ferroelectrics **102**, 1, 297–308 (1990). DOI: 10.1080/00150199008221489.

[441] Benedek, N.A. & Fennie, C.J.: "Why Are There So Few Perovskite Ferroelectrics?" The Journal of Physical Chemistry C **117**, 26, 13339–13349 (2013). DOI: 10.1021/jp402046t.

[442] Körbel, S.; Marques, M.A.L. *et al.*: "Stability and electronic properties of new inorganic perovskites from high-throughput ab initio calculations". J. Mater. Chem. C **4**, 3157–3167 (2016). DOI: 10.1039/C5TC04172D.

[443] Schmidt, J.; Shi, J. *et al.*: "Predicting the Thermodynamic Stability of Solids Combining Density Functional Theory and Machine Learning". Chemistry of Materials **29**, 12, 5090–5103 (2017). DOI: 10.1021/acs.chemmater.7b00156.

[444] Hwang, J.; Rao, R.R. *et al.*: "Perovskites in catalysis and electrocatalysis". Science **358**, 6364, 751–756 (2017). DOI: 10.1126/science.aam7092.

[445] Tonui, P.; Oseni, S.O. *et al.*: "Perovskites photovoltaic solar cells: An overview of current status". Renewable and Sustainable Energy Reviews **91**, 1025–1044 (2018). DOI: 10.1016/j.rser.2018.04.069.

[446] Teh, Y.W.; Chee, M.K.T. *et al.*: "An insight into perovskite-based photocatalysts for artificial photosynthesis". Sustainable Energy Fuels **4**, 973–984 (2020). DOI: 10.1039/C9SE00526A.

[447] Goodenough, J.B.: "Electronic and ionic transport properties and other physical aspects of perovskites". Reports on Progress in Physics **67**, 11, 1915–1993 (2004). DOI: 10.1088/0034-4885/67/11/r01.

[448] Zhu, J.; Li, H. *et al.*: "Perovskite Oxides: Preparation, Characterizations, and Applications in Heterogeneous Catalysis". ACS Catalysis **4**, 9, 2917–2940 (2014). DOI: 10.1021/cs500606g.

[449] Blum, V.; Gehrke, R. *et al.*: "Ab initio molecular simulations with numeric atom-centered orbitals". Computer Physics Communications **180**, 11, 2175–2196 (2009). DOI: 10.1016/j.cpc.2009.06.022.

[450] van Lenthe, E.; Baerends, E.J. *et al.*: "Relativistic total energy using regular approximations". The Journal of Chemical Physics **101**, 11, 9783–9792 (1994). DOI: 10.1063/1.467943.

[451] Perdew, J.P.; Ruzsinszky, A. *et al.*: "Restoring the Density-Gradient Expansion for Exchange in Solids and Surfaces". Physical Review Letters **100**, 136406 (2008). DOI: 10.1103/PhysRevLett.100.136406. Erratum: Physical Review Lett. 102, 039902 (2009).

[452] Lenz, M.O.; Purcell, T.A.R. *et al.*: "Parametrically constrained geometry relaxations for high-throughput materials science". npj Computational Materials **5**, 1, 123 (2019). DOI: 10.1038/s41524-019-0254-4.

[453] Broyden, C.G.: "The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations". IMA Journal of Applied Mathematics **6**, 1, 76–90 (1970). DOI: 10.1093/imamat/6.1.76.

[454] Fletcher, R.: "A new approach to variable metric algorithms". The Computer Journal **13**, 3, 317–322 (1970). DOI: 10.1093/comjnl/13.3.317.

[455] Goldfarb, D.: "A Family of Variable-Metric Methods Derived by Variational Means". Mathematics of Computation **24**, 109, 23–26 (1970). DOI: 10.1090/S0025-5718-1970-0258249-6.

[456] Shanno, D.F.: "Conditioning of Quasi-Newton Methods for Function Minimization". Mathematics of Computation **24**, 111, 647–656 (1970). DOI: 10.1090/S0025-5718-1970-0274029-X.

[457] Birch, F.: "Finite Elastic Strain of Cubic Crystals". Physical Review **71**, 809–824 (1947). DOI: 10.1103/PhysRev.71.809.

[458] Murnaghan, F.D.: "The Compressibility of Media under Extreme Pressures". Proceedings of the National Academy of Sciences **30**, 9, 244–247 (1944). DOI: 10.1073/pnas.30.9.244.

[459] Heyd, J.; Scuseria, G.E. *et al.*: "Hybrid functionals based on a screened Coulomb potential". The Journal of Chemical Physics **118**, 18, 8207–8215 (2003). DOI: 10.1063/1.1564060.

[460] Heyd, J. & Scuseria, G.E.: "Efficient hybrid density functional calculations in solids: Assessment of the Heyd-Scuseria-Ernzerhof screened Coulomb hybrid functional". The Journal of Chemical Physics **121**, 3, 1187–1192 (2004). DOI: 10.1063/1.1760074.

[461] Pokluda, J.; Černý, M. *et al.*: "Ab initio calculations of mechanical properties: Methods and applications". Progress in Materials Science **73**, 127–158 (2015). DOI: 10.1016/j.pmatsci.2015.04.001.

[462] Cahill, D.G.; Watson, S.K. *et al.*: "Lower limit to the thermal conductivity of disordered crystals". Physical Review B **46**, 6131–6140 (1992). DOI: 10.1103/PhysRevB.46.6131.

[463] Snyder, G.J. & Toberer, E.S.: "Complex thermoelectric materials". Nature Materials **7**, 2, 105–114 (2008). DOI: 10.1038/nmat2090.

[464] Pugh, S.: "XCII. Relations between the elastic moduli and the plastic properties of polycrystalline pure metals". The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **45**, 367, 823–843 (1954). DOI: 10.1080/14786440808520496.

[465] Gilman, J.J.: "Why Silicon Is Hard". Science **261**, 5127, 1436–1439 (1993). DOI: 10.1126/science.261.5127.1436.

[466] Kaner, R.B.; Gilman, J.J. *et al.*: "Designing Superhard Materials". Science **308**, 5726, 1268–1269 (2005). DOI: 10.1126/science.1109830.

[467] Lan, H. & Venkatesh, T.: "On the relationships between hardness and the elastic and plastic properties of isotropic power-law hardening materials". Philosophical Magazine **94**, 1, 35–55 (2014). DOI: 10.1080/14786435.2013.839889.

[468] Šimůnek, A. & Vackář, J.: "Hardness of Covalent and Ionic Crystals: First-Principle Calculations". Physical Review Lett. **96**, 085501 (2006). DOI: 10.1103/PhysRevLett.96.085501.

[469] Gao, F.: "Hardness of cubic solid solutions". Scientific Reports **7**, 1, 40276 (2017). DOI: 10.1038/srep40276.

[470] Wang, J.; Yang, X. *et al.*: "New methods for prediction of elastic constants based on density functional theory combined with machine learning". Computational Materials Science **138**, 135–148 (2017). DOI: 10.1016/j.commatsci.2017.06.015.

[471] Vacher, R. & Boyer, L.: "Brillouin Scattering: A Tool for the Measurement of Elastic and Photoelastic Constants". Physical Review B **6**, 639–673 (1972). DOI: 10.1103/PhysRevB.6.639.

[472] Dil, J.G.: "Brillouin scattering in condensed matter". Reports on Progress in Physics **45**, 3, 285–334 (1982). DOI: 10.1088/0034-4885/45/3/002.

[473] Furmanchuk, A.; Agrawal, A. *et al.*: "Predictive analytics for crystalline materials: bulk modulus". RSC Adv. **6**, 95246–95251 (2016). DOI: 10.1039/C6RA19284J.

[474] Xie, T. & Grossman, J.C.: "Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties". Physical Review Lett. **120**, 145301 (2018). DOI: 10.1103/PhysRevLett.120.145301. URL: http://megnet.crystals.ai/.

[475] Loader, C.: *Local Regression and Likelihood*. Statistics and Computing. Springer New York, New York, Berlin, Heidelberg (1999). ISBN: 978-0-387-22732-0. DOI: 10.1007/b98858.

[476] Park, C.W. & Wolverton, C.: "Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery". Physical Review Materials **4**, 063801 (2020). DOI: 10.1103/PhysRevMaterials.4.063801.

[477] Mentel, L.: "mendeleev – A Python resource for properties of chemical elements, ions and isotopes, ver. 0.6.0" (2014). URL: https://github.com/lmmentel/mendeleev.

[478] Cardarelli, F.: *Materials Handbook: A Concise Desktop Reference*. Springer London, London (2008). ISBN: 978-1-84628-669-8. DOI: 10.1007/978-1-84628-669-8.

[479] Rumble, J. (ed.): *CRC Handbook of Chemistry and Physics*. CRC Press, Boca Raton, Florida, 101st ed. (2020). ISBN: 9780367417246. URL: http://hbcponline.com.

[480] Blöchl, P.E.: "Projector augmented-wave method". Physical Review B **50**, 17953–17979 (1994). DOI: 10.1103/PhysRevB.50.17953.

[481] Kresse, G. & Joubert, D.: "From ultrasoft pseudopotentials to the projector augmented-wave method". Physical Review B **59**, 1758–1775 (1999). DOI: 10.1103/PhysRevB.59.1758.

[482] Perdew, J.P.; Burke, K. *et al.*: "Generalized Gradient Approximation Made Simple". Physical Review Letters **77**, 3865–3868 (1996). DOI: 10.1103/PhysRevLett.77.3865.

[483] Voigt, W.: *Lehrbuch der Kristallphysik*, vol. 12. Vieweg+Teubner Verlag, Wiesbaden (1928). DOI: 10.1007/978-3-663-15884-4.

[484] Reuss, A.: "Berechnung der Fließgrenze von Mischkristallen auf Grund der Plastizitätsbedingung für Einkristalle". ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik **9**, 1, 49–58 (1929). DOI: 10.1002/zamm.19290090104.

[485] Hill, R.: "The Elastic Behaviour of a Crystalline Aggregate". Proceedings of the Physical Society. Section A **65**, 5, 349–354 (1952). DOI: 10.1088/0370-1298/65/5/307.

[486] Liu, A.Y. & Cohen, M.L.: "Prediction of New Low Compressibility Solids". Science **245**, 4920, 841–842 (1989). DOI: 10.1126/science.245.4920.841.

[487] Gilman, J.J.: "Chemical and physical "hardness"". Materials Research Innovations **1**, 2, 71–76 (1997). DOI: 10.1007/s100190050023.

[488] Yang, W.; Parr, R.G. *et al.*: "New relation between hardness and compressibility of minerals". Physics and Chemistry of Minerals **15**, 2, 191–195 (1987). DOI: 10.1007/BF00308783.

[489] Li & Wu, P.: "Correlation of Bulk Modulus and the Constituent Element Properties of Binary Intermetallic Compounds". Chemistry of Materials **13**, 12, 4642–4648 (2001). DOI: 10.1021/cm0104203.

[490] Hermansen, C.; Matsuoka, J. *et al.*: "Densification and plastic deformation under microindentation in silicate glasses and the relation to hardness and crack resistance". Journal of Non-Crystalline Solids **364**, 40–43 (2013). DOI: 10.1016/j.jnoncrysol.2012.12.047.

[491] Villars, P.; Cenzual, K. *et al.*: "Data-driven atomic environment prediction for binaries using the Mendeleev number: Part 1. Composition AB". Journal of Alloys and Compounds **367**, 1, 167–175 (2004). DOI: 10.1016/j.jallcom.2003.08.060. Proceedings of the VIII International Conference on Crystal Chemistry of Intermetallic Compounds.

[492] Meija, J.; Coplen, T.B. *et al.*: "Atomic weights of the elements 2013 (IUPAC Technical Report)". Pure and Applied Chemistry **88**, 3, 265–291 (2016). DOI: 10.1515/pac-2015-0305.

[493] Wang, M.; Audi, G. *et al.*: "The AME2016 atomic mass evaluation (II). Tables, graphs and references". Chinese Physics C **41**, 3, 030003 (2017). DOI: 10.1088/1674-1137/41/3/030003.

[494] Bratsch, S.G. & Lagowski, J.: "Predicted stabilities of monatomic anions in water and liquid ammonia at 298.15 K". Polyhedron **5**, 11, 1763–1770 (1986). DOI: 10.1016/S0277-5387(00)84854-8.

[495] Guo, Y. & Whitehead, M.A.: "Electron affinities of alkaline-earth and actinide elements calculated with the local-spin-density-functional theory". Physical Review A **40**, 28–34 (1989). DOI: 10.1103/PhysRevA.40.28.

[496] Felfli, Z.; Msezane, A.Z. *et al.*: "Resonances in low-energy electron elastic cross sections for lanthanide atoms". Physical Review A **79**, 012714 (2009). DOI: 10.1103/PhysRevA.79.012714.

[497] Andersen, T.: "Atomic negative ions: structure, dynamics and collisions". Physics Reports **394**, 4, 157–313 (2004). DOI: 10.1016/j.physrep.2004.01.001.

[498] Pearson, R.G.: "Absolute electronegativity and hardness: application to inorganic chemistry". Inorganic Chemistry **27**, 4, 734–740 (1988). DOI: 10.1021/ic00277a030.

[499] Kramida, A.; Ralchenko, Y. *et al.*: "NIST Atomic Spectra Database (version 5.3)". Available: https://physics.nist.gov/asd (2015). DOI: 10.18434/T4W30F. [Online; accessed 13-April-2015].

[500] Slater, J.C.: "Atomic Shielding Constants". Physical Review **36**, 57–64 (1930). DOI: 10.1103/PhysRev.36.57.

[501] Parr, R.G. & Pearson, R.G.: "Absolute hardness: companion parameter to absolute electronegativity". Journal of the American Chemical Society **105**, 26, 7512–7516 (1983). DOI: 10.1021/ja00364a005.

[502] Cordero, B.; Gómez, V. *et al.*: "Covalent radii revisited". Dalton Trans. 2832–2838 (2008). DOI: 10.1039/B801115J.

[503] Ghosh, D.C. & Biswas, R.: "Theoretical Calculation of Absolute Radii of Atoms and Ions. Part 1. The Atomic Radii". International Journal of Molecular Sciences **3**, 2, 87–113 (2002). DOI: 10.3390/i3020087.

[504] Ghosh, D.C.: "A new scale of electronegativity based on absolute radii of atoms". Journal of Theoretical and Computational Chemistry **04**, 01, 21–33 (2005). DOI: 10.1142/S0219633605001556.

[505] Schwerdtfeger, P. & Nagle, J.K.: "2018 Table of static dipole polarizabilities of the neutral elements in the periodic table". Molecular Physics **117**, 9-12, 1200–1225 (2019). DOI: 10.1080/00268976.2018.1535143.

[506] Ku, H.T.; Ku, M.C. *et al.*: "Generalized Power Means and Interpolating Inequalities". Proceedings of the American Mathematical Society **127**, 1, 145–154 (1999). URL: http://www.jstor.org/stable/118925.

[507] Knopoff, L.: "Approximate Compressibility of Elements and Compounds". Physical Review **138**, A1445–A1447 (1965). DOI: 10.1103/PhysRev.138.A1445.

[508] Cohen, M.L.: "Calculation of bulk moduli of diamond and zinc-blende solids". Physical Review B **32**, 7988–7991 (1985). DOI: 10.1103/PhysRevB.32.7988.

[509] Liu, X.; Wang, H. *et al.*: "A simple bulk modulus model for crystal materials based on the bond valence model". Phys. Chem. Chem. Phys. **19**, 22177–22189 (2017). DOI: 10.1039/C7CP03739B.

[510] Lam, P.K.; Cohen, M.L. *et al.*: "Analytic relation between bulk moduli and lattice constants". Physical Review B **35**, 9190–9194 (1987). DOI: 10.1103/PhysRevB.35.9190.

[511] Sekar, M.; Chandra Shekar, N. *et al.*: "Structural stability of ultra-incompressible Mo2B: A combined experimental and theoretical study". Journal of Alloys and Compounds **654**, 554–560 (2016). DOI: 10.1016/j.jallcom.2015.09.128.

[512] Meredig, B.; Antono, E. *et al.*: "Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery". Molecular Systems Design & Engineering **3**, 819–825 (2018). DOI: 10.1039/C8ME00012C.

[513] Yuan, J.; Stanev, V. *et al.*: "Recent advances in high-throughput superconductivity research". Superconductor Science and Technology **32**, 12, 123001 (2019). DOI: 10.1088/1361-6668/ab51b1.

[514] Udrescu, S.M. & Tegmark, M.: "AI Feynman: A physics-inspired method for symbolic regression". Science Advances **6** (2020). DOI: 10.1126/sciadv.aay2631.

[515] Fan, J.; Samworth, R. *et al.*: "Ultrahigh Dimensional Feature Selection: Beyond The Linear Model". J. Mach. Learn. Res. **10**, 2013–2038 (2009). URL: http://www.jmlr.org/papers/volume10/fan09a/fan09a.pdf.

[516] Breen, P.: "Algorithms for sparse approximation". School of Mathematics, University of Edinburgh, Year **4** (2009). URL: https://www.zess.uni-siegen.de/myoffer/files/file_5_52170211052012.pdf.

Publications

Some chapters of this thesis are partially taken from the publication:

- Regler, B.; Scheffler, M.; Ghiringhelli, L.M.: “TCMI: a non-parametric mutual-dependence estimator for multivariate continuous distributions” (under review).

  ARXIV: arXiv:2001.11212 [stat.ML]
  Interactive notebook: https://nomad-lab.eu/services/aitoolkit

  Addendum: After submission of the dissertation, the scientific paper entitled “TCMI: a non-parametric mutual-dependence estimator for multivariate continuous distributions” was accepted and published in the journal Data Mining and Knowledge Discovery, to be found under the DOI (10.1007/s10618-022-00847-y). Parts of Chapter 3 are loosely based on this paper. However, the paper is a stand-alone publication and also presents applications and experiments not covered in the dissertation.

Further publications and an interactive tutorial based on this thesis are in preparation. During my PhD, I also worked on the following projects:

- Regler, B.; Ghiringhelli, L.M.; Scheffler, M.: Periodic Table (Analytics Toolkit). An interactive notebook for visualizing the atomic properties of a data set in the periodic table. This notebook was created during the 2nd NoMaD Analytics Hackathon (2017). URL: https://gitlab.rzg.mpg.de/nomad-lab/Hackaton-PTable.
- Regler, B.; Sastre, A.; Mohamed, F.; Ghiringhelli, L.M.; Scheffler, M.: NoMaD Query GUI (Analytics Toolkit). Front-end application for searching materials data in the Novel Materials Discovery (NoMaD) archive (2017). URL: https://gitlab.rzg.mpg.de/nomad-lab/notebook-nomad-query.

Acknowledgements

This thesis would not have been possible without the invaluable guidance and support of my thesis supervisors Dr. Luca M. Ghiringhelli and Prof. Dr. Matthias Scheffler. It was a great pleasure to work with them and be part of the group researching emerging topics in materials science. I thank both of them for their training and discussions throughout these years. I would also like to thank Prof. Dr. Jens Eisert for agreeing to be my co-supervisor and reviewer for my thesis.

I am grateful to all who have accompanied me along the way. My special thanks goes to Emre Ahmetcik, Danilo S. Brambila, and Andreas Leitherer, with whom I had fascinating scientific discussions. My heartfelt thanks goes to all the people at the institute with whom I shared lovely years.

Working at the thesis was not only an exciting time, but also an enriching experience. My research benefited from an initial collaboration with the May Planck Institute for Informatics in Saarbrücken. I am grateful to Dr. Mario Boley, Panagiotis Mandros, and Prof. Dr. Jilles Vreeken for teaching me the information-theoretic foundations and for inspiring discussions, especially in the early stages of my PhD.

I owe the deepest gratitude to my family for their constant unconditional support and encouragement during these many years of study. I am indebted to my wife Tanja, and my daughter Malea for their love, support, and understanding even in the difficult and challenging times and for all the time I spent working and not with them.

Benjamin Regler
Berlin, October 2021

This document, containing 54 figures, 28 tables, 4 listings, and 516 references (possibly with multiple citations), was typeset with the LaTeX documentation system using Bitstream Charter, Computer Modern, and TeX Gyre Heros type faces. Figures and illustrations were created with Matplotlib and Ti*k*Z.