Liu, Xiangyue

# Machine learning and reaction dynamics: From spectroscopic constants of diatomic molecules to buffer gas chemistry

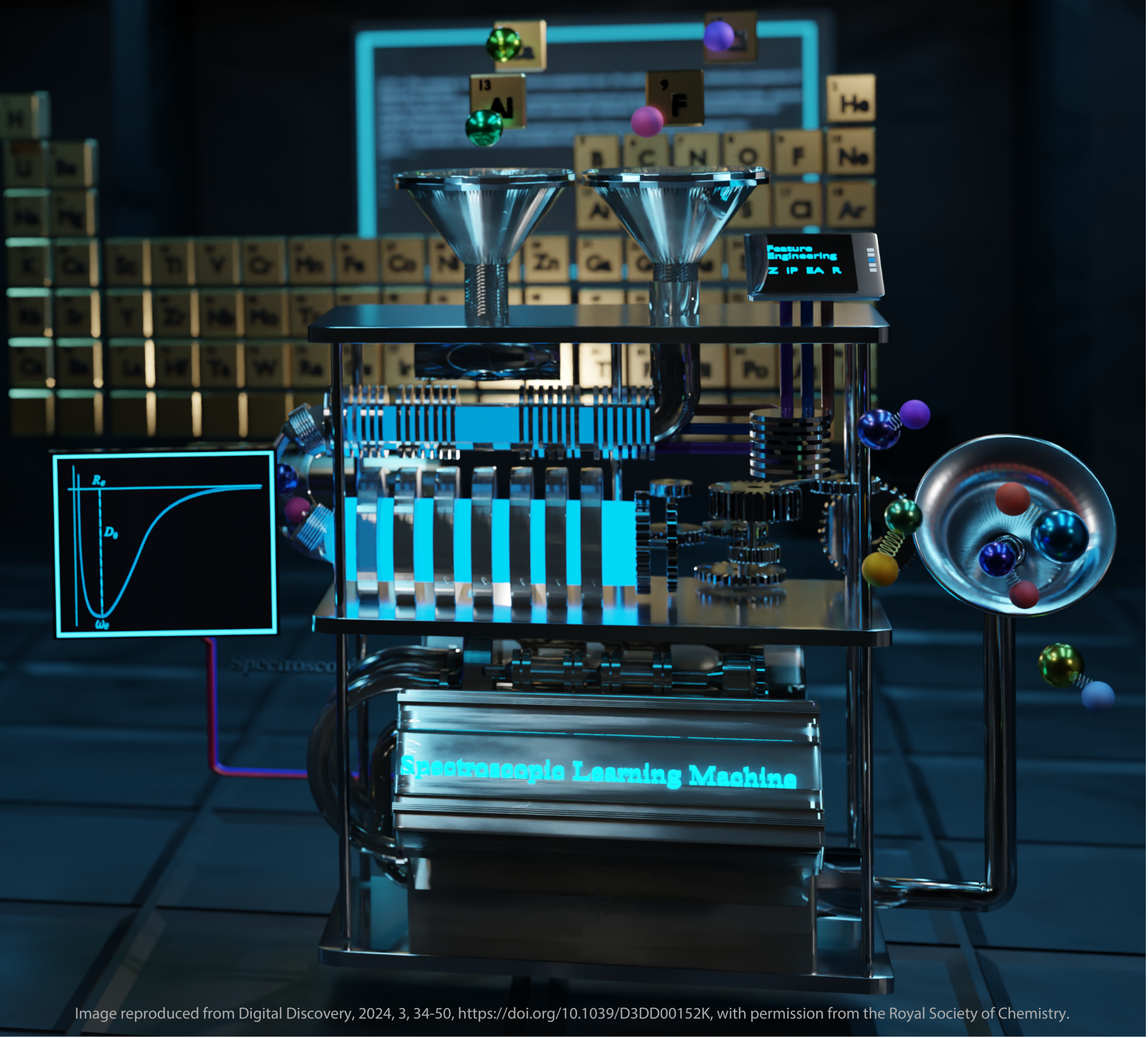


Image reproduced from Digital Discovery, 2024, 3, 34-50, https://doi.org/10.1039/D3DD00152K, with permission from the Royal Society of Chemistry.

# Machine learning and reaction dynamics: From spectroscopic constants of diatomic molecules to buffer gas chemistry

**Dissertation**

zur Erlangung des Grades eines Doctor rerum naturalium

(Dr. rer. nat.)

am Fachbereich Physik der Freien Universität Berlin

vorgelegt von

Xiangyue Liu

Berlin 2023

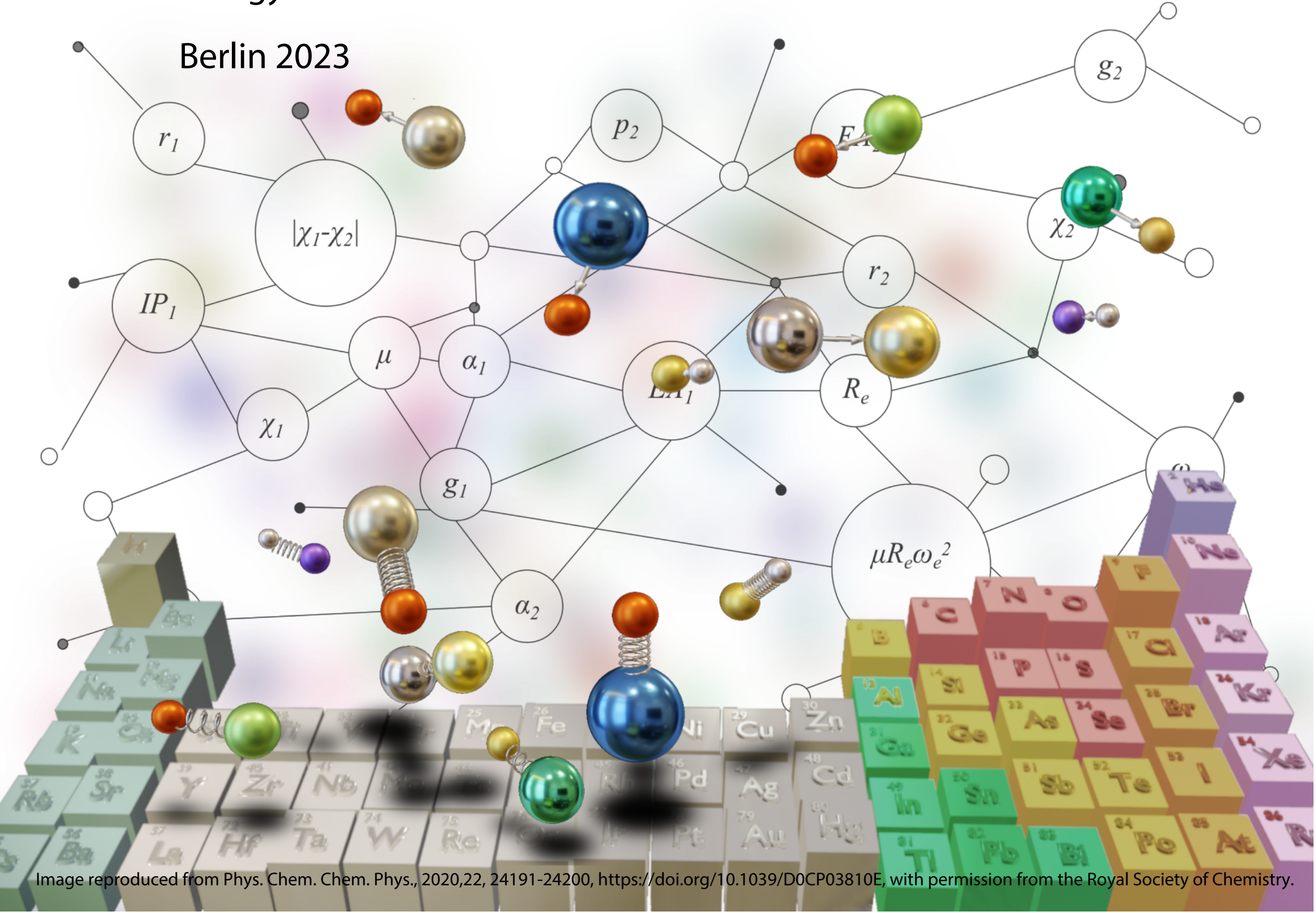


Image reproduced from Phys. Chem. Chem. Phys., 2020,22, 24191-24200, https://doi.org/10.1039/D0CP03810E, with permission from the Royal Society of Chemistry.

Erstgutachter/in: Prof. Dr. Gerard Meijer

Zweitgutachter/in: Prof. Dr. Piet Brouwer

Tag der Disputation: 18.01.2024

# Summary

This thesis explores the spectroscopic properties and chemistry of diatomic molecules, which hold significant promise for applications in areas like quantum information and ultracold chemistry.

Firstly, the Diatomic Molecular Spectroscopy Database, accessible through a dynamic website, has been implemented. This database predominantly consolidates spectroscopic information while enabling the computation and visualization of Franck-Condon factors, and is adaptable for user contributions. Based on this database, machine learning models have been built to effectively reveal relationships among spectroscopic constants, with input features based on constituent atoms' group and period. Similarly, a comprehensive dataset of contemporary experimental electric dipole moments has been created. Utilizing this dataset, it has been shown that a machine learning model can accurately predict dipole moments using spectroscopic constants.

The availability of precise spectroscopic data allows for a rigorous assessment of advanced quantum chemistry methods. Specifically, we investigated the accuracy of coupled-cluster with single, double, and perturbative triple excitations [CCSD(T)] in predicting electric dipole moments when combined with different basis sets. Additionally, the hyperfine constants for the $a^3\Pi$ state of aluminum monofluoride (AlF) have been computed and compared to experimental values. Our study underscores the significance of a thorough evaluation encompassing both experimental and theoretical methodologies.

AlF and calcium monofluoride (CaF), among other metal monofluorides, have emerged as highly promising options for experiments involving laser cooling and trapping of cold molecules. We have compared the efficiency of different fluorine-donor molecules producing AlF and CaF through metal atom ablation in a buffer gas cell. Additionally, we present an efficient machine learning method for fitting the potential energy surface of AlF-AlF system, trained on relevant configurations from molecular dynamics simulations at the CCSD(T) level.

# Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit der Erforschung der spektroskopischen Eigenschaften und Chemie von zweiatomigen Molekülen, die für Anwendungen in Bereichen wie Quanteninformation und ultrakalte Chemie vielversprechend sind.

Zunächst wurde eine Datenbank zur Spektroskopie zweiatomiger Moleküle implementiert, die über eine dynamische Website zugänglich ist. Diese Datenbank konsolidiert hauptsächlich spektroskopische Informationen und ermöglicht gleichzeitig die Berechnung und Visualisierung des Franck-Condon-Faktors und ist für Benutzerbeiträge anpassbar.

Basierend auf dieser Datenbank wurden Machine-Learning-Modelle entwickelt, um die Beziehungen zwischen den spektroskopischen Konstanten zweiatomiger Moleküle aufzudecken, wobei die Eingabemerkmale auf der Gruppe und der Periode der beteiligten Atome basieren. Ebenso wurde ein umfassender Datensatz mit aktuellen experimentellen elektrischen Dipolmomenten erstellt. Unter Verwendung dieses Datensatzes wurde gezeigt, dass ein Machine-Learning-Modell Dipolmomente genau vorhersagen kann, indem es spektroskopische Konstanten verwendet.

Die Verfügbarkeit präziser spektroskopischer Daten ermöglicht eine gründliche Bewertung fortschrittlicher quantenchemischer Methoden. Insbesondere untersuchten wir die Genauigkeit des gekoppelten Clusteransatzes mit einfachen, doppelten und perturbativen dreifachen Anregungen (CCSD(T)) bei der Vorhersage von elektrischen Dipolmomenten in Kombination mit verschiedenen Basissätzen. Zusätzlich wurden die Hyperfeinstrukturkonstanten für den $a^3\Pi$-Zustand von Aluminiummonofluorid (AlF) berechnet und mit experimentellen Werten verglichen. Unsere Studie betont die Bedeutung einer gründlichen Bewertung, die sowohl experimentelle als auch theoretische Methoden umfasst.

AlF und Calciummonofluorid (CaF) haben sich neben anderen metallischen Monofluoriden als äußerst vielversprechende Optionen für Experimente zur Laserkühlung und zum Einfangen von kalten Molekülen herausgestellt. Wir haben die Effizienz verschiedener Fluor-Donor-Moleküle verglichen, die AlF und CaF durch Metallatom-ablation in einer Puffergaszelle erzeugen. Darüber hinaus präsentieren wir eine effiziente Methode des maschinellen Lernverfahrens zur Anpassung der potenziellen Energiefläche des AlF-AlF-Systems, das auf relevanten Konfigurationen aus Molekulardynamik-Simulationen auf dem CCSD(T)-Niveau trainiert wurde.

# Acknowledgement

I would like to extend my heartfelt gratitude to Jesús for his unwavering support in both my academic journey and personal life. His profound knowledge of theoretical physics and boundless enthusiasm for scientific research are truly admirable, and his receptiveness to innovative problem-solving approaches has been invaluable. He always provided unreserved support to all my naive questions. At the same time, he supports me not only as a doctoral supervisor but also as a person of integrity who can always encourage me when I encounter difficulties. I would especially like to thank him for his tolerance and support for my family. I can hardly put it into words, but I feel incredibly lucky to be his student because I know he backs me up no matter what.

At the same time, I am very grateful to Gerard. Although he has a lot to do as a director, he always actively participates in group meetings and patiently answers various questions. As a student, we can communicate with him without any barriers and look forward to his help. Although he mainly conducts experimental work, he has an open mind for our theoretical work and supports us straightforwardly.

The invaluable contributions of the experimental team at the AlF project have been instrumental to my research. I extend my deepest thanks to Stefan Truppe, Nicole, Sidney, Max, Simon, Sebastian, and Russell for their graciousness in sharing experimental insights and knowledge. Additionally, I am grateful to Christian Schewe for his invaluable guidance.

Engaging in discussions with André has been a privilege. His breadth of knowledge and sense of humor are truly impressive. I extend my thanks to Uwe, Wieland, Sandra, Gert, Bernd, and Henrik for their friendly exchanges, which have enriched my academic experience. I also would like to thank Stefan Willitsch for his insightful discussions and his patience in our collaboration.

The Molecular Physics department has been an exceptional environment to work in, characterized by a welcoming, inclusive, and supportive atmosphere that exceeds all expectations. I extend my thanks to Ms. Manuela Misch, Ms. Karin Grassow, and Ms. Evelyn Prohn for their contributions to fostering this conducive atmosphere and for their assistance in daily routines.

To my dear friends Nicole and Nadia, I am grateful for the enriching conversations we have had about scientific research and life. Your friendship has been a source of great joy and inspiration. I am also grateful to Marjan for her friendship and for providing me with invaluable suggestions.

I would also like to extend my gratitude to Prof. Piet Brouwer and Prof. Felix von Oppen for their friendly help with my registration and for their wonderful master's courses.

Lastly, my sincere gratitude goes to my family, especially Weiqi, for his unwavering support in both my research and everyday life. My mother's dedication to chemical experimentation is a huge inspiration, and my father's enthusiasm for various hobbies fuels my curiosity. My grandfather's love for natural science instilled in me a deep appreciation for the wonders of the natural world. Thanks also to Elaine for giving me the space to focus on my work and for her exceptional talent in bringing laughter to my days.

# CONTENTS

II Chemistry of diatomic fluorides

# 1

# INTRODUCTION

## 1.1 Importance of accurate spectroscopic information for diatomic molecules

Accurate and comprehensive diatomic molecular spectroscopic data have long been vital in a wide variety of applications. Specifically, this information serves as a crucial reference for advancing quantum chemistry methods. Diatomic molecules, being considered the "simplest" molecules, offer valuable insights into the chemical bonds between different elements, potentially shedding light on the properties of polyatomic molecules and materials. Consequently, precise spectroscopic data for diatomic molecules is highly desirable and has been used for benchmarking.

In particular, equilibrium molecular constants, such as the equilibrium internuclear distance, bond dissociation energy, ionization potential, harmonic/anharmonic vibrational frequency, etc., can be directly compared to results from quantum chemistry. These properties play a crucial role in developing quantum chemistry in various ways. For instance, energetic properties, such as bond dissociation energy and electron affinity, can be employed to assess the extent to which electron correlation energy is captured by a specific level of the electron correlation method [1, 2, 3, 4]. Additionally, comparing computed diatomic equilibrium internuclear distances with experimental values helps in understanding the convergence patterns of computed geometric properties as electron correlation methods and basis set sizes are refined. These behaviors also hold significance in the advancement of basis set development [5, 6, 7, 8].

Certainly, it is also viable to benchmark quantum chemistry methods by contrasting computed spectroscopic constants with those from established theoretical references, such as the coupled cluster with single, double, and perturbative triple excitations [CCSD(T)]. Purely theoretical reference datasets have gained notable traction in contemporary quantum chemistry advancements. Developers now prefer to scrutinize their models against a substantial dataset, generated theoretically at the same level of theory and utilizing consistent basis sets [9], rather than relying on a limited set of molecules [10]. This approach is favored as it allows for better control over reference errors. Meanwhile, the creation of such reference datasets has become significantly more accessible compared to past decades, courtesy of the advancements in supercomputing technology. As a result, it has become customary, especially in the development of density functionals [11, 12, 9].

However, as elucidated in the insightful and thought-provoking discussions in [13], the limited inclusion of experimental data in the presently favored reference datasets does not indicate the insignificance of experimental references. Instead, it stems from the pragmatic challenges of aligning theoretical predictions with experimental outcomes. Indeed, there are instances where theoreticians may need more expertise to interpret experimental measurements accurately, including understanding the associated errors and determining the appropriate comparison method [13]. Additionally, they often face constraints in terms of time availability for extensive literature searches [13]. This can lead to challenges and potential misinterpretations during the benchmarking, especially when experimental data has not been appropriately preprocessed. A pertinent example can be found in the discussions surrounding the accuracy of coupled cluster methods in predicting dissociation energies of 3d transition metals in [1, 14, 4, 15].

Meanwhile, as highlighted in [16], obtaining updated experimental data, which often offer higher precision and lower uncertainties, can be challenging. Remarkably, even in the present day, a significant number of researchers heavily lean on the enduringly successful and comprehensive work of Huber and Herzberg [17], published several decades ago, to obtain the equilibrium spectroscopic constants [18], due to the scarcity of alternative datasets. In the case of properties tied to energy derivatives, such as dipole moments, the absence of an accurate and all-encompassing dataset necessitates benchmarking efforts to rely on theoretical references [19].

With the hope of addressing the aforementioned issues to some extent, we have introduced the diatomic molecular spectroscopy database [20]. It primarily incorporates equilibrium spectroscopic constants sourced mainly from Huber and Herzberg [17]. Importantly, the data has been organized for easy access and download via a user-friendly website. Additionally, this database is designed to accommodate user contributions, making it adaptable to future updates. Separately, we have curated an updated compilation of experimental ground-state electric dipole moments [21]. Based on this dataset, we have benchmarked the accuracy of the CCSD(T) method in predicting electric dipole moments. We employed a range of basis sets, considering the inclusion or exclusion of diffuse functions, and analyzed their performance. As shown in the corresponding chapters, our results suggest the paramount significance of a meticulous assessment encompassing experimental and theoretical methodologies.

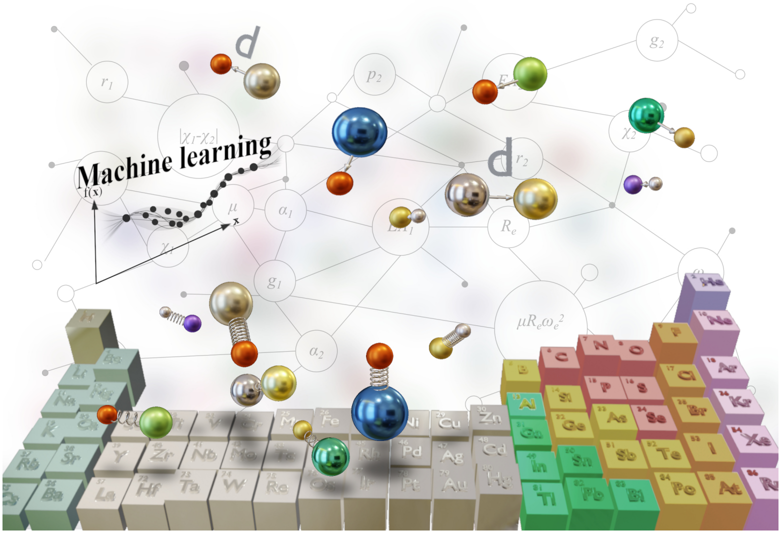


Figure 1: The results in this thesis indicate that with the incorporation of suitable atomic and molecular features, machine learning methods can proficiently forecast the spectroscopic constants and dipole moments of diatomic molecules. These selected features encompass a blend of molecular properties, contributing to enhancing our understanding of the fundamental essence of spectroscopic constants.

On the other hand, machine learning techniques provide a powerful tool for uncovering relationships between diverse properties based on provided datasets. In this context, we have utilized the previously introduced datasets to apply machine learning methods, allowing us to discern the connections between spectroscopic constants. Through thoughtful engineering of input features, the resulting machine learning models offer interpretability. Remarkably, they demonstrate the ability to accurately predict spectroscopic constants without the necessity for labor-intensive quantum chemistry calculations.

## 1.2 Diatomic fluorides relevant for laser cooling

Molecules that have been laser-cooled exhibit distinct properties and find applications in various fields, such as precision measurements in fundamental physics [22, 23] and the development of novel platforms for quantum information processing. They also open new avenues for studying ultracold molecular collisions and chemical reactions through precise control over their initial quantum states.

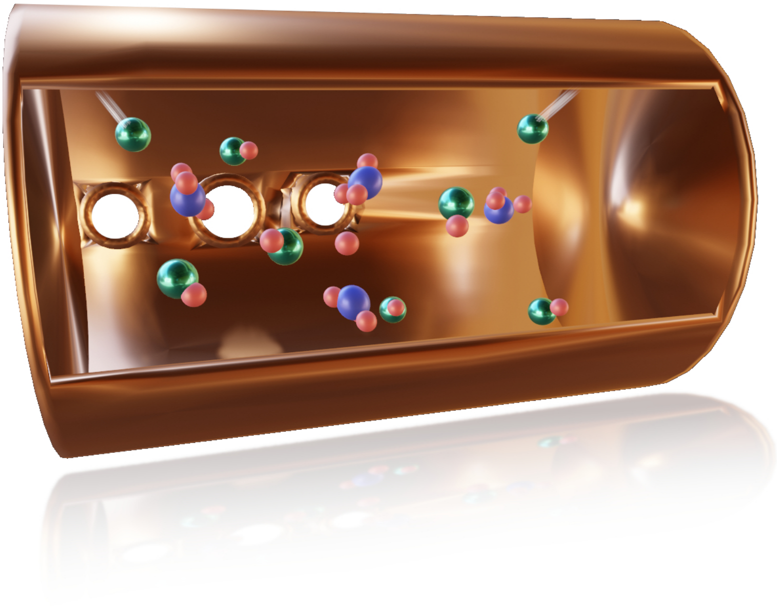


Figure 2: The efficient production of AlF and CaF molecules has been achieved in buffer gas cells. This thesis delves into the chemistry of reactions responsible for generating AlF and CaF molecules, alongside an investigation of the properties exhibited by the resulting AlF-AlF dimers.

The intricate internal structure of molecules poses a challenge to trapping. The feasibility of directly laser-cooling and trapping a molecule depends on its electric, vibrational, and rotational characteristics, as these factors dictate the efficiency of photon scattering in the cycling process, and consequently, the complexity of the cooling laser system. Notably, having a nearly-diagonal Franck-Condon matrix and corresponding vibrational branching is highly desirable. In such cases, the equilibrium internuclear distances in the ground and excited electronic states are very close, allowing for efficient direct excitation of electrons to the excited state during optical cycling. Meanwhile, the diagonal Franck-Condon factors reduce the loss channels via vibrational transitions [24]. Additionally, molecules with simple hyperfine structures and rotational splittings are preferred, although technical solutions can address these complexities [24]. The absence of intermediate electronic levels is also preferred, as it simplifies the cooling scheme.

The first successful trapping of a diatomic molecule was achieved with strontium monofluoride (SrF) using magneto-optical traps, capitalizing on its $A^2\Pi_{1/2}$-$X^2\Sigma^+$ electronic transition in the cycling scheme [25]. Molecules like CaF [26], MgF [27], YO [28], AlF [29, 30], and AlCl [31] also show promise for laser cooling. The potential for laser-cooling polyatomic molecules has been explored in several molecules, including CaOH and $CaOCH_3$ [24]. In this thesis, the focus of investigation lies on the chemistry of diatomic fluorides, specifically AlF and CaF.

## 1.3 Overview of the thesis

This thesis comprises two main segments. The first segment centers on the spectroscopic constants of diatomic molecules. In **Chapter 2**, we introduce the implementation and functionality of the diatomic molecular spectroscopy database, characterized by a dynamic structure with a user-friendly website interface. **Chapter 3** leverages the spectroscopic constants cataloged in the database to investigate the relationship between these constants and the "periodicity" of the constituent atoms. Specifically, we focus on the equilibrium internuclear distance in both ground and first-excited states, as well as the harmonic vibrational frequency. Additionally, we explore the relationship between the ground-state dissociation energy and other spectroscopic constants. To discover these relationships, we employ machine learning (ML) regression methods, allowing for the automatic modeling of the connection between input and output variables in a versatile manner in the presence of reference datasets. **Chapter 4** constructs a comprehensive dataset of ground-state electric dipole moment, and extends the ML approach to the dipole moment, aiming to ascertain its potential correlation with other spectroscopic constants. We delve into not only those molecules exhibiting a clear relationship between their dipole moments and other spectroscopic constants but also those that deviate from this pattern. Next, we examine the accuracy of *ab initio* quantum chemistry methods to reproduce experimentally measured spectroscopic constants. **Chapter 5** employs the spectroscopic information compiled in Chapters 2 and 4 to benchmark, for selected molecules, the accuracy of the coupled cluster with single, double, and perturbative triple excitations [CCSD(T)] method in predicting electric dipole moments. This method has served as a benchmark standard for the implementation of other electronic structure theory approaches. Finally, **Chapter 6** involves the computation of hyperfine constants for AlF using quantum chemistry methods, which are then compared with experimental measurements.

The second part of this thesis is dedicated to the study of the chemistry of diatomic fluorides, with a specific focus on AlF and CaF. **Chapter 7** directs attention towards the production of AlF and CaF in buffer gas sources. Through *ab initio* molecular dynamics simulations, we compare the efficiency of producing AlF and CaF from different fluoride-donor molecules. Additionally, once AlF molecules are formed, the formation of long-lived complexes is deemed unfavorable as it compromises the utility of ultracold molecules in many applications, leading to molecular loss. Therefore, in **Chapter 8**, we develop an accurate hybrid *ab initio*-machine learning potential energy surface for the AlF-AlF system, a crucial advancement for investigating the formation of AlF-AlF complexes, playing a major role in the stability of AlF molecules in the ultracold regime.

Finally, **Chapter 9** concludes this thesis.

Part I

# THE SPECTROSCOPIC CONSTANTS OF DIATOMIC MOLECULES

# 2

# THE DIATOMIC MOLECULAR SPECTROSCOPY DATABASE

## 2.1 Introduction

The ability to control internal and external degrees of freedom of molecules is the main driving force in atomic, molecular, and optical physics due to its applications in quantum information sciences [32], cold and ultracold chemistry [33, 34], coherent control, and the search of new particles and fields beyond the standard model of particle physics [35, 36]. Most of these applications require a cold sample of molecules in a well-defined quantum state. In this regard, laser cooling is the most prominent tool to bring down an ensemble of molecules to the cold and ultracold regime. However, laser cooling can only be efficiently applied to molecules with nearly vertical Franck-Condon factors (FCFs). These factors are contingent on the spectroscopic properties of both the ground and excited electronic states. Consequently, building a comprehensive database encompassing spectroscopic constants and FCFs for various states will aid in the identification of prime candidates for molecular laser cooling.

Several valuable databases have been made available to the public through various websites, including well-known resources like HITRAN [37], ExoMol [38], NIST Chemistry WebBook [39], and OSDB [40], among others [41, 42, 43, 44, 38, 45, 46], as reviewed in [16]. These platforms primarily deliver spectral data for molecules, tailored explicitly for astrophysics and atmospheric physics applications. For instance, platforms like HITRAN [37] and ExoMol offer comprehensive rovibronic line lists for tens of diatomic molecules. Additionally, DiRef [42] offers downloadable reference papers for diatomic molecules, primarily sourced from 1974 to 2000. However, most existing databases do not typically include information on the spectroscopic constants of diatomic molecules. Only the NIST Chemistry WebBook offers such data, although not always retrievable in convenient, readable formats like XML, JSON (JavaScript Object Notation), or comma-separated values (CSV).

In this work, we implement a user-friendly database linked to an interactive website that collates spectroscopic constants for polar diatomic molecules, encompassing their ground and first excited electronic states. Additionally, we provide calculations of FCFs, assuming a Morse potential shape for all the relevant states [1].

Our database relies on the spectroscopic constants for diatomic molecules sourced from the authoritative work of Huber and

[1] This chapter is written based on reference [20]: Xiangyue Liu, Stefan Truppe, Gerard Meijer, and Jesús Pérez-Ríos. The diatomic molecular spectroscopy database. *Journal of Cheminformatics*, 12(1):1–8, 2020, https://doi.org/10.1186/s13321-020-00433-8.

Herzberg [17], the most comprehensive compendium of molecular spectroscopy data for diatomic molecules. The website is interactive, enabling users to upload new data, subject to approval by the web administrators. Consequently, the database is dynamic and has the potential for continuous expansion as new spectroscopic measurements become available.

## 2.2 Functionality of the database

Our database provides access to the spectroscopic constants of diatomic molecules' ground and first excited states. These constants can be easily retrieved in user-friendly formats through the website. The website is built on the Linux, Apache, MySQL, and PHP (LAMP) web service stacks. Specifically, we use Linux as the operating system, Apache as the HTTP Server, MySQL as the database management system, and PHP as the programming language for the web. The website offers several key functionalities, such as querying and contributing to the spectroscopic data and calculating the FCFs. PHP and MySQL are responsible for data querying and editing in the database and user information management. Dynamic generation of webpages is accomplished using PHP in conjunction with HTML/CSS and JavaScript. Notably, JavaScript is crucial in swiftly calculating and visualizing the FCFs. Additionally, users can register and upload new data to the database.

It is worth emphasizing that MySQL provides a robust yet user-friendly database management system known for its high efficiency in handling various database operations. Customized, dynamic, data-driven websites are enabled when seamlessly integrated with PHP, facilitating database management, including creation, access, and operation. Additionally, both MySQL and PHP offer robust security features, ensuring the integrity and protection of data when implemented correctly.

This project is licensed under the Free-Libre/Open Source Software (FLOSS) license Apache License 2.0, enabling unrestricted and open access to the source codes and facilitating efficient collaboration in software maintenance.

### 2.2.1 Database construction

The spectroscopic constants of diatomic molecules are sourced from Huber and Herzberg [17] and stored in a MySQL database. The molecules in the database, along with the states of these molecules, are indexed with non-null integer numbers: "idMol" and "idAll_in", respectively, within the "molecule_data" table. "idAll_in" serves as the primary key for the table. Each state is represented by a symbol in Latex format, stored as a string in the database. In addition, the database retains the reduced mass of each molecule as a non-null floating-point value. This incorporation of reduced mass serves to differentiate isotopes within the constituent atoms.

The spectroscopic constants comprise several key parameters for each molecule and state, including:

- Minimum electronic energy ($T_e$) measured in cm$^{-1}$.

- Harmonic frequency ($\omega_e$) measured in $\text{cm}^{-1}$.
- First anharmonic correction ($\omega_e x_e$) measured in $\text{cm}^{-1}$.
- Equilibrium rotational constant ($B_e$) measured in $\text{cm}^{-1}$.
- Anharmonic correction to the rotational constant ($\alpha_e$) measured in $\text{cm}^{-1}$.
- Centrifugal distortion constant ($D_e$) measured in $\text{cm}^{-1}$.
- Binding energy ($D_0$) measured in eV.
- Equilibrium internuclear distance ($R_e$) measured in angstroms (Å).
- Ionization potential ($IP$) measured in eV.

"NULL" if the data is unavailable. Additionally, the database records the reference from which the data are obtained, along with the reference date. This meticulous record-keeping ensures that users can compare the dates of different measurements for the same molecule and effortlessly trace back to the original source. Moreover, each record in the database includes contribution information, including the contributor's identification ("id_user") and the date of data input into the database.

### 2.2.2 Search in the database

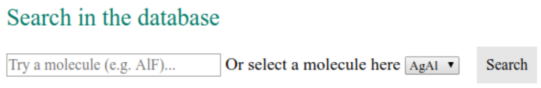
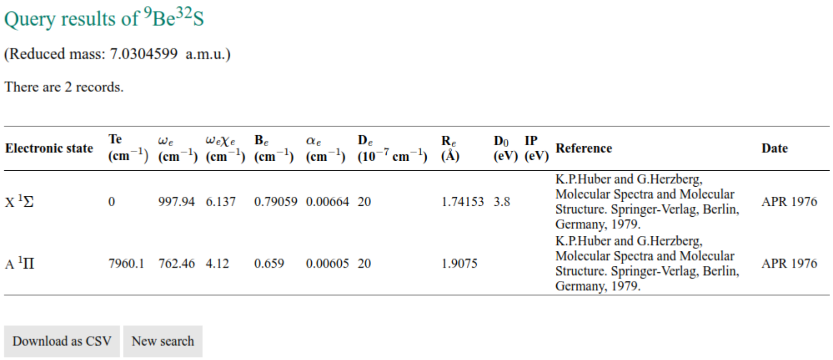

Search in the database

Figure 3: General search menu. The user needs to introduce the chemical formula of a molecule. Figure reproduced from ref. [20].

The website allows users to retrieve spectroscopic constants from the database by searching with the chemical formula of molecules, as shown in Fig. 3. An example of a search results is displayed in a table that can be conveniently downloaded in CSV format as displayed in Fig. 4), making it easy for users to process the data.



Query results of $^{9}Be^{32}S$

(Reduced mass: 7.0304599 a.m.u.)

There are 2 records.

| Electronic state | Te ($cm^{-1}$) | $\omega_e$ ($cm^{-1}$) | $\omega_e\chi_e$ ($cm^{-1}$) | $B_e$ ($cm^{-1}$) | $\alpha_e$ ($cm^{-1}$) | $D_e$ ($10^{-7}$ $cm^{-1}$) | $R_e$ (Å) | $D_0$ (eV) | IP (eV) | Reference | Date |
|---|---|---|---|---|---|---|---|---|---|---|---|
| X $^1\Sigma$ | 0 | 997.94 | 6.137 | 0.79059 | 0.00664 | 20 | 1.74153 | 3.8 |  | K.P.Huber and G.Herzberg, Molecular Spectra and Molecular Structure. Springer-Verlag, Berlin, Germany, 1979. | APR 1976 |
| A $^1\Pi$ | 7960.1 | 762.46 | 4.12 | 0.659 | 0.00605 | 20 | 1.9075 |  |  | K.P.Huber and G.Herzberg, Molecular Spectra and Molecular Structure. Springer-Verlag, Berlin, Germany, 1979. | APR 1976 |

Figure 4: Example output after a search. The results are presented in a table and can be downloaded to a CSV file. Figure reproduced from ref. [20].

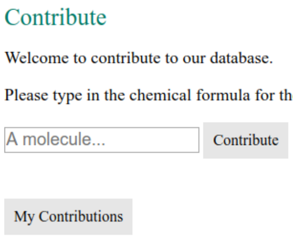
Furthermore, as introduced in Sec. 2.3.2, the website enables the calculation of FCFs for selected states of the molecules using the spectroscopic constants of the ground and excited states. This feature provides users with valuable insights into the favored vibrational transitions between different electronic states of the molecules, enhancing their understanding of molecular spectroscopy.

### 2.2.3 User contributions to the database

Following an initial search for a given molecule to access its existing spectroscopic information, users have the option to contribute new information to the database, as illustrated in Fig. 5. The user can upload spectroscopic data via an HTML form (Fig. 6), wherein the reduced mass and reference information must be provided as non-null fields. The electronic states are expected to be uploaded in Latex format, for example, X $^1\Sigma^+$.

After submission (Fig. 7), contributions are inserted into the database only after receiving authorization from web managers (Fig.8). In case of rejection by the manager, the contributor will be notified of the reason via email.

Additionally, users have the option to check their own contributions to the database, ensuring transparency and accountability throughout the contribution process.

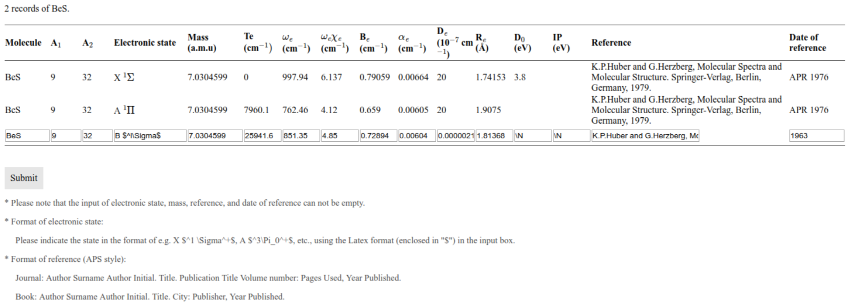


Contribute

Welcome to contribute to our database.

Please type in the chemical formula for the molecule:

Figure 5: Users can search for a molecule before contributing, and check their contributions to the database.



2 records of BeS.

| Molecule | $A_1$ | $A_2$ | Electronic state | Mass (a.m.u) | Te ($cm^{-1}$) | $\omega_e$ ($cm^{-1}$) | $\omega_e\chi_e$ ($cm^{-1}$) | $B_e$ ($cm^{-1}$) | $\alpha_e$ ($cm^{-1}$) | $D_e$ ($10^{-7}$ $cm^{-1}$) | $R_e$ (Å) | $D_0$ (eV) | IP (eV) | Reference | Date of reference |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BeS | 9 | 32 | X $^1\Sigma$ | 7.0304599 | 0 | 997.94 | 6.137 | 0.79059 | 0.00664 | 20 | 1.74153 | 3.8 | | K.P.Huber and G.Herzberg, Molecular Spectra and Molecular Structure. Springer-Verlag, Berlin, Germany, 1979. | APR 1976 |
| BeS | 9 | 32 | A $^1\Pi$ | 7.0304599 | 7960.1 | 762.46 | 4.12 | 0.659 | 0.00605 | 20 | 1.9075 | | | K.P.Huber and G.Herzberg, Molecular Spectra and Molecular Structure. Springer-Verlag, Berlin, Germany, 1979. | APR 1976 |
| BeS | 9 | 32 | B $^I\Sigma$ | 7.0304599 | 25941.6 | 851.35 | 4.85 | 0.72894 | 0.00604 | 0.0000021 | 1.81368 | \N | \N | K.P.Huber and G.Herzberg, Mc | 1963 |

* Please note that the input of electronic state, mass, reference, and date of reference can not be empty.

* Format of electronic state:

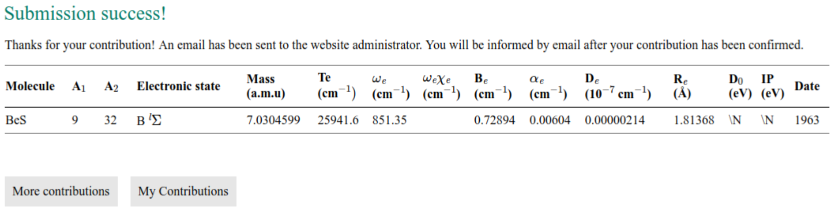
Please indicate the state in the format of e.g. X $^1 \Sigma^+$, A $^3\Pi_0^+$, etc., using the Latex format (enclosed in "$") in the input box.

* Format of reference (APS style):

Journal: Author Surname Author Initial. Title. Publication Title Volume number: Pages Used, Year Published.

Book: Author Surname Author Initial. Title. City: Publisher, Year Published.

Figure 6: Contribution menu. The user needs to fill an HTML table with new data. Figure reproduced from ref. [20].

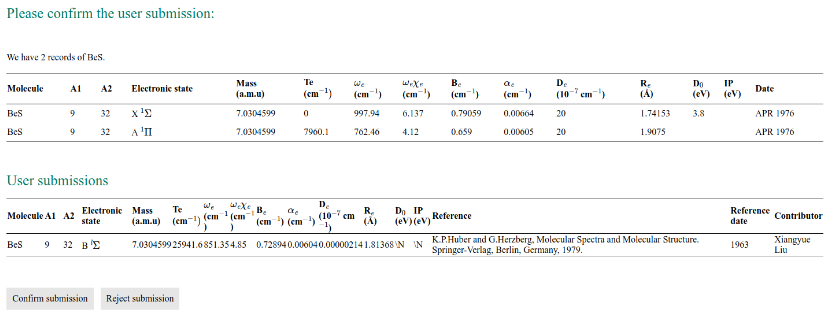


Submission success!

Thanks for your contribution! An email has been sent to the website administrator. You will be informed by email after your contribution has been confirmed.

| Molecule | $A_1$ | $A_2$ | Electronic state | Mass (a.m.u) | Te ($cm^{-1}$) | $\omega_e$ ($cm^{-1}$) | $\omega_e\chi_e$ ($cm^{-1}$) | $B_e$ ($cm^{-1}$) | $\alpha_e$ ($cm^{-1}$) | $D_e$ ($10^{-7}$ $cm^{-1}$) | $R_e$ (Å) | $D_0$ (eV) | IP (eV) | Date |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BeS | 9 | 32 | B $^I\Sigma$ | 7.0304599 | 25941.6 | 851.35 | | 0.72894 | 0.00604 | 0.00000214 | 1.81368 | \N | \N | 1963 |

Figure 7: After submission, the users can look at their contributions, while an email is sent to the web managers to authorize the contributions. Figure reproduced from ref. [20].



Please confirm the user submission:

We have 2 records of BeS.

| Molecule | A1 | A2 | Electronic state | Mass (a.m.u) | Te ($cm^{-1}$) | $\omega_e$ ($cm^{-1}$) | $\omega_e\chi_e$ ($cm^{-1}$) | $B_e$ ($cm^{-1}$) | $\alpha_e$ ($cm^{-1}$) | $D_e$ ($10^{-7}$ $cm^{-1}$) | $R_e$ (Å) | $D_0$ (eV) | IP (eV) | Date |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BeS | 9 | 32 | X $^1\Sigma$ | 7.0304599 | 0 | 997.94 | 6.137 | 0.79059 | 0.00664 | 20 | 1.74153 | 3.8 | | APR 1976 |
| BeS | 9 | 32 | A $^1\Pi$ | 7.0304599 | 7960.1 | 762.46 | 4.12 | 0.659 | 0.00605 | 20 | 1.9075 | | | APR 1976 |

User submissions

| Molecule | A1 | A2 | Electronic state | Mass (a.m.u) | Te ($cm^{-1}$) | $\omega_e$ ($cm^{-1}$) | $\omega_e\chi_e$ ($cm^{-1}$) | $B_e$ ($cm^{-1}$) | $\alpha_e$ ($cm^{-1}$) | $D_e$ ($10^{-7}$ $cm^{-1}$) | $R_e$ (Å) | $D_0$ (eV) | IP (eV) | Reference | Reference date | Contributor |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BeS | 9 | 32 | B $^l\Sigma$ | 7.0304599 | 25941.6 | 851.35 | 4.85 | 0.72894 | 0.00604 | 0.00000214 | 1.81368 | \N | \N | K.P.Huber and G.Herzberg, Molecular Spectra and Molecular Structure. Springer-Verlag, Berlin, Germany, 1979. | 1963 | Xiangyue Liu |

Figure 8: The web managers can compare the user contribution to the existing records in the database, and decide to confirm or reject the contribution. Figure reproduced from ref. [20].

## 2.2.4 The application programming interface (API)

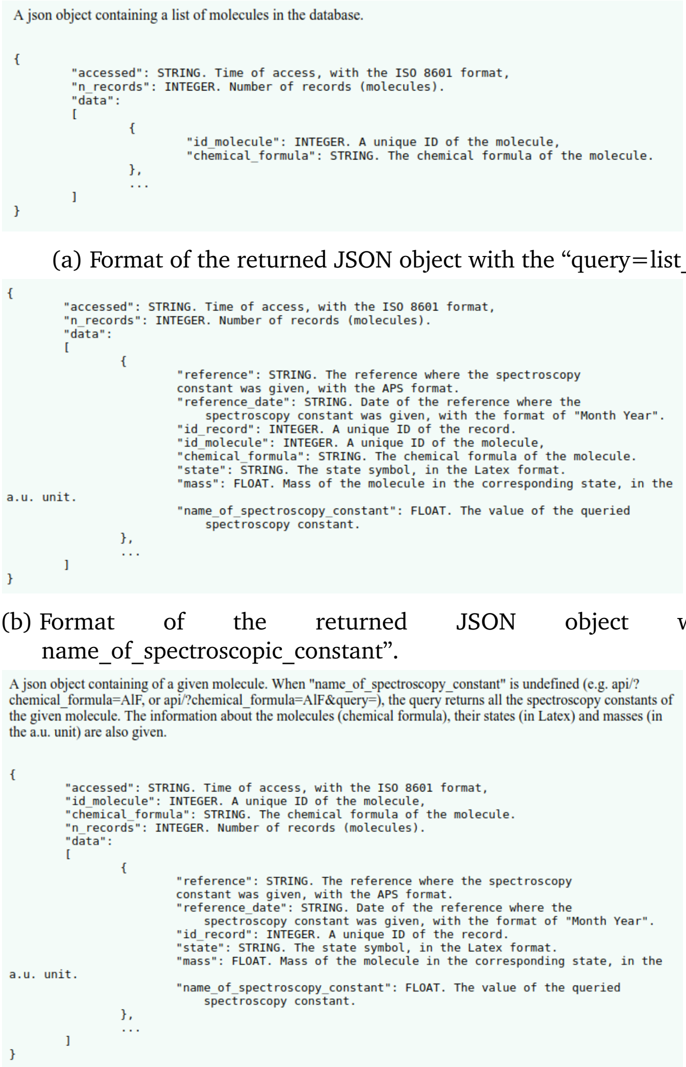
An Application Programming Interface (API) is provided to allow users to query the spectroscopic information of molecules in the database. The API returns JSON objects in the format shown in Fig. 9.

By accessing "/api/?query=list_molecules", users can obtain a list of all molecules present in the database. This list comprises an array of objects where each molecule is indexed by "id_molecule" and labeled with its chemical formula.

The API endpoint "/api/?query=name_of_spectroscopic_constant" provides the spectroscopic constants of the ground and excited states for all molecules that have data available for the specified constant.

For more detailed information about a specific molecule, users can search using "/api/? chemical_formula=chemical _formula&
query= name_of_spectroscopic_constant". For instance, entering "api/?chemical _formula=AlF&query=Be" will return a JSON object containing the $B_e$ value of AlF.

In case the "name_of_spectroscopic_constant" is undefined (e.g., "api/? chemical_formula=AlF", or "api/?chemical_formula=AlF&query="), the query will return all spectroscopic constants available for the specified molecule. The information about the molecules (chemical formula), their states (in Latex format), and masses (in atomic units) are also provided.



A json object containing a list of molecules in the database.

```
{
        "accessed": STRING. Time of access, with the ISO 8601 format,
        "n_records": INTEGER. Number of records (molecules).
        "data":
        [
                {
                        "id_molecule": INTEGER. A unique ID of the molecule,
                        "chemical_formula": STRING. The chemical formula of the molecule.
                },
                ...
        ]
}
```

(a) Format of the returned JSON object with the “query=list_molecules”.

```
{
        "accessed": STRING. Time of access, with the ISO 8601 format,
        "n_records": INTEGER. Number of records (molecules).
        "data":
        [
                {
                        "reference": STRING. The reference where the spectroscopy
                        constant was given, with the APS format.
                        "reference_date": STRING. Date of the reference where the
                            spectroscopy constant was given, with the format of "Month Year".
                        "id_record": INTEGER. A unique ID of the record.
                        "id_molecule": INTEGER. A unique ID of the molecule,
                        "chemical_formula": STRING. The chemical formula of the molecule.
                        "state": STRING. The state symbol, in the Latex format.
                        "mass": FLOAT. Mass of the molecule in the corresponding state, in the
a.u. unit.
                        "name_of_spectroscopy_constant": FLOAT. The value of the queried
                            spectroscopy constant.
                },
                ...
        ]
}
```

(b) Format of the returned JSON object with “query= name_of_spectroscopic_constant”.

A json object containing of a given molecule. When "name_of_spectroscopy_constant" is undefined (e.g. api/?chemical_formula=AlF, or api/?chemical_formula=AlF&query=), the query returns all the spectroscopy constants of the given molecule. The information about the molecules (chemical formula), their states (in Latex) and masses (in the a.u. unit) are also given.

```
{
        "accessed": STRING. Time of access, with the ISO 8601 format,
        "id_molecule": INTEGER. A unique ID of the molecule,
        "chemical_formula": STRING. The chemical formula of the molecule.
        "n_records": INTEGER. Number of records (molecules).
        "data":
        [
                {
                        "reference": STRING. The reference where the spectroscopy
                        constant was given, with the APS format.
                        "reference_date": STRING. Date of the reference where the
                            spectroscopy constant was given, with the format of "Month Year".
                        "id_record": INTEGER. A unique ID of the record.
                        "state": STRING. The state symbol, in the Latex format.
                        "mass": FLOAT. Mass of the molecule in the corresponding state, in the
a.u. unit.
                        "name_of_spectroscopy_constant": FLOAT. The value of the queried
                            spectroscopy constant.
                },
                ...
        ]
}
```

(c) Format of the returned JSON object with the query keyword “chemical_formula = chemical_formula”.

Figure 9: Sample returns of API’s. Figures reproduced from ref. [20].

## 2.3 Implementation of the database

### 2.3.1 Querying Data

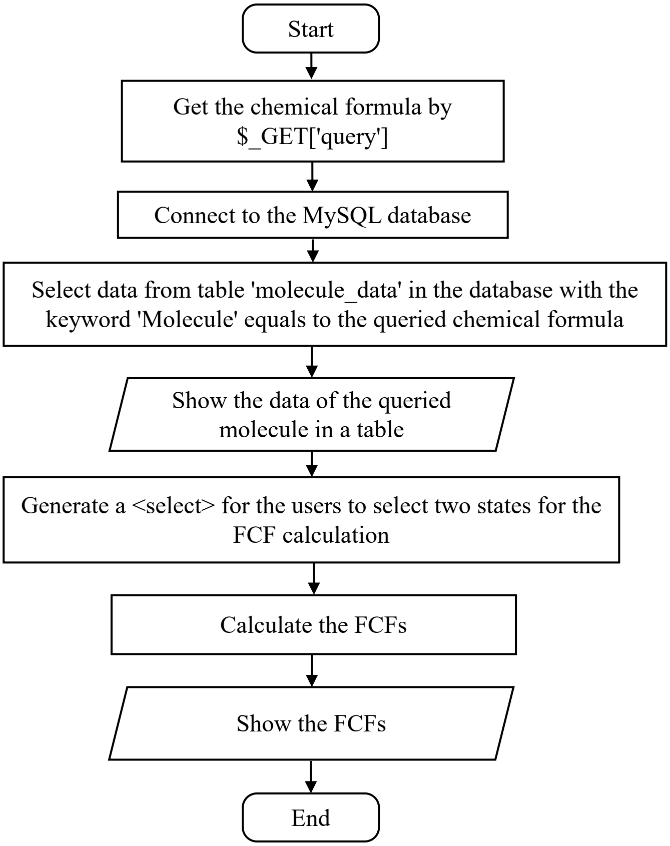

Start
Get the chemical formula by $_GET['query']
Connect to the MySQL database
Select data from table 'molecule_data' in the database with the keyword 'Molecule' equals to the queried chemical formula
Show the data of the queried molecule in a table
Generate a <select> for the users to select two states for the FCF calculation
Calculate the FCFs
Show the FCFs
End

Figure 10: Flowchart of search_data.php. Figure reproduced from ref. [20].

The primary purpose of the website is to enable users to search the database for the spectroscopic constants of a specific diatomic molecule. This functionality is implemented in “search_data.php”, as illustrated in Fig.10, showing the flowchart of the search engine in the database. The query keyword is the chemical formula of the molecule, obtained through the HyperText Transfer Protocol (HTTP) GET method from the input field. Subsequently, a query is performed in the MySQL table “molecule_data”, retrieving rows that contain the spectroscopic information of the queried molecule. If the molecule exists in the database, the query results are displayed in a table and can be downloaded in CSV format. Furthermore, an HTML <select> tag is dynamically generated based on the available electronic states of this molecule, allowing users to choose two electronic states for calculating the Franck-Condon factors.

### 2.3.2 Calculation of Franck-Condon factor (FCF)

Under the Born-Oppenheimer approximation, during an electronic transition of a molecule, the nuclei, heavier than the electrons, can be considered to maintain their configuration with minimal alteration. In this scenario, the Franck-Condon principle states that the intensity of an electronic transition is proportional to the square of the overlap between the vibrational wavefunctions, $|\Psi_v\rangle$, of the two states of the transition as [47]

$$|\langle\Psi_v|\Psi_{v'}\rangle|^2. \tag{1}$$

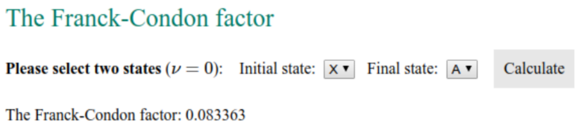
Therefore, the FCFs provide valuable information about the preferred vibrational transitions between different electronic states of a molecule.

From Eq.(1), it is clear that the FCFs are very sensitive to vibrational wavefunctions. Hence, it is necessary to count on accurate interatomic potentials. In the case of vibrational levels near the minimum well of the interatomic potentials, the influence of second and higher anharmonic corrections is negligible. Accordingly, it is possible to describe the interatomic potential via a Morse potential, the potential of choice for this work. Next, we solve numerically the time-independent Schrödinger equation via a discrete variable representation (DVR) method [48, 49] [2]. In DVR, the basis sets are associated with a specific set of quadrature points, ensuring that the potential matrix is diagonal, while the kinetic energy operator contains non-diagonal terms. For this implementation, the number of DVR quadrature points is set to 200, which yields vibrational wavefunctions with an error of less than 0.1%. The overlap between the vibrational wavefunctions is calculated numerically using the trapezoidal rule.

The user can obtain FCFs, calculated on-the-fly, for any of the molecules in the database. The results can be displayed as bar and density plots,

[2] The DVR method offers efficient and accurate solutions to quantum mechanical problems involving a small number of particles [50].

similar to those presented in Fig.12, with the assistance of the JavaScript library D3.js[51, 52].

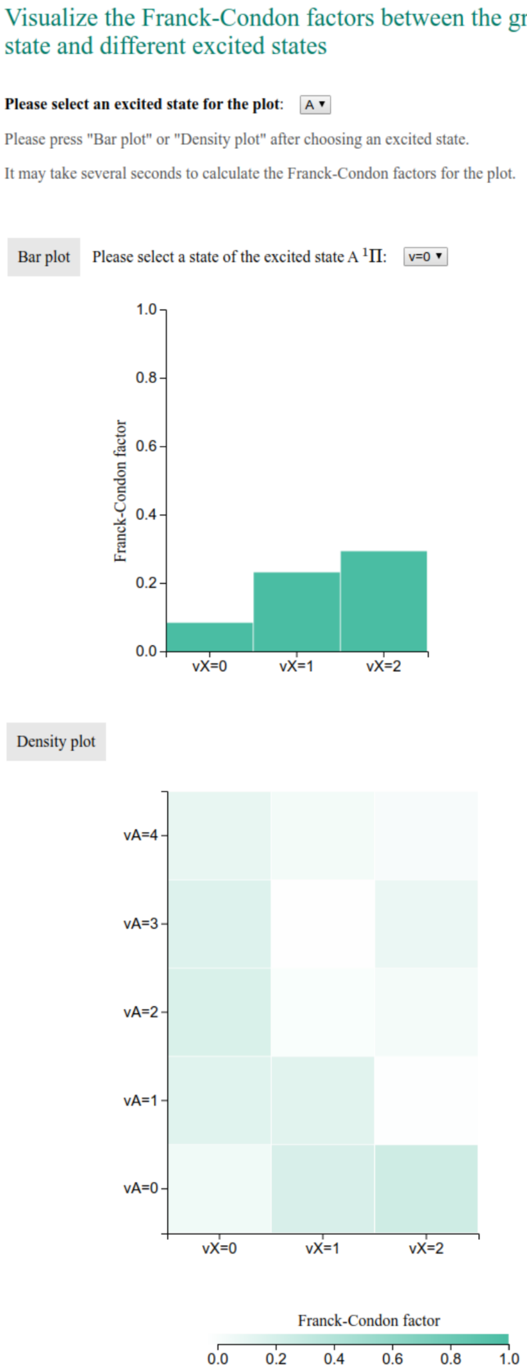


The Franck-Condon factor

**Please select two states** ($\nu = 0$): Initial state: X ▾ Final state: A ▾ Calculate

The Franck-Condon factor: 0.083363

Figure 11: Calculation of the Franck-Condon factor. Figure reproduced from ref. [20].



Visualize the Franck-Condon factors between the ground state and different excited states

**Please select an excited state for the plot**: A ▾

Please press "Bar plot" or "Density plot" after choosing an excited state.

It may take several seconds to calculate the Franck-Condon factors for the plot.

Bar plot Please select a state of the excited state A $^1\Pi$: v=0 ▾

Density plot

Figure 12: Visualization of the Franck-Condon factor between different states. Figure reproduced from ref. [20].

### 2.3.3 User contribution

Users can register on the website to contribute new spectroscopic data to the database through a web page interface containing input forms. This contribution feature is implemented using several PHP scripts, including “contribution_main.php”, “contribution_data.php”, “contribution_submit.php”. Additionally, the web manager authorization is managed through “contribution_confirm.php” and “contribution_reject.php”. After the submission is confirmed, the data is uploaded into the database via“contribution_insert_data.php”. The flowcharts illustrating the contribution process are displayed in Figs. 13-17.

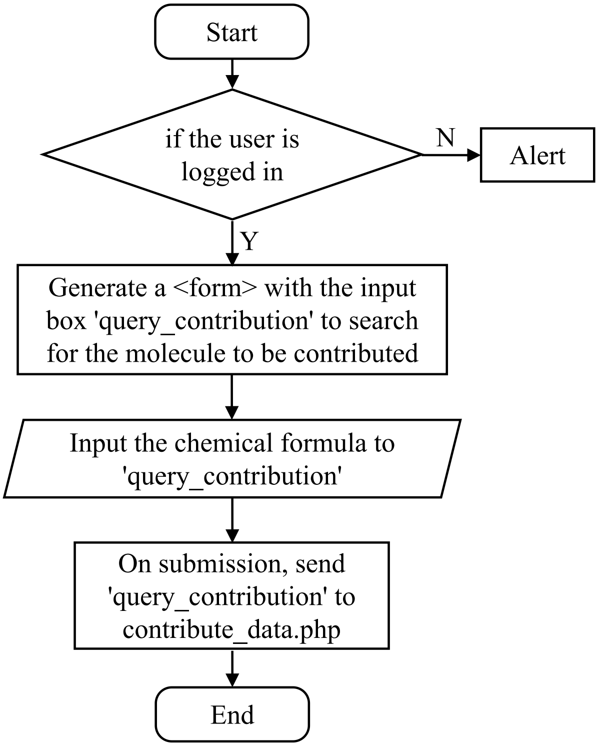

Start
if the user is logged in
N
Alert
Y
Generate a <form> with the input box 'query_contribution' to search for the molecule to be contributed
Input the chemical formula to 'query_contribution'
On submission, send 'query_contribution' to contribute_data.php
End

Figure 13: Flowchart of contribution_main.php. Figure reproduced from ref. [20].

In "contribution_main.php" (Fig. 13), logged-in users are presented with an input field named "query_contribution" where they can enter the chemical formula of the molecule they wish to contribute to the database.

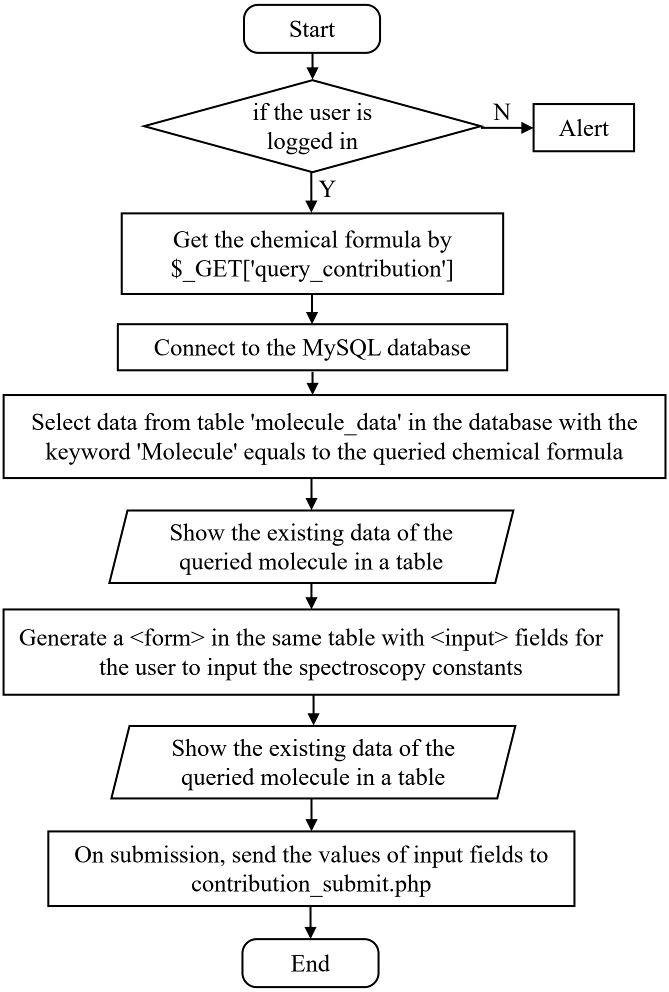

Start
if the user is logged in
N
Alert
Y
Get the chemical formula by $_GET['query_contribution']
Connect to the MySQL database
Select data from table 'molecule_data' in the database with the keyword 'Molecule' equals to the queried chemical formula
Show the existing data of the queried molecule in a table
Generate a <form> in the same table with <input> fields for the user to input the spectroscopy constants
Show the existing data of the queried molecule in a table
On submission, send the values of input fields to contribution_submit.php
End

Figure 14: Flowchart of contribute_data.php. Figure reproduced from ref. [20].

Upon submission, the input field action is directed to “contribute_data.php” (Fig. 14), which retrieves data from the “molecule_data” table in the database corresponding to the queried chemical formula. The results are displayed in a table.

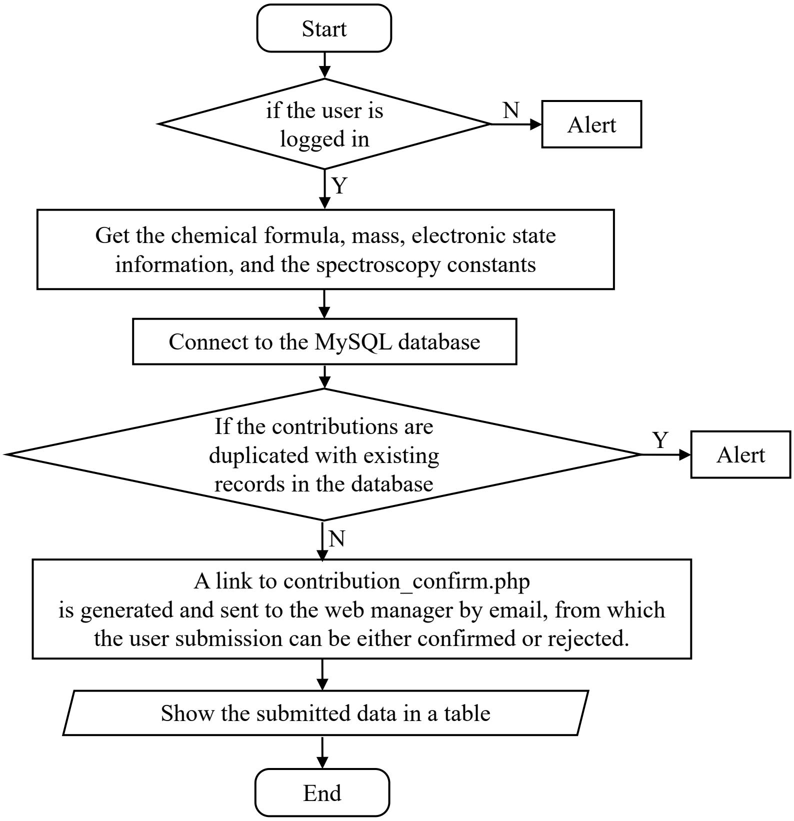

Start
if the user is logged in
N
Alert
Y
Get the chemical formula, mass, electronic state information, and the spectroscopy constants
Connect to the MySQL database
If the contributions are duplicated with existing records in the database
Y
Alert
N
A link to contribution_confirm.php is generated and sent to the web manager by email, from which the user submission can be either confirmed or rejected.
Show the submitted data in a table
End

Figure 15: Flowchart of contribution_submit.php. Figure reproduced from ref. [20].

An HTML <form> is generated within the same table for users to input the spectroscopic constants. This form’s action is set to “contribution_submit.php” (Fig. 15), where the input data is processed, and a duplicate check is performed. The duplicate check compares the spectroscopic constants in the current submission with the existing data in the database.

The newly uploaded data is presented in a table to the user, along with a link to "contribution_confirm.php", which is sent to the web managers via email.

Upon receiving the link, "contribution_confirm.php" presents the existing data in the database alongside the user's submission in tables. This comparison enables the web managers to review and decide whether to confirm or reject the contribution (Fig.16).

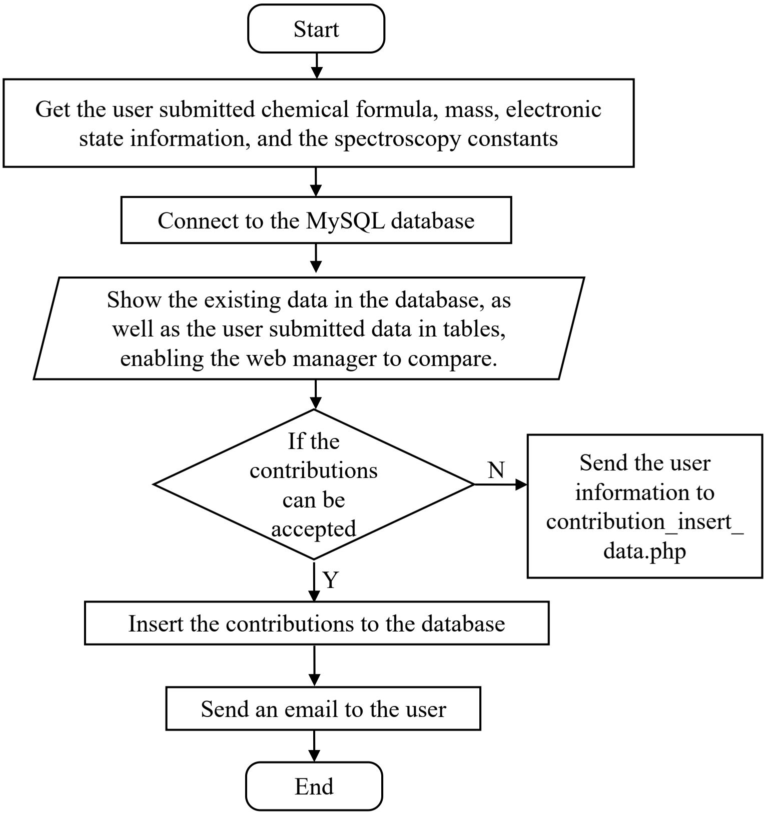

Start
Get the user submitted chemical formula, mass, electronic state information, and the spectroscopy constants
Connect to the MySQL database
Show the existing data in the database, as well as the user submitted data in tables, enabling the web manager to compare.
If the contributions can be accepted
N
Send the user information to contribution_insert_ data.php
Y
Insert the contributions to the database
Send an email to the user
End

Figure 16: Flowchart of contribution_confirm.php. Figure reproduced from ref. [20].

The data provided by the user undergoes a two-step validation process. First, it is checked whether the journal referenced by the user is indexed in the Web of Science. Second, the user-provided data is verified against the data in the referenced paper. If the data is correct, the contribution is accepted, ensuring high-quality data on the website. In the case of rejection, "contribution_reject.php" generates an HTML form for the web managers to enter the reasons for rejection (Fig.17), which is then communicated to the contributor via email.

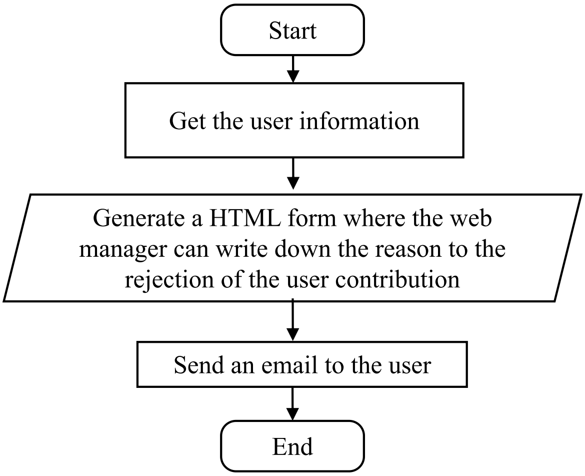

Start
Get the user information
Generate a HTML form where the web manager can write down the reason to the rejection of the user contribution
Send an email to the user
End

Figure 17: Flowchart of contribution_reject.php. Figure reproduced from ref. [20].

### 2.3.4 Website and database security

Security and data protection are the main issues in web page design. A webpage involving SQL queries and insertion operations needs a security protocol to avoid SQL injection attacks. Specifically, we undertake several security measures after receiving a query with the chemical formula as the keyword:

- We check the length of the keyword to ensure it is not longer than 4 characters, which is impossible for diatomic molecules.
- To prevent SQL injection, we encode the keyword using "mysqli_real_ escape_string()".

With mysqli_ real_escape_string(), all user inputs are appropriately sanitized. This function escapes special characters in the input data, ensuring that malicious SQL queries are neutralized.

- Prepared statements are utilized for SQL SELECT and INSERT operations, ensuring the separation of SQL code from user input. These statements separate the SQL code from user input, preventing potential SQL injection attacks. User input is treated as parameters rather than being directly inserted into the SQL query, providing an extra layer of security.
- We adopt a pre-checking approach to avoid potential security threats regarding the APIs. Before executing the query, we verify whether the provided keyword is part of the allowed list. This way, we avoid potential unauthorized queries and guarantee that only valid input is processed. These security measures collectively safeguard the integrity of the website's database and protect against SQL injection vulnerabilities.
- The protection to passwords has been implemented with "password_ hash()", with a fixed cost and automatically set salt. The abovementioned SQL injection protections are also made.
- Webpages, including "contribution_insert_data.php" and "contribution_reject_email.php" are accessible only for administrators, to avoid anyone being able to use them without any authentication, for example writing data to the database, or sending spam.

By combining these security measures, the website effectively guards against SQL injection attacks, ensuring the integrity and safety of the database and protecting user data from potential threats.

The complete database is accessible through the API in the website https://rios.mp.fhi.mpg.de. The source code is available at https://github.com/hlslxy/DMSD under the Apache License. The database as well as its mirrored copy, are maintained by the Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), a service organization that collaborates with the University of Göttingen and the Max Planck Society as a data and IT service center.

## 2.4 Conclusion and outlook

The diatomic molecule spectroscopic database offers open access to both ground and excited state spectroscopic constants for polar diatomic molecules, along with Franck-Condon factors that characterize transitions between various electronic states. This database operates dynamically, enabling registered users to contribute spectroscopic data.

Currently, the database encompasses a modest subset of the potential diatomic polar molecules found across the periodic table. As of April 2020, the database has 608 records, comprising 130 molecules in a $\Sigma$ ground state, 34 molecules with a $\Pi$ ground state, and 5 molecules with a $\Delta$ ground state. Currently (April 2023), the database has 177 molecules, 134 of which have a $\Sigma$ ground state, 35 have a $\Pi$ ground state, 7 have a $\Delta$ ground state, and 1 has a $\Phi$ ground state.

Further extensions of this work may entail:

- The incorporation of charged diatomic molecules.
- The inclusion of homogeneous diatomic molecules.
- The potential for users to upload data from CSV files.

Building on this research, the Database of Spectroscopic Constants of Diatomic Molecules (DSCDM) [53] has been recently developed. It encompasses 344 diverse neutral diatomic molecules, both heterogeneous and homogeneous, featuring the most current and accurate spectroscopic constants. Additionally, it provides machine-learning predictors for spectroscopic constants.

# 3

# ON THE RELATIONSHIP BETWEEN SPECTROSCOPIC CONSTANTS OF DIATOMIC MOLECULES: A MACHINE LEARNING APPROACH

## 3.1 Background and methods

In the early stages of the development of molecular spectroscopy within chemical physics during the 1920s [54], researchers made intriguing empirical observations regarding various spectroscopic properties [55, 56, 57]. Specifically, they noted a correlation between the equilibrium distance ($R_e$) and the harmonic vibrational frequency ($\omega_e$) in diatomic molecules. Over time, this relationship between $R_e$ and $\omega_e$ became more pronounced, and additional empirical connections among spectroscopic constants were uncovered [58, 59, 60, 61, 62, 63, 64, 65]. However, these empirical relationships were typically valid only for specific atomic numbers or groups of constituent atoms. These findings prompted the development of realistic diatomic molecular potentials [57, 66, 67, 68, 69, 70] and sparked discussions within the physical chemistry community about the "periodicity" of diatomic molecules [71].

The advent of quantum chemistry brought insights into the underlying physics behind empirical relations among spectroscopic constants. Notably, through the application of the Hellmann-Feynman theorem, researchers were able to establish a direct connection between $\omega_e$ and the electronic density at $R_e$ [72, 73, 74, 75]. Consequently, a first-principles-based explanation, involving a few free parameters, emerged to account for the observed empirical relationships among spectroscopic constants [76, 77, 78, 79, 80, 81, 82, 83]. Nevertheless, the derived relations based on electronic density remained applicable only to specific subsets of molecules. To this day, a comprehensive set of general relations for spectroscopic constants of diatomic molecules in terms of the properties of their constituent atoms has remained elusive.

The accuracy of quantum chemistry methods relies on finite basis sets that have been optimized for individual elements within specific constraints [84, 85]. Meanwhile, an accurate depiction of the electronic structure of the system is imperative. This is accomplished through a hierarchical approach that encompasses various treatments of electron correlation [84, 85]. On the other hand, widely employed Kohn-Sham density functional theory (DFT) methods necessitate precise electron exchange-correlation density functionals. Non-empirical density functionals are formulated with specific constraints and may incorporate numerous free parameters [86, 87, 88, 89]. In contrast, semi-empirical density functionals adopt more adaptable functional forms, often characterized by multiple coefficients that are tailored to match various experimental or theoretical reference properties [86, 90].

In a different vein, machine learning (ML) techniques unveil underlying relationships from data, often referred to as the "training set", and subsequently construct predictive models based on these relationships. These models can offer quantitative predictions for other systems that share similar underlying physical principles. Moreover, ML provides the potential to uncover correlations between diverse properties within the system under investigation [91].

This study seeks universal relationships between spectroscopic constants in heteronuclear diatomic molecules, applicable to a wide range of molecular species. Our findings are based on the application of state-of-the-art machine learning (ML) models to a conventional dataset comprising experimental spectroscopic constants for diatomic molecules. In particular, we employ the Gaussian process regression (GPR) model [92] to predict key parameters, namely $R_e$, $\omega_e$, and the binding energy, $D_0$, as functions of the constituent atoms' group and period. Additionally, our models are capable of predicting $R_e$ and $\omega_e$ for the A-excited electronic state of a given molecule [1].

Our approach, based on an ML perspective, extends the conventional wisdom that some chemical properties of a system are contingent on the group and period of the constituent atoms, finding accurate predictions of spectroscopic constants based on atomic properties. Hence, our results can be viewed as a stepping stone toward finding universal relationships between spectroscopic constants, a *dream* that is over a century old.

### 3.1.1 The dataset

In this study, our primary focus is on heteronuclear molecules, given their significance in laser cooling applications within the field of cold and ultracold chemistry [94, 95, 34]. Our dataset comprises essential spectroscopic constants included in the dataset introduced in Chapter 2, namely $R_e$, $\omega_e$, and $D_0$, for the ground electronic state of heteronuclear diatomic molecules. Specifically, it includes experimental values of $R_e$ and $\omega_e$ for 256 heteronuclear diatomic molecules, while experimentally determined values of $D_0$ are available for 197 of these molecules.

[1] This chapter is written based on reference [93]: Xiangyue Liu, Stefan Truppe, Gerard Meijer, and Jesús Pérez-Ríos. On the relationship between spectroscopic constants of diatomic molecules: A machine learning approach. *RSC advances*, 11(24):14552–14561, 2021, https://doi.org/10.1039/D1RA02061G, with permission from the Royal Society of Chemistry.

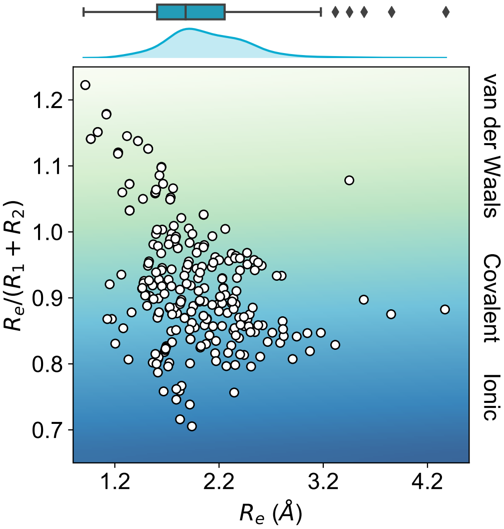


Figure 18: Ratio of the equilibrium distance, $R_e$, to the sum of the atomic radii of the atoms forming a molecule, $R_1 + R_2$, vs. $R_e$. The background color indicates the nature of the molecular bond in each of the molecules. The density in the upper part of the figure shows the kernel density distribution of $R_e$. The box plot shows the minimum, the maximum, the sample median, and the first and third quarterlies of $R_e$. The empirical atomic radii of the atoms are taken from Ref. [96]. Figure reproduced from ref. [93].

To the best of our knowledge, this dataset represents one of the most comprehensive collections of experimental ground state properties for heteronuclear diatomic molecules[2]. Figure 18 illustrates the distribution of equilibrium distances and their ratios to the sum of the atomic radii of the constituent atoms, denoted as $R_1 + R_2$, within the dataset. Notably, most molecules exhibit equilibrium distances ranging from 1.4 Å to 3.8 Å, with a peak occurrence at approximately 1.7 Å. Furthermore, an examination of the values of $R_e/(R_1 + R_2)$ reveals that the molecules in the dataset exhibit a diverse range of bond types, encompassing covalent, van der Waals, and ionic bonds.

[2] Recently, I have been involved in a new effort towards a more extensive database (https://dscdm.physics.stonybrook.edu) [53]

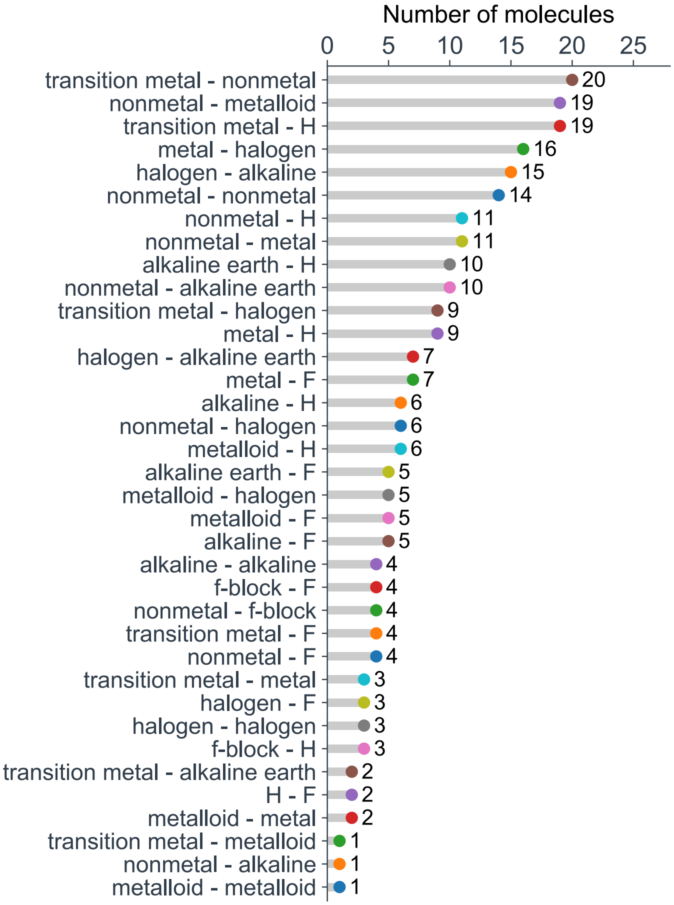


Figure 19: Molecules in the dataset classified by the types of their constituent atoms. Figure reproduced from ref. [93].

We have categorized the dataset based on the constituent atoms present within each molecule, and the results are presented in Figure 19. Upon analysis, we observed that the dataset predominantly comprises a diverse range of metal and non-metal halides, hydrides, and metalloid compounds. Over 20% of the dataset comprises transition metal compounds, including elements from the f-block. Consequently, the present dataset is inclusive, extending beyond the realm of main-group diatomic molecules and encompassing more intricate and complex atoms from a chemical perspective.

Furthermore, we also investigated a set of 131 molecules for which $R_e$ and $\omega_e$ data are available for the A-excited electronic state. The A-state dataset primarily features metal and non-metal compounds, including transition metal compounds and several f-block compounds.

### 3.1.2 Machine learning method

The quest to uncover universal relationships among spectroscopic constants is closely linked to the challenge of understanding how atomic and molecular properties collectively contribute to defining a molecule's spectroscopic property, denoted as $y = f(\mathbf{x})$. In this context, $\mathbf{x} = (x_1, x_2, ..., x_n)$ represents a collection of various atomic properties of the constituent atoms or molecular properties, where $n$ indicates the number of input features relevant to the specific problem at hand. Here, we adopt a GPR approach to model those relationships.

GPR adopts a Bayesian perspective and posits a prior distribution over the space of functions. This approach allows for a more flexible and data-driven exploration of the relationships between spectroscopic constants. A GPR is defined as

$$f(\mathbf{x}_i) \sim GP(m(\mathbf{x}_i), K(\mathbf{x}_i, \mathbf{x}_j)), \tag{2}$$

with a joint multivariate-Gaussian distribution, centered at $m(\mathbf{x_i})$ and characterized by the covariance function $K(\mathbf{x}_i, \mathbf{x}_j)$, which specifies the correlation (or "similarity") between data points [92]. The spectroscopic properties, denoted as $y$, are modeled as follows:

$$P(y_i | f(\mathbf{x}_i), \mathbf{x}_i) \sim \mathcal{N}(y_i | \mathbf{h}(\mathbf{x}_i)^T \beta + f(\mathbf{x}_i), \sigma_y^2). \tag{3}$$

where $\mathcal{N}(\mu, \sigma)$ represents a normal distribution of mean $\mu$ and standard deviation $\sigma$. Here, the basis functions, represented as $\mathbf{h}(\mathbf{x}_i)$, project the set $\{\mathbf{x}_i\}$ into a new, possibly higher-dimensional feature space characterized by coefficients $\beta$. The term $\sigma_y$ encompasses the noise present in the observations [92, 97].

The training set denoted as $\mathscr{D} = \{(\mathbf{x}_i, y_i) | i = 1, \cdots, N\}$, comprises $N$ observations and serves to constrain the distribution of available functions through Bayes' theorem. For prediction purposes, the mean of the posterior distribution is employed. The specific functional forms of $K(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{h}(\mathbf{x})$ can be tuned during the optimization of GPR models, and an optimal model can be selected based on the cross-validation performances. More details about the GPR methods can be found in the Appendix.

While GPR is a powerful tool for modeling non-linear relationships between variables, it can be beneficial to engineer features to make the regression problem more amenable to a linear model. By introducing non-linear terms or transformations of the input features, one can potentially capture complex relationships more cheaply. This may lead to a model that is easier to interpret and understand and can also reduce the risk of overfitting. In this study, we have tested different input features $\mathbf{x}$ generated from atomic or molecular properties and compared the performance.

### 3.1.3 Model performance evaluation

In training and evaluating regression models, as customary in machine learning, we divide the ground state dataset into two subsets: the training set and the test set. The training set is used for learning a given spectroscopic constant based on the atomic properties of the constituent atoms. In contrast, the test set consists of molecules that have not been part of the learning process and are thus new to the regression algorithm. For example, when learning the equilibrium internuclear distance ($R_e$) and the harmonic vibrational frequency ($\omega_e$), the training set comprises 231 molecules, while the test set contains 25 molecules. For learning $\log \frac{D_0}{R_e^3 Z_1 Z_2}$, the dataset is split into a training set of 172 molecules and a test set of 25 molecules. Finally, in the case of excited states spectroscopic constants, the training set encompasses 106 molecules, and the test set includes 25 molecules.

From a machine learning perspective, the present dataset may be considered relatively small. When splitting the dataset into training and test sets, there's a potential for the training set to not fully represent the underlying data distribution, potentially leading to bias in test set performance. At the same time, it would be crucial to examine whether the relationships between spectroscopic constants, as modeled from different subsets of molecules, hold true universally. To this end, we have developed a Monte Carlo (MC) approach. This approach stratifies the dataset into 25 strata based on the true values of the labels ($R_e$, $\omega_e$, and $\log \frac{D_0}{R_e^3 Z_1 Z_2}$ in this work).

Our MC approach involves two loops, as illustrated in panel (a) of Figure 20, one for training and another for evaluation of the models. We divide the dataset into training and test sets in the outer loop. The training set is used for model learning, while the test set is reserved for model evaluation. In the inner loop, we further divide the training set into five stratified folds for cross-validation (CV) during hyperparameter optimization. As depicted in panel (b), in the outer loop, we employ an MC approach to perform the training/test split. We randomly select 25 test molecules from the dataset, which has been previously stratified into 25 strata based on the levels of the true values of the labels. This stratification helps maintain the proportional composition of the dataset upon splitting [98]. A regression model is trained in each MC step, providing predictions for the training and test sets. Consequently, we report the mean and standard deviation of the predictions for each molecule when they are used in both the training and test sets, derived from all 1000 MC steps for model performance evaluation and 500 MC steps for generating learning curves.

The performance of the models is assessed using three distinct estimators. The first estimator is the mean absolute error (MAE), defined as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - y_i^*|, \tag{4}$$

where $y_i^*$ are the true values, $y_i$ are the predictions, and $N$ is the number of observations. The second estimator is the root mean square error (RMSE), which is given by

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - y_i^*)^2}. \tag{5}$$

The last estimator is the normalized error $r_E$, defined as the ratio of the RMSE to the range of $y$,

$$r_E = \frac{\mathrm{RMSE}}{y_{max} - y_{min}}. \tag{6}$$

### 3.1.4 The learning curves

Learning curves illustrate the training and test performance of a model as a function of the training set size $N$. These curves provide valuable insights into a model's bias and variance. Moreover, they enable us to discern whether the model's performance benefits from an increase in the training set size.

For each data point on the learning curve, the training process is executed with the aid of 500 different training/test splittings, accomplished through the MC approach.

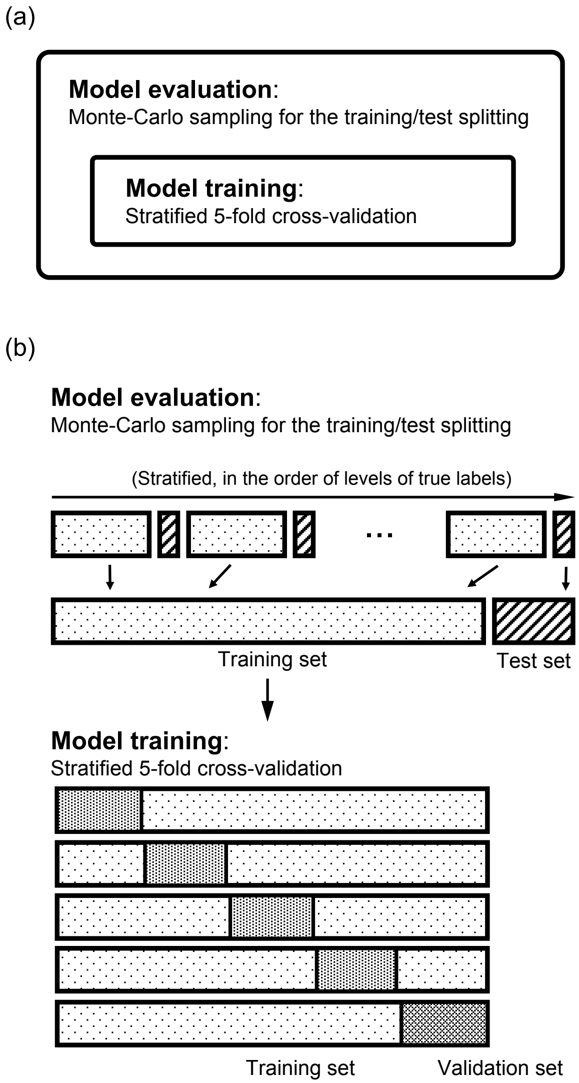

(a)
Model evaluation:
Monte-Carlo sampling for the training/test splitting
Model training:
Stratified 5-fold cross-validation
(b)
Model evaluation:
Monte-Carlo sampling for the training/test splitting
(Stratified, in the order of levels of true labels)
...
Training set
Test set
Model training:
Stratified 5-fold cross-validation
Training set
Validation set

Figure 20: Scheme of the training/test set splitting in the model evaluation. (a) There are two loops: The outer loop for the model performance evaluation, and the inner loop for the training of model and hyperparameter optimization. (b) In the outer loop, the data are stratified based on the true values of the labels, and each stratum is randomly split into training and test sets. In learning the properties, the training sets are further split into training and validation sets to perform a stratified 5-fold cross-validation. Figure reproduced from ref. [93].

## 3.2 The quest of relationships between spectroscopic constants of diatomic molecules

As soon as molecular spectroscopy became an indispensable tool for analyzing the unique characteristics of molecules, researchers began to amass a wealth of molecular spectra. In the process, they discovered approximate relationships among spectroscopic constants, suggesting that these constants might be correlated on empirical grounds. Notably, in the case of hydrogen halides [55, 99, 100, 101], it was observed that the equilibrium distance and the harmonic vibrational frequency were related by the expression $R_e^2 \omega_e^2 m = \text{const}$, where $m$ represents the reduced mass of the molecule. This correlation later evolved into Badger's rule [59], expressed as $R_e^i \omega_e^2 m = \text{const}$, with $i$ as a natural number.

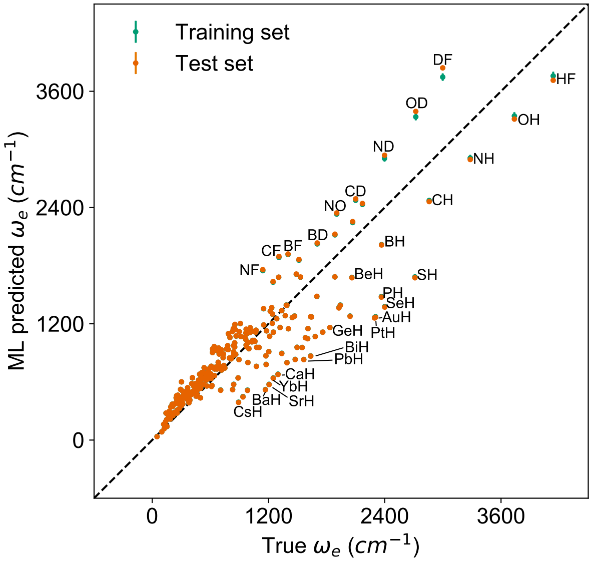


Figure 21: Predictions of $\omega_e$ with $R_e^{-2}$ by a linear regression model. Figure reproduced from ref. [93].

Conversely, Mecke and Birge, in their study encompassing 16 molecules, including homonuclear molecules and molecular ions, found that the expression $R_e^2 \omega_e = \text{const}$ provided a better description of the observed spectra [56, 102]. Applying a linear regression with this relationship on the current dataset to predict $\omega_e$ as a function of $R_e^{-2}$, the test-set RMSE is $297.3 \pm 1.4$ cm$^{-1}$, and $r_E$ is $7.27 \pm 0.03\%$, as shown in Fig. 21. In this figure, the largest errors come from the predictions of hydrides, deuterides, and several fluorides. This model underestimates $\omega_e$ of hydrides while overestimating $\omega_e$ of deuterides because $R_e$ does not have an obvious isotope effect.

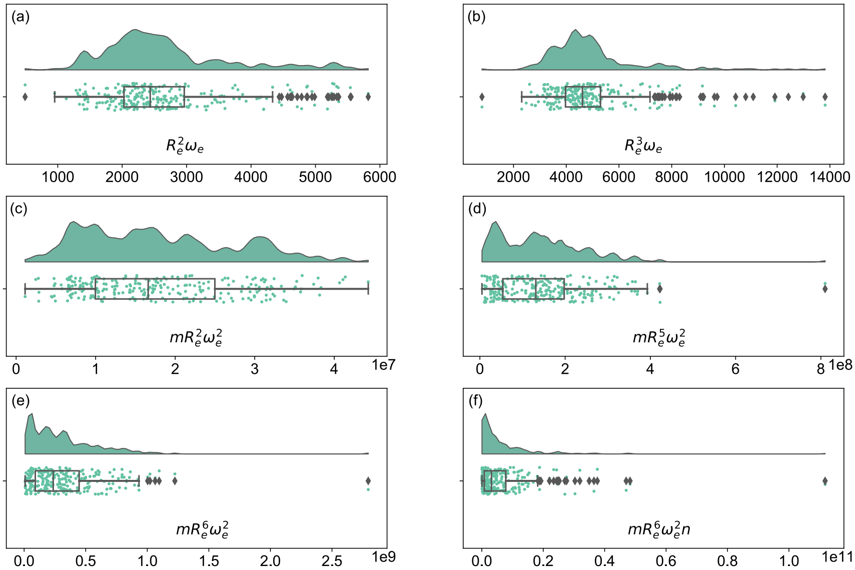


Figure 22: Distribution and box plots of $R_e^a\omega_e^b$ with different powers combined with the reduced mass $m$ and number of valence electrons $n$. Figure reproduced from ref. [93].

Similarly, Morse proposed an empirical relationship in line with a specific functional form for the interatomic interaction, expressed as $R_e^3\omega_e =$ const [57]. Furthermore, more intricate relationships between equilibrium distance and vibrational harmonic frequency were suggested, such as $mR_e^6\omega_e^2n^a$, where $n$ denotes the number of valence electrons, and $a$ is a rational number [70]. However, when examined with a more extensive dataset, as in our current study, none of these empirical relationships prove to be universally applicable, as depicted in Figure 22.

At the same time, as more comprehensive spectroscopic data on molecules became accessible and advanced, precise quantum chemistry tools were developed, enabling researchers to seek a first-principles explanation for the empirically observed relationships among spectroscopic constants. Leading this endeavor, Parr and his colleagues investigated the electron density within molecules as the underlying source of these relationships. Their model posits that the electron density created by one atom in the vicinity of another atom is equivalent at the equilibrium distance, which corresponds to the sum of the atomic radii. Specifically, within a molecule, the electron density of atom 1 at the position of atom 2 is expressed as [80]

$$\rho_1(2) = CZ_1 \exp\{(-\xi R_1)\}, \tag{7}$$

where $C$ is a fitting parameter, $\xi$ represents the decay constant of the electron density. Within this model, a connection emerges between the atomic numbers of the two atoms, denoted as $Z_1$ and $Z_2$, and the equilibrium internuclear distance, $R_e$, of a diatomic molecule, expressed as [80, 103, 82, 83]

$$Z_1 Z_2 = A \exp(\xi R_e), \tag{8}$$

where $A$ is a free parameter. According to this relationship, $R_e$ is linearly dependent on $\log(Z_1 Z_2)$, as

$$R_e = \xi^{-1} \log Z_1 Z_2 - \xi^{-1} \log A. \tag{9}$$

Nevertheless, the validity of this relationship has solely been assessed for molecules featuring atoms originating from the same group of the periodic table [82]. Indeed, as shown in Fig. 23, the above linear relationship no longer holds in the current dataset, giving a mean test-set RMSE of $0.3591 \pm 0.0006$ Å and $r_E$ being $10.41 \pm 0.01\%$.

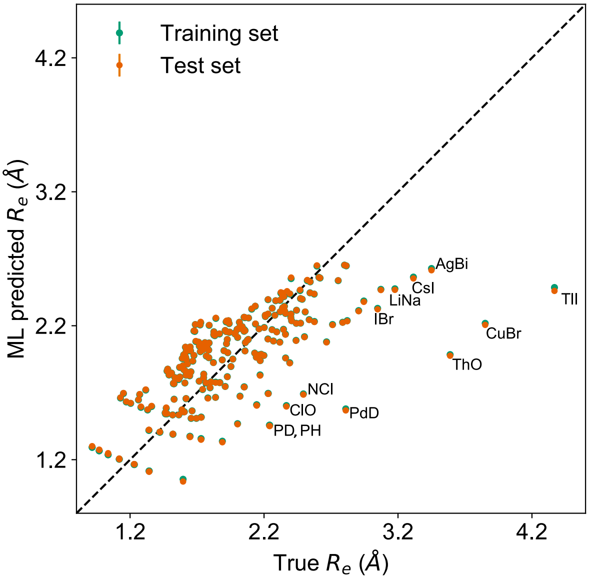


Figure 23: Predictions of $R_e$ with $\log(Z_1Z_2)$ by a linear regression model. Figure reproduced from ref. [93].

Anderson, Parr, and their colleagues also proposed a relationship between $\omega_e$ and $R_e$ [82], expressed as

$$m\omega_e^2 = 4\pi C Z_1 Z_2 e^{-2R_e}, \tag{10}$$

based on the Born-Oppenheimer approximation, the electron density described in Eq. (7), and the Hellman-Feynman theorem. Eq. (10) allows one to express the harmonic vibrational frequency in terms of the equilibrium distance and atomic properties as

$$\omega_e = \sqrt{\frac{C' Z_1 Z_2 e^{-2R_e}}{m}}. \tag{11}$$

where $m$ is the reduced mass of the molecule. Applying this relationship to predict $\omega_e$ using a linear regression model with $\sqrt{Z_1 Z_2 e^{-2R_e}/m}$ as independent variable, as shown in Fig. 24, the mean test RMSE is $529.5 \pm 1.2$ cm$^{-1}$, and $r_E$ is $12.95 \pm 0.03\%$.

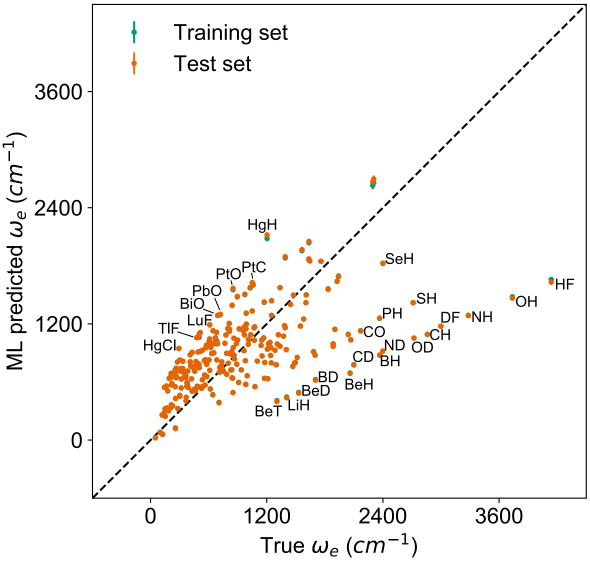


Figure 24: Predictions of $\omega_e$ with $\sqrt{Z_1 Z_2 e^{-2R_e}/m}$ by a linear regression model. Figure reproduced from ref. [93].

In a similar vein, by extending the connection between equilibrium distance and harmonic vibrational frequency, one can establish a relationship between the atomic number $Z_i$, $R_e$, and the dissociation energy $D_e$ [80, 103, 82, 83], as

$$\frac{D_e}{R_e^l} = 4\pi C Z_1 Z_2 \exp(-\xi' R_e), \tag{12}$$

which can be rewritten as

$$\log \frac{D_e}{R_e^l Z_1 Z_2} = -\xi' R_e + \log(4\pi C). \tag{13}$$

For the derivation of Eq. (12), it is assumed that $D_e = Am\omega_e^2 R_e^l$ without any additional justification [83]. In Eq. (13), the values are $l = 3$ and $\xi' = 0.97$. Eq. (13) has been tested on a dataset of 150 molecules, yielding satisfactory results, although no further characterization of the model's performance was provided to assess its quality objectively. Finally, by employing the relationship between the dissociation energy, $D_e$, and the binding energy, $D_0$,

$$D_e = D_0 + \frac{1}{2}\hbar\omega_e - \frac{1}{4}\hbar\omega_e x_e, \tag{14}$$

where $\omega_e x_e$ represents the first anharmonic correction to the harmonic vibrational frequency, it should be feasible to construct a linear regression model for $\log \frac{D_0}{R_e^l Z_1 Z_2}$. The predictions are shown in Fig. 25. We notice that most of the outliers are highly ionic molecules. This is due to the fact that Parr and coworkers presumedan exponentially decaying functional form for the electron density that is descriptive of covalent molecules.

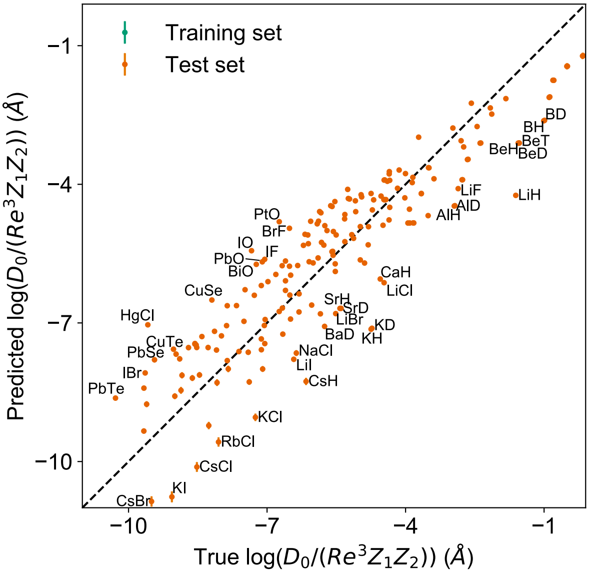


Figure 25: Predictions of $\log \frac{D_0}{R_e^3 Z_1 Z_2}$ with $R_e$ by a linear regression model. Figure reproduced from ref. [93].

## 3.3 Prediction of spectroscopic constants with Gaussian process regression

### 3.3.1 Learning ground state spectroscopic constants

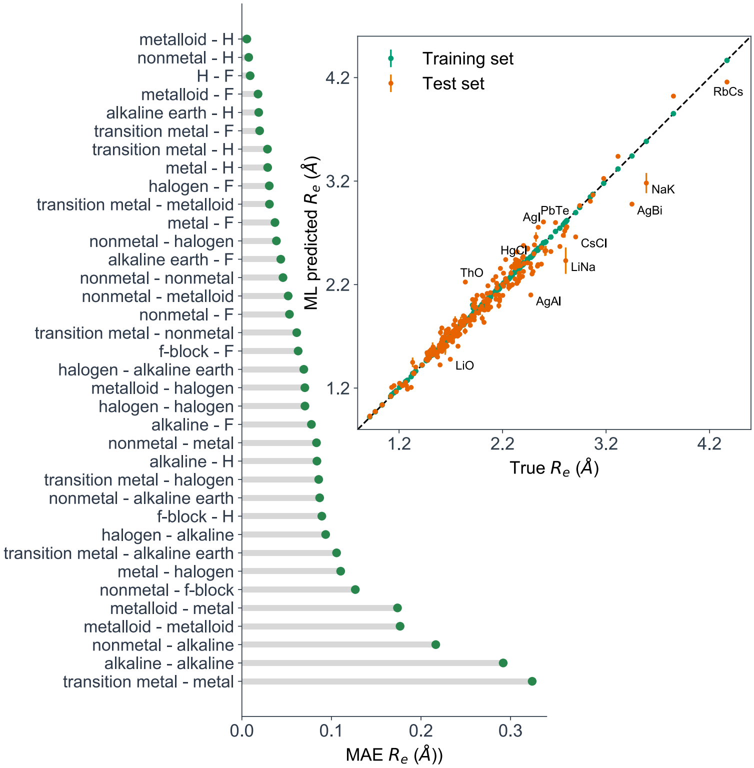


Figure 26: GPR performance on predicting $R_e$ using $(g_1, g_2, p_1, p_2)$ as input features classified by the types of the constituent atoms. In particular, the MAE of the test set is reported. The inset shows the test set predictions of $R_e$ versus the true values. The values shown are the average of predictions from 1000 MC sampled training/test splittings. The GPR model gives predictions of the test and training sets. Shown are the mean and standard derivation of each molecule's predictions when used as training data (green symbols) and test data (orange symbols).

Inspired by the concept of molecular periodicity (as discussed in Ref.[71] and related references), we incorporate the group, denoted as $g_k$, and the period, denoted as $p_k$, of each atom within a molecule (where $k$ can be 1 or 2) as input features for a GPR model.

This model is employed to predict various combinations of spectroscopic constants, including $R_e$, $\omega_e$, and $\log\left(\frac{D_0}{R_e^3 Z_1 Z_2}\right)$, as elaborated upon in Section 3.2. Furthermore, training sets are permuted before being utilized by the learning algorithm to ensure that the GPR models maintain permutational invariance. This ensures that the relevant properties remain unaffected when two atoms in a molecule are interchanged.

The performance of our GPR model in predicting ground state $R_e$, based on the input features $(g_1, g_2, p_1, p_2)$, is depicted in Fig. 26. This Figure presents the MAE associated with each distinct type of molecule. The majority of molecules are accurately described by the GPR model, with the exception of transition metal-metal and bi-alkali molecules, as evident in the inset of Fig. 26.

To further quantify the performance of the GPR model, we calculate the RMSE of predicted $R_e$ based on 1000 randomly selected test sets. This results in an RMSE of 0.0968 $\pm$0.0070 Å(Table 1), along with relative error $r_E$ of 2.80 $\pm$0.20%. These results demonstrate that our model's performance improves as the number of molecules in the training set $N$ increases. This is clearly illustrated in the learning curve in panel (a) of Fig. 27. Notably, the model's performance has not yet converged for $N = 231$, indicating that further improvement can be achieved by incorporating additional data into the training set.

In the quest to predict $\omega_e$, we have identified $(R_e^{*-1}, g_1^{iso}, g_2^{iso}, p_1, p_2, \bar{g})$ as the most effective combination of features. Here, $R_e^*$ represents the predicted equilibrium distance derived from $(g_1, g_2, p_1, p_2)$, while $g_k^{iso}$ encodes information regarding the hydrogen isotopes of the $k$-th atom in the molecule. Additionally, $\bar{g}$ represents the average of the groups of the two atoms involved. However, we observe significantly improved performance when using the actual $R_e$ value. The GPR model's performance is shown in the inset of Fig. 28, where it is evident that the predicted values closely align with the true values. Indeed, the MAE and RMSE for the test set stand at $46.7 \pm 0.6$ cm$^{-1}$ and $73.4 \pm 0.2$ cm$^{-1}$, respectively, while $r_E$ is calculated to be $1.80 \pm 0.005$%, as indicated in Table 1.

Despite the impressive performance of the GPR model, certain molecules still pose challenges in accurate prediction, as illustrated in Fig. 28. Notably, these challenging cases encompass HF, DF, and HgH. The notable errors in predicting $\omega_e$ for HF and DF can be attributed to their unique bonding mechanisms compared to other halogen hydrides.

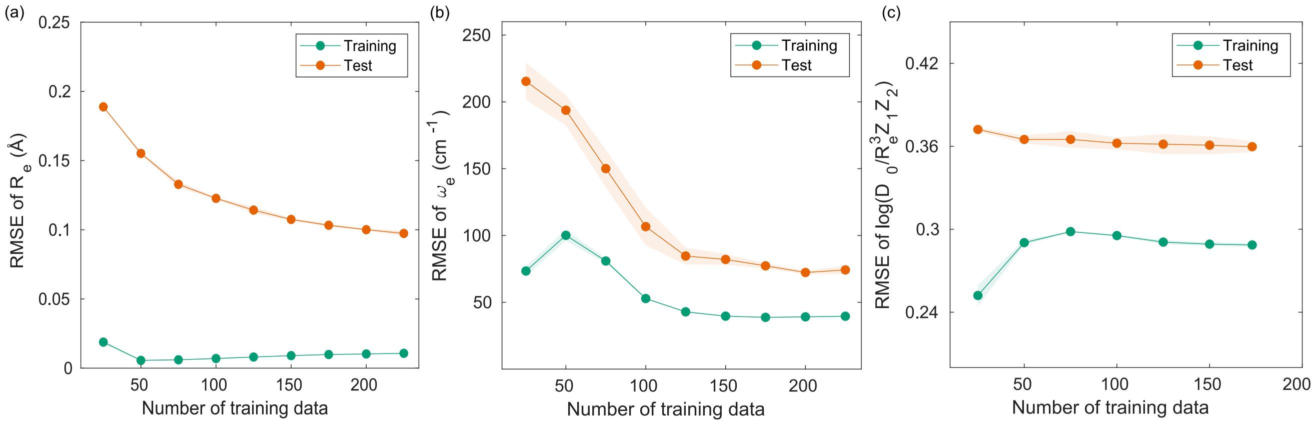


Figure 27: Performance of the GPR models as a function of the training set size $N$. (a) The learning curve of $R_e$ as a function of the size of the training set, predicted with the groups and periods of the two atoms, $(g_1, g_2, p_1, p_2)$. (b) Learning curve of $\omega_e$ as a function of the size of training set, using the equilibrium internuclear distance $R_e$, as well as the groups and periods and the average of groups of the two atoms $({R_e}^{-1},\ g_1^{iso},\ g_2^{iso},\ p_1,\ p_2,\ \bar{g})$ as the input feature. (c) Learning curve of $\log\left(\frac{D_0}{R_e^3 Z_1 Z_2}\right)$ as a function of the size of training set, using the equilibrium internuclear distance $R_e$, as well as the averages of groups and periods of the two atoms $(R_e, \bar{g}, \bar{p})$ as the input feature. The shade around the points denotes the variance of the errors regarding the MC method.

Table 1: Regression model predictions of $R_e$, $\omega_e$, and $D_0$. $g_i$ and $p_i$ represent the group and period of the $i$-th atom, respectively. $g_i^{iso}$ stands for the group encoding the information of isotopes of hydrogen, and $\bar{p}$, $\bar{g}$ are the average of groups and periods of the two atoms, respectively. $R_e^*$ is the GPR-predicted value from $(g_1, g_2, p_1, p_2)$.

| Property | Model | Feature | Test MAE | Test RMSE | Test $r_E$ (%) |
|---|---|---|---|---|---|
| $R_e$ (Å) | GPR | $(g_1, g_2, p_1, p_2)$ | $0.0662 \pm 0.0037$ | $0.0968 \pm 0.0070$ | $2.80 \pm 0.20$ |
| | LR | $\log(Z_1 Z_2)$ | $0.2605 \pm 0.0018$ | $0.3591 \pm 0.0006$ | $10.41 \pm 0.01$ |
| $\omega_e(\mathrm{cm}^{-1})$ | GPR | $(R_e^{-1}, g_1, g_2, p_1, p_2)$ | $126.7 \pm 2.1$ | $207.2 \pm 2.6$ | $5.07 \pm 0.06$ |
| | | $(R_e^{*-1}, g_1, g_2, p_1, p_2)$ | $152.5 \pm 3.6$ | $227.5 \pm 4.6$ | $5.56 \pm 0.11$ |
| | | $(R_e^{-1}, g_1^{iso}, g_2^{iso}, p_1, p_2)$ | $61.5 \pm 2.9$ | $142.8 \pm 7.0$ | $3.49 \pm 0.17$ |
| | | $(R_e^{*-1}, g_1^{iso}, g_2^{iso}, p_1, p_2)$ | $96.9 \pm 2.9$ | $176.0 \pm 13.1$ | $4.30 \pm 0.32$ |
| | | $(R_e^{-1}, g_1^{iso}, g_2^{iso}, p_1, p_2, \bar{p})$ | $67.5 \pm 1.0$ | $151.8 \pm 9.5$ | $3.71 \pm 0.2$ |
| | | $(R_e^{*-1}, g_1^{iso}, g_2^{iso}, p_1, p_2, \bar{p})$ | $101.8 \pm 5.4$ | $188.7 \pm 25.4$ | $4.61 \pm 0.62$ |
| | | $(R_e^{-1}, g_1^{iso}, g_2^{iso}, p_1, p_2, \bar{g})$ | $46.7 \pm 0.6$ | $73.4 \pm 0.2$ | $1.80 \pm 0.005$ |
| | | $(R_e^{*-1}, g_1^{iso}, g_2^{iso}, p_1, p_2, \bar{g})$ | $81.0 \pm 0.82$ | $121.8 \pm 0.8$ | $2.98 \pm 0.02$ |
| | LR | $\sqrt{Z_1 Z_2 e^{-2R_e}/m}$ | $376.5 \pm 6.6$ | $529.4 \pm 1.2$ | $12.95 \pm 0.03$ |
| | | $R_e^{-2}$ | $209.6 \pm 5.4$ | $297.3 \pm 1.4$ | $7.27 \pm 0.03$ |
| $\log \frac{D_0}{R_e^3 Z_1 Z_2}$ | GPR | $(R_e, \bar{g}, \bar{p})$ | $0.249 \pm 0.008$ | $0.357 \pm 0.007$ | $3.52 \pm 0.07$ |
| | | $(R_e^*, \bar{g}, \bar{p})$ | $0.270 \pm 0.006$ | $0.451 \pm 0.007$ | $4.45 \pm 0.07$ |
| | LR | $R_e$ | $0.833 \pm 0.004$ | $1.018 \pm 0.014$ | $10.03 \pm 0.14$ |

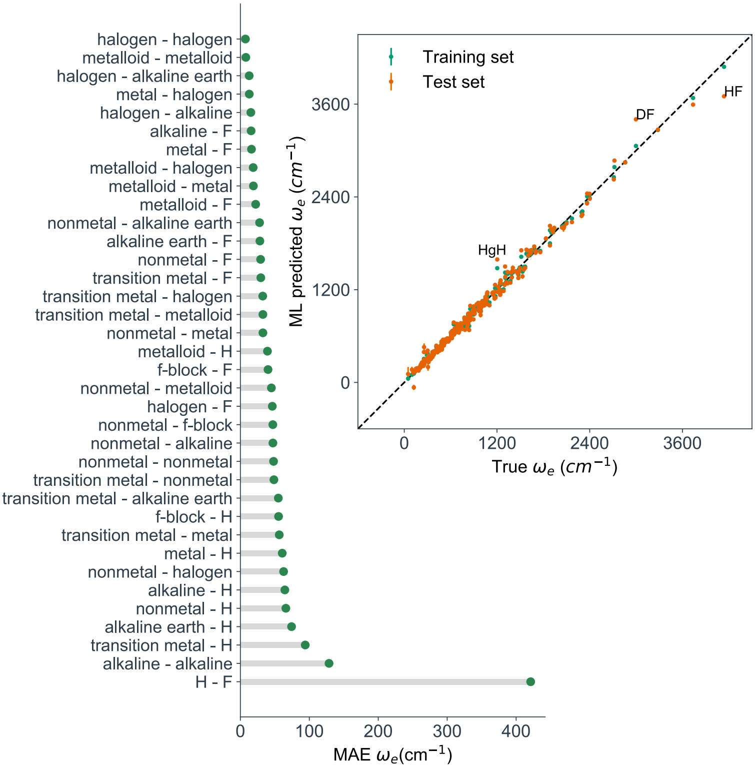


Figure 28: GPR performance based on the MAE predicting $\omega_e$ for molecules in the test set using $(R_e{}^{-1}, g_1^{iso}, g_2^{iso}, p_1, p_2, \bar{g})$ as input features classified by the types of the constituent atoms. The inset shows the test set predictions of $\omega_e$ compared with respect to the true values. The values shown are the average of predictions from 1000 MC sampled training/test splittings. The GPR model as learned from the training set gives predictions of the test and training set. Shown are the mean and standard derivation of each molecule's predictions when used as training data (green symbols) and test data (orange symbols).

Among the features $(R_e{}^{-1}, g_1^{iso}, g_2^{iso}, p_1, p_2, \bar{g})$, the introduction of the average of groups, denoted as $\bar{g}$ and defined as $\bar{g} = \frac{g_1+g_2}{2}$, proves to be instrumental in learning $\omega_e$. In particular, incorporating $\bar{g}$ results in a significant reduction of approximately 20% in the model's MAE compared to predictions solely using $(R_e{}^{-1}, g_1^{iso}, g_2^{iso}, p_1, p_2)$ as input features, as summarized in Table 1. Furthermore, the standard deviation of the MC training/test splitting predictions becomes notably smaller, indicating that the model exhibits greater robustness across various types of molecules within the dataset.

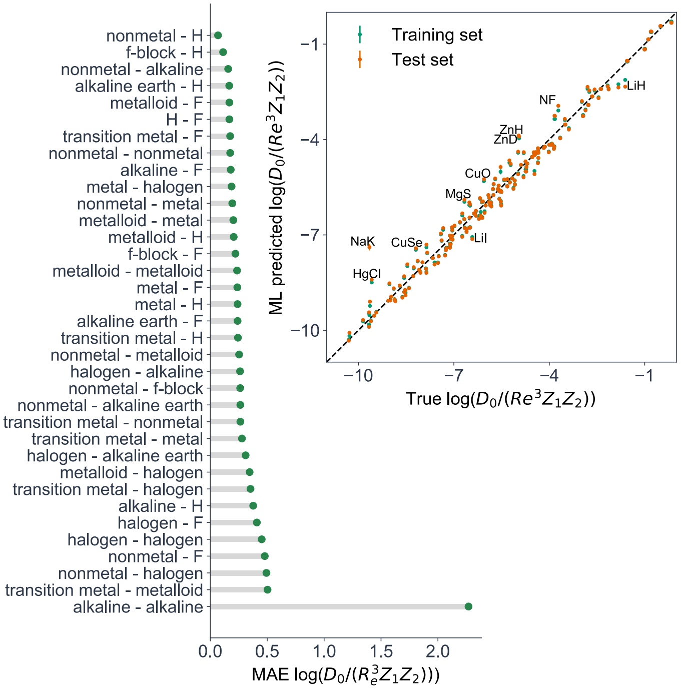


Figure 29: GPR performance on predicting $\log \frac{D_0}{R_e^3 Z_1 Z_2}$ using $(R_e,\ \bar{g},\ \bar{p})$ as input features classified by the types of the constituent atoms. In particular, the MAE of the test set is reported. The inset shows the test-set predictions of $\log \frac{D_0}{R_e^3 Z_1 Z_2}$ compared with respect to the true values. The values shown are the average of predictions from 1000 MC sampled training/test splittings. The GPR model gives predictions of the test and training set. Shown are the mean and standard derivation of each molecule's predictions when used as training data (green symbols) and test data (orange symbols).

Remarkably, the introduction of $\bar{g}$ yields the most significant improvements in the predictions for bi-alkali molecules, where the MAE can be reduced by a factor of three. While HF and DF remain challenging cases for the model, introducing $\bar{g}$ does lead to a 2-fold reduction in prediction errors for these molecules. On the contrary, introducing the average of periods, denoted as $\bar{p}$ and defined as $\bar{p} = \frac{p_1+p_2}{2}$, does not contribute to improving the model's performance. This suggests that $\omega_e$ depends on the total number of valence electrons of the two atoms rather than the number of electron shells.

Motivated by the pioneering work of Anderson, Parr, and colleagues [80, 103, 82, 83], we have delved into the prediction of $\log \frac{D_0}{R_e^3 Z_1 Z_2}$ using GPR. The results are presented in Fig. 29. In particular, the inset of the figure illustrates the GPR model's predictions of $\log \frac{D_0}{R_e^3 Z_1 Z_2}$ against its true values, demonstrating strong performance with an RMSE of $0.357 \pm 0.007$ and a $r_E$ of $3.52 \pm 0.07\%$, as detailed in Table 1. For this prediction, the GPR model is supplied with the input features $(R_e, \bar{g}, \bar{p})$. It exhibits rapid convergence concerning the size of the training set, with notable stability achieved around $N = 150$, as depicted in panel (c) of Fig. 27. The primary outlier is NaK, which is a van der Waals molecule. In this case, $D_0$ for NaK is overestimated. This discrepancy may be attributed to the unique nature of NaK as the sole bi-alkali molecule in the dataset possessing a $D_0$ value. Additionally, there are some outliers featuring first-row elements and 3d transition metals.

A comprehensive summary of the GPR models' performance in predicting various combinations of ground state spectroscopic constants is provided in Table 1. We compare the model's performance against the models proposed by Parr, Anderson, and colleagues [80, 103, 82, 83], as reviewed in Sec. 3.2.

Remarkably, our GPR model demonstrates superior performance compared to the linear model (denoted as LR in the table) based on specific functional forms of the electron density within the molecule. In some instances, the GPR model achieves a relative error that is five times better than the linear model. This highlights the effectiveness of utilizing the group and period (which are correlated with the number of valence electrons and the number of electron shells, respectively) of constituent atoms within a molecule as valuable indicators of spectroscopic constants, as opposed to relying on simple functional forms for the electron density, as employed in the models proposed by Parr, Anderson, and colleagues [80, 103, 82, 83].

It is worth noting that when predicting $R_e$ and $\omega_e$, the inclusion of the groups and periods of each atom in the molecule is essential. In contrast, for predicting $\log \frac{D_0}{R_e^3 Z_1 Z_2}$, the model performs well with only the average of the group and period of the two atoms involved. This suggests that $\log \frac{D_0}{R_e^3 Z_1 Z_2}$ is more strongly correlated with the additive properties of groups and periods for the two atoms rather than the differences between the two atoms arising from their distinct groups.

To further assess the generalizability of our ML approach, we have selected 26 molecules from the dataset that were unseen by the ML algorithm. These molecules include CoO[104], CrC[105], InBr[106], IrSi[107], MgD[108], MoC[109], NbC[109], NiBr[110], NiC[111], NiO[112], NiS[113], PbI[114], PdC[109], RuC[109], RuF[115], ScBr[110], SnI[114], TiBr[110], UF[116], UO[117], WC[118], YC[109], ZnBr[110], ZrC[109], ZrCl[119], and ZrF[119]. The GPR model's MAE in predicting the ground state $R_e$ for this additional test set is 0.066 Å, with an average relative error (defined as the absolute errors of each molecule divided by their true $R_e$) of 3.3%. Notably, for CrC, InBr, MgD, ZnBr, and ZrCl, the relative errors are less than 1%. Within this extra test set, experimental ground state $\omega_e$ values are available for 14 molecules: InBr, MoC, NbC, NiC, NiO, NiS, PbI, PdC, RuC, SnI, UO, WC, YC, and ZnBr. The GPR model achieves a MAE of 30 cm$^{-1}$ (4%) in predicting these values. For RuC and ZnBr, the relative errors are below 1%, and for NiS and MoC, the relative errors are below 2%. Additionally, for MoC, NbC, PbI, SnI, YC, and ZrC, where experimental binding energy data is available, the GPR model achieves a MAE of 0.32 eV (7.6%) in predicting $D_0$. Overall, our models perform quite well on this extra test set, demonstrating their robustness and generalizability.

Furthermore, we have checked some molecules in the above-mentioned extra test set whose multireference configuration interaction (MRCI) results are available. For MoC, it has been determined that $R_e = 1.676$ and $\omega_e = 1008 \pm 9$ cm$^{-1}$ experimentally in Ref. [109], MRCI results lead to $R_e = 1.693$ and $\omega_e = 971$ cm$^{-1}$ in Ref. [120], whereas our GPR model gives $R_e = 1.62 \pm 0.02$ and $\omega_e = 1020 \pm 10$ cm$^{-1}$, which leads to a more precise prediction of $\omega_e$. For RuC, the experimental $R_e = 1.6079$ and $\omega_e = 1102$ cm$^{-1}$ are found in Ref. [109], the MRCI calculation of Ref. [121] gives $R_e = 1.616$ and $\omega_e = 1085$ cm$^{-1}$, and our GPR model gives $R_e = 1.63 \pm 0.04$ and $\omega_e = 1111 \pm 8$ cm$^{-1}$. As a result, for some molecules, ML can be as accurate as MRCI. Therefore, although the ML models might not be as accurate as MRCI for all the molecules, considering that ML models can give predictions of thousands of molecules in one shot within seconds, they can be helpful in a rough screening of molecules for certain applications before the more expensive quantum chemistry calculations.

### 3.3.2 Learning the first excited state spectroscopic constants

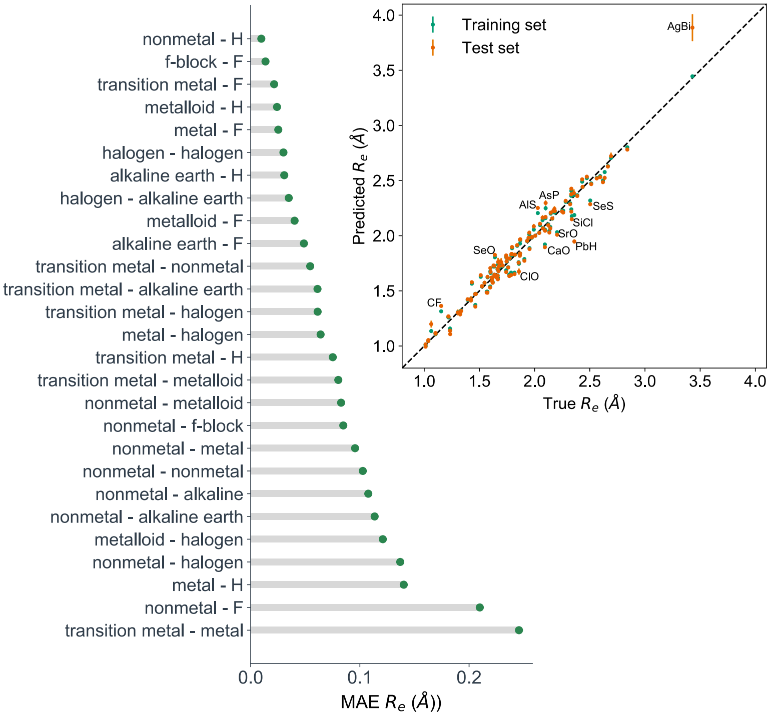


Figure 30: The test set MAE predicting A excited electronic state $R_e$ by GPR, using $(g_1, g_2, p_1, p_2, R_e(X), \mathrm{D(IP,EA)})$ as input features, classified by the types of the constituent atoms. The inset shows the test-set predictions of the A-excited electronic state $R_e$ compared with respect to the true values. The values shown are the average of predictions from 1000 MC sampled training/test splittings. The GPR model as learned from the training set gives predictions of the test and training set. Shown are the mean and standard derivation of each molecule's predictions when used as training data (green symbols) and test data (orange symbols).

To model the equilibrium internuclear distance $R_e$ of the A excited electronic state for different molecules, we have found that it is necessary to utilize atomic features of the two constituent atoms, including $g_1$, $g_2$, $p_1$, $p_2$, D(IP,EA), and the ground state $R_e(X)$ when constructing the GPR models. Interestingly, including D(IP, EA), which is defined as

$$\mathrm{D(IP,EA)} = \begin{cases} \mathrm{EA}_2 - \mathrm{IP}_1, & \text{if } \chi_1 < \chi_2 \\ \mathrm{EA}_1 - \mathrm{IP}_2, & \text{otherwise} \end{cases}$$

has proven to improve the predictions (Table 2). Here, $\mathrm{IP}_i$, $\mathrm{EA}_i$, and $\chi_i$ are the ionization potential, electron affinity, and electronegativity of atom $i$, respectively. Hence, D(IP,EA) provides a qualitative measure of the electron transfer between the two constituent atoms.

The resulting test-set MAE, RMSE, and $r_E$ for the A-state $R_e$ are $0.0691 \pm 0.0062$, $0.098 \pm 0.0097$, and $5.32 \pm 0.53$, respectively. The results are displayed in Fig. 30, showing similar outliers as those of the ground state $R_e$: transition metal-metal compounds. To predict $\omega_e$ for the A excited electronic state, it is found to be crucial to include not only the ground state $R_e^{-1}(X)$ but also the A state $R_e^{-1}(A)$.

Additionally, incorporating the ground state $\omega_e(X)$ as an input feature improves the model's performance. The outcomes are displayed in Fig. 31, where the combination of features ($\omega_e(X)$, $R_e^{-1}(X)$, $R_e^{-1}(A)$, $g_1$, $g_2$, $p_1$, $p_2$) yields an RMSE of $105.1 \pm 1.1$ and $r_E = 11.0 \pm 0.12\%$. Notably, including the average of groups $\bar{g}$ or isotope information does not provide further enhancement model performance, as this information is already encompassed within the ground state $\omega_e$.

A summary of our models' performance in predicting the spectroscopic constants for the A excited electronic state, including $R_e$ and $\omega_e$, is summarized in Table 2. Comparatively, the errors associated with predicting these properties for the excited state are roughly twice as large as those for the ground state. This suggests that predicting excited state properties is inherently more challenging, which may be attributed, in part, to more experimental uncertainties associated with excited states compared to ground states. Nevertheless, similar to ground state molecules, we observe a correlation between $\omega_e$ and the inverse of $R_e(A)$ for the A excited electronic state. This finding aligns with the historical notion of a relationship between $R_e$ and $\omega_e$ in the early days of molecular spectroscopy, as discussed in Section 3.2.

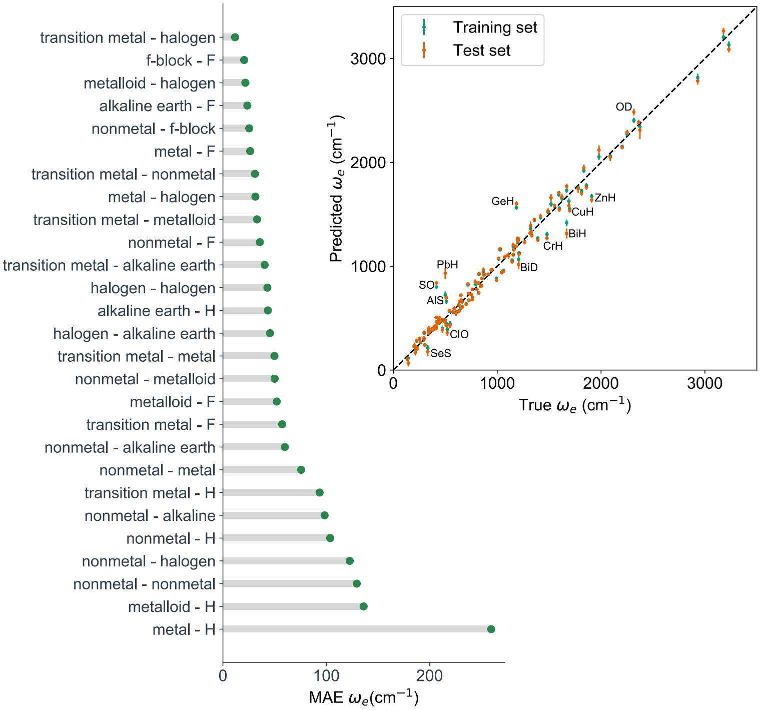


Figure 31: The test-set MAE predicting A excited electronic state $\omega_e$ by GPR, using $(\omega_e(X), R_e^{-1}(X), R_e^{-1}(A), g_1, g_2, p_1, p_2)$ as input features, classified by the types of the constituent atoms. The inset shows the test-set predictions of A-excited electronic state $\omega_e$ compared with respect to the true values. The values shown are the average of predictions from 1000 MC sampled training/test splittings. The GPR model as learned from the training set gives predictions of the test and training set. Shown are the mean and standard derivation of each molecule's predictions when used as training data (green symbols) and test data (orange symbols).

Table 2: Regression model predictions of the A excited electronic state $R_e$ and $\omega_e$. $g_i$ and $p_i$ are the groups and periods of the $i$-th atom, respectively, whereas $g_i^{iso}$ stand for the group encoding the information of isotopes of hydrogen. $\bar{p}$, $\bar{g}$ are the average of groups and periods of the two atoms, respectively. $R_e(X)$ and $R_e(A)$ refer to the ground state and A-state $R_e$, respectively. $\omega_e(X)$ refers to the ground state $\omega_e$.

| Property | Model | Feature | Test MAE | Test RMSE | Test $r_E$ (%) |
|---|---|---|---|---|---|
| $R_e$ (Å) | GPR | $(R_e(X)$, $g_1$, $g_2$, $p_1$, $p_2)$ | $0.0783 \pm 0.0018$ | $0.107 \pm 0.0026$ | $5.81 \pm 0.14$ |
| | | $(R_e(X)$, $g_1$, $g_2$, $p_1$, $p_2$, $D(IP,EA)$ | $0.0691 \pm 0.0062$ | $0.098 \pm 0.0097$ | $5.32 \pm 0.53$ |
| | | ) | | | |
| $\omega_e(\mathrm{cm}^{-1})$ | GPR | $(\omega_e(X)$, $R_e^{-1}(X)$, $R_e^{-1}(A)$, $g_1^{iso}$, $g_2^{iso}$, $p_1$, $p_2$, $\bar{g})$ | $71.8 \pm 1.4$ | $107.9 \pm 4.4$ | $11.3 \pm 0.46$ |
| | | $(\omega_e(X)$, $R_e^{-1}(X)$, $R_e^{-1}(A)$, $g_1^{iso}$, $g_2^{iso}$, $p_1$, $p_2)$ | $70.4 \pm 0.9$ | $105.1 \pm 1.5$ | $11.0 \pm 0.15$ |
| | | $(\omega_e(X)$, $R_e^{-1}(X)$, $R_e^{-1}(A)$, $g_1$, $g_2$, $p_1$, $p_2)$ | $70.6 \pm 0.9$ | $105.1 \pm 1.1$ | $11.0 \pm 0.12$ |

## 3.4 Conclusion

In summary, our study demonstrates that the GPR model can effectively reveal relationships between the main spectroscopic constants of diatomic molecules. This finding reaffirms the century-old vision of Kratzer and Mecke [56, 55]. These relationships are relatively independent of the nature of the chemical bond in diatomic molecules. Specifically, we have shown that using only the group and period of the constituent atoms within a molecule as input features, one can predict certain combinations of spectroscopic constants with an error of less than 5%. Therefore, our GPR models can efficiently learn from relevant datasets and accurately predict the values of spectroscopic constants. Moreover, we have demonstrated that GPR can also effectively learn spectroscopic relationships for excited electronic states of molecules, achieving an error of less than 11%.

It is worth noting that machine learning methods are sometimes perceived as mere fitting techniques or black-box algorithms from which it is challenging to gain meaningful insights. This perception is only partially accurate. As demonstrated in our study, we have been able to extract valuable insights from our machine-learning approach:

- It has been a common assumption that certain molecular properties can be qualitatively predicted based on the positions of the constituent atoms in the periodic table [122]. However, these predictions have typically been qualitative rather than quantitative. For example, one might be able to anticipate the type of bond in a molecule but not accurately predict its dissociation energy. Machine learning has allowed us to break through this limitation, demonstrating that it is indeed possible to make reasonably accurate quantitative predictions of spectroscopic constants by considering only the group and period of the constituent atoms.

- We have discovered that $\omega_e$ and $R_e$ exhibit a significant dependence on the number of valence electrons and the number of electron shells in the atoms that comprise a molecule. Simultaneously, the average count of valence electrons emerges as a vital factor in characterizing $\omega_e$. Furthermore, $\log \frac{D_0}{R_e^3 Z_1 Z_2}$ showcases a reliance on the average number of valence electrons and electron shells within the molecule.

- The potential to acquire knowledge about the properties of excited electronic states in diatomic molecules could pave the way for predicting Franck-Condon factors relevant to intriguing transitions. These transitions, in turn, hold promise for directly influencing the cooling processes of ultracold molecules [123, 124, 125, 126, 94].

Our approach has the potential to propel spectroscopy into the information era, which is particularly noteworthy, contributing to a more profound comprehension of spectroscopic properties. Additionally, our findings may offer valuable insights for developing features and geometric representations within material science. Along these lines, we have envisioned different points that can boost machine learning applications in spectroscopy and the quest for universal relationships between spectroscopic constants:

- Inclusion of more molecules. There exists a sizable pool of approximately 7,000 heteronuclear molecules, yet our GPR models solely rely on a subset of 256 among them. This choice is largely due to the limited availability of spectroscopic data, encompassing only about 3% of the potential heteronuclear diatomic molecules. This stark contrast underscores the vast untapped potential within the domain of diatomic molecule spectroscopy.

  As we gather more data, the predictive accuracy of our GPR models is poised to improve, even before it reaches the point of convergence in its learning curve. Moreover, it promises to enhance the collective knowledge base concerning the fundamental attributes of diatomic molecules. From our vantage point, this endeavor serves as a catalyst, stimulating data science-driven investigations in diatomic molecule spectroscopy.

  Based on this idea, it has recently been shown that more accurate GPR models can be obtained when using a dataset including both homogeneous and heterogeneous diatomic molecules with updated spectroscopic constants[127]. From this approach, we acquire fresh insights into the various forms of chemical bonding.

- Inclusion of input uncertainty. In the standard GPR models, the inputs are considered to be accurate and do not contain any measurement of their uncertainty. The output values are subject to some random noise following a Gaussian distribution, characterized by a constant variance, i.e. the noise level is assumed to be the same across all data points.

  However, in reality, as in the cases of our works on spectroscopic constants, the measurements are rarely perfectly precise, and the input points are associated with different uncertainties. Therefore, it can be helpful to include the input uncertainty during regressions. Indeed, it has been shown in [128] that the performance of the regression models can be improved when the input points are not treated as a fixed value, but instead as a random variable following a Gaussian distribution with approximations.

- Causal machine learning. In the current models, we learn the relationships between spectroscopic constants in relation to each other. Looking ahead, it could be intriguing to explore causal effects using advanced machine learning methods tailored for causal inference and find analytical expressions linking spectroscopic constants.

# 4

# A DATA-DRIVEN APPROACH TO DETERMINE DIPOLE MOMENTS OF DIATOMIC MOLECULES

## 4.1 Introduction

In the previous chapter, we have demonstrated the existence of universal relationships among spectroscopic constants. These relationships hold regardless of the specific nature of the molecular bond. However, despite being a fundamental molecular property, the electric dipole moment of molecules has not received significant attention in previous studies investigating connections among spectroscopic constants.

Only recently have researchers begun to explore the relationship between the dipole moment and molecular spectroscopic constants. Notably, Hou and Bernath have made noteworthy strides in understanding the dipole moment in the context of spectroscopic constants [129, 130]. Their findings have implications for the conventional expression of the dipole moment, denoted as $d$ and commonly taught in introductory chemistry courses, given by

$$d = qR_e, \tag{15}$$

where $q$ is the effective charge and $R_e$ corresponds to the equilibrium bond length of the molecule. However, Eq. (15) does not adequately capture the underlying physics of the dipole moment in numerous molecules [129, 130]. Subsequent research by Hou and colleagues has revealed that the dipole moment of certain molecules can be predicted more accurately by considering the effective charge, which is determined through quantum chemistry calculations, and the spectroscopic constants of molecules.

In this work, we introduce a data-driven methodology for assessing dipole moments in diatomic molecules and examining their correlation with spectroscopic constants. We demonstrate that, by assembling the most comprehensive compilation of dipole moment data for diatomic molecules to date (to the best of our knowledge) into a dataset, we can predict the dipole moment of diatomic molecules with a relative error of less than 5% based on atomic and molecular properties. Fig. 33 shows the number of molecules in our dataset categorized by the types of constituent atoms [1].

Our findings indicate that predicting a molecule's dipole moment solely from atomic properties is not feasible, although this is achievable for spec-

[1] This chapter is written based on reference [21]: Xiangyue Liu, Stefan Truppe, Gerard Meijer, and Jesús Pérez-Ríos. A data-driven approach to determine dipole moments of diatomic molecules. *Phys. Chem. Chem. Phys.*,22:24191–24200, 2020, https://doi.org/10.1039/D0CP03810E, with permission from the Royal Society of Chemistry.

troscopic constants, as demonstrated in the previous chapter. Instead, it is imperative to incorporate molecular characteristics.

### 4.1.1 An overview on the nature of the electric dipole moment of molecules

The nature of the electric dipole moment in molecules is a long-standing subject in quantum chemistry that has captivated the chemical physics community for nearly a century. The earliest attempt to elucidate the essence of the electric dipole moment in molecules can be attributed to Linus Pauling in the 1930s [131]. Specifically, after thoroughly examining hydrogen halide molecules, Pauling postulated that a molecule's dipole moment is intricately linked to the prevalence of its ionic structure compared to its covalent structure at the molecule's equilibrium bond length.

According to this model, the dipole moment arises as a consequence of the charge transfer occurring between the atoms within the molecule. Hence, the greater the extent of charge transfer, the larger the dipole moment becomes. The quantification of this charge transfer is achieved through the measure of ionic character (IC), which is defined by

$$\mathrm{IC} = \frac{d}{eR_e}. \tag{16}$$

Here, $e$ represents the electron charge. By comparing Eqs.(16) and (15), it becomes evident that the ionic character is analogous to the effective charge, denoted as $q$, positioned at the center of each of the atoms composing the molecule, as defined by Eq.(15). However, it should be noted that Pauling's model does not anticipate a 100% ionic character for molecules entirely ionic, such as alkali metal halides. Despite the slight deviation of Pauling's model when predicting dipole moments, it is important to underscore that Pauling recognized that a molecule's dipole moment must be intricately linked to other molecular properties through the molecular bond.

The next significant advancement came with the introduction of a novel concept known as the homopolar dipole moment, denoted as $d_h$, pioneered by Mulliken. Specifically, Mulliken recognized that due to the varying sizes of atomic orbitals, their overlap results in a charge displacement relative to the midpoint of the equilibrium bond length, influencing the molecule's electric dipole moment [132]. Furthermore, Mulliken observed that the asymmetry in the charge distribution of hybrid orbitals gives rise to what is termed the atomic dipole moment, represented as $d_a$.

The models of Mulliken and Pauling were summarized and further expanded upon by Coulson [133], who proposed the ultimate expression for the dipole moment of a diatomic molecule as

$$d = eR_e + d_a + d_h + d_p, \tag{17}$$

where $d_p$ represents the contribution arising from the polarization of atomic orbitals to the molecule's dipole moment. It is important to note that while Eq. (17) offers greater precision compared to Eq. (15), it necessitates input from quantum chemistry calculations. For a comprehensive overview of the Pauling and Mulliken models, we recommend consulting the extensive review by Klessinger [134].

The models proposed by Pauling and Mulliken have long been embraced by the physical chemistry community and incorporated into introductory chemistry courses, despite the fact that neither model provides a fully satisfactory explanation. Recently, Hou and Bernath [129, 130], following an examination of experimentally determined dipole moments for a wide range of molecules and employing quantum chemistry calculations, have put forward a new perspective regarding the electric dipole moment of a molecule, suggesting that it should be expressed as

$$d = qR_d. \tag{18}$$

In this expression, $q$ denotes the effective charge, and $R_d$ represents an effective length, which is contingent on fundamental spectroscopic constants of the molecule, with the condition that $R_d$ is less than $R_e$. Both Eq. (18) and Eq. (17) hinge on quantum chemistry calculations, specifically relying on outcomes derived from a natural bond orbital analysis. Therefore, the electric dipole moment of diatomic molecules remains without a completely satisfactory and precise elucidation in terms of fundamental spectroscopic constants.

## 4.2 The dataset

The dataset utilized in this study comprises ground-state dipole moments of 162 polar diatomic molecules, with 139 of them featuring both information on the equilibrium bond length, denoted as $R_e$, and the harmonic vibrational frequency denoted as $\omega_e$. The detailed dataset is given in Table 12 of the Appendix, and it represents the most extensive compilation of dipole moments for diatomic molecules to our knowledge.

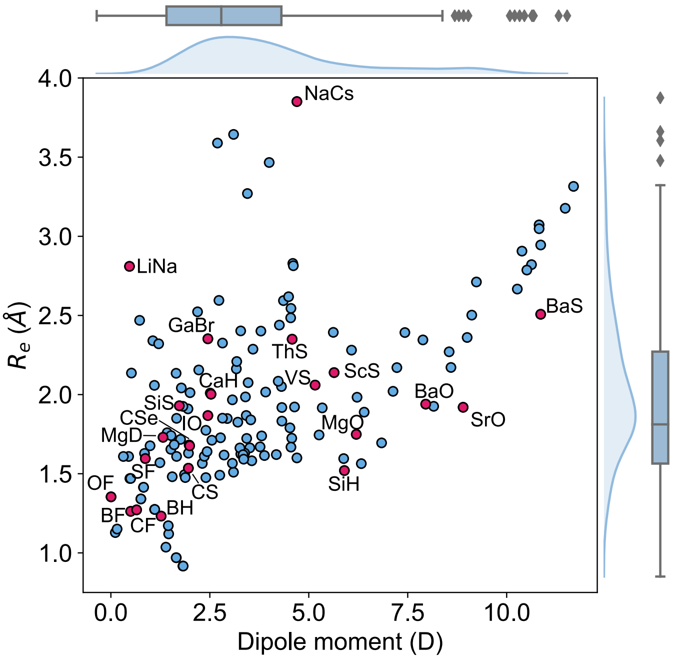


Figure 32: The equilibrium bond length $R_e$ versus the electric dipole moment of the molecules in the dataset. The blue-filled circles are the molecules that can be learned by the GPR model in this work. The red-filled circles indicate the molecules that can hardly be described by the GPR model in this work. These molecules are labeled by their chemical formula. The density in the right part and upper part of the figure shows the kernel density distribution of $R_e$ and dipole moments, respectively. The box plot shows the minimum, the maximum, the sample median, and the first and third quarterlies of $R_e$ (right) and dipole moments (top). Figure reproduced from ref. [21].

For a more comprehensive understanding of the dataset's characteristics, we present in Fig. 32 a graphical representation of the equilibrium bond length, $R_e$, versus the electric dipole moment of these diatomic molecules. The density plots and box plots provide insights into the distribution of $R_e$ (on the right) and the dipole moment, $d$ (at the top). The equilibrium bond lengths of the molecules span a range from 0.9 to 3.9 Å, with a median value of approximately 1.5 Å. Most of the molecules exhibit equilibrium bond lengths between 1.2 and 3.2 Å. Regarding the dipole moment values in the dataset, they span from 0.0043 Debye to 11.69 Debye, with a median value of roughly 2.45 Debye. This diversity in dipole moments reflects the broad spectrum of molecules encompassed in the dataset.

The dataset can also be classified based on the types of atoms comprising the molecules, as illustrated in Fig. 33. This figure reveals that most molecules in the dataset exhibit a highly ionic bond, formed between a transition metal and a nonmetal atom. The second most prevalent category of molecules consists of combinations involving a halogen atom and an alkaline metal atom, resulting in an ionic bond. The remaining molecules display a range of bond types, spanning from partially ionic to highly ionic, highlighting the dataset's remarkable diversity.

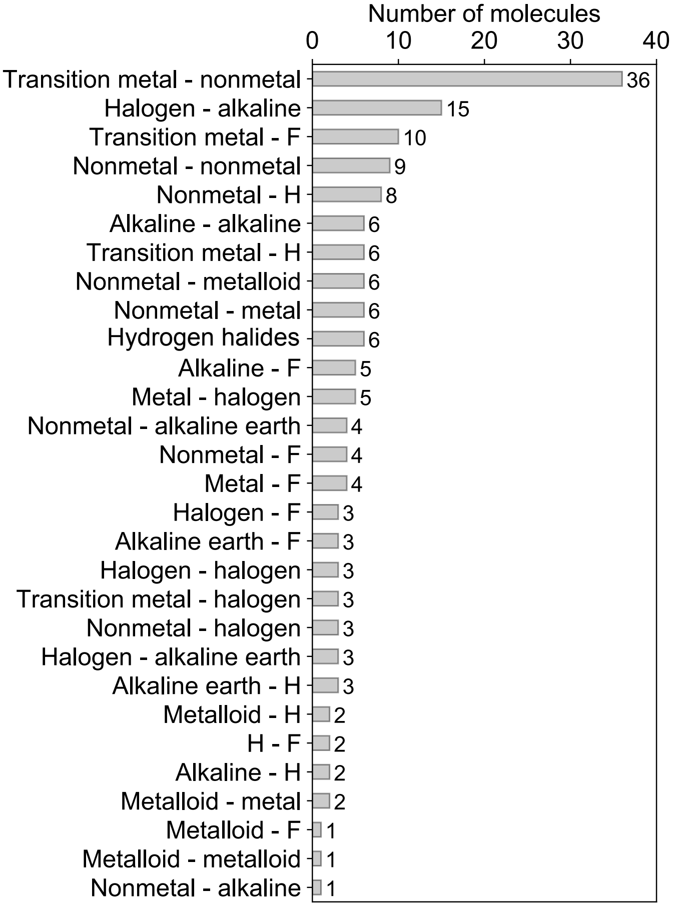


Figure 33: Molecules in the whole dataset classified by the types of their constituent atoms. Figure reproduced from ref. [21].

## 4.3 Machine learning approach

### 4.3.1 Gaussian process regression

Discovering relationships between the dipole moment and spectroscopic constants can be approached as a regression problem. In this context, the objective is to establish a mapping from the input atomic and molecular features, denoted as $\mathbf{x}$, to the target property, which, in this case, is the electric dipole moment. This mapping is represented by a function, denoted as $y = f(\mathbf{x})$. Similar to the approach introduced in Sec.3.1, we employ Gaussian Process Regression (GPR) to approximate the function $f(\mathbf{x})$.

### 4.3.2 Model evaluation

To learn dipole moments, the dataset is partitioned into training and test sets with a Monte Carlo (MC) approach, in the same way as introduced in Sec. 3.1. In this data-driven approach, GPR learns the connection between the input features and dipole moments by observing the training set, while the predictive accuracy of the GPR models is assessed using the test set. In this study, a total of 20 molecules are designated for the test set, with the remainder assigned to the training set. The assessment of the GPR model's performance is conducted using the mean absolute error (MAE), the root mean square error (RMSE), and the normalized error, $r_E$, as introduced in Sec. 3.1.

## 4.4 Prediction of dipole moments

### 4.4.1 Performance of the dipole moment models

A GPR has been built to predict the dipole moments of diatomic molecules, utilizing a variety of atomic and molecular properties as input features. Specifically, we have considered the following atomic properties: electron affinity (EA) obtained from references [135, 136, 137], ionization potential (IP) sourced from reference [138], electronegativity ($\chi$), and polarizability ($\alpha$), both extracted from reference [135]. These properties are chosen due to their relevance to the intrinsic chemical nature of the dipole moment, associated with the polarity of a molecular orbital in molecular-orbital bond theory or the ionic character of the molecular bond within valence-bond theory [133]. Additionally, we have incorporated molecular properties, such as the reduced mass ($\mu$), equilibrium bond length ($R_e$), and harmonic vibrational frequency ($\omega_e$).

Table 3: GPR Predictions on the ground-state dipole moments. $g_i$, $p_i$, $\mathrm{EA}_i$, $\mathrm{IP}_i$, $\chi_i$, $\alpha_i$ are groups, periods, electron affinity, ionization potential, electronegativity and polarizability of the atom $i$, respectively. $\mu$ is the reduced mass of a molecule. For these results, we employ 118 from the dataset out of the 139 molecules having values for both $R_e$ and $\omega_e$.

| Feature | Test RMSE (D) | Test MAE (D) | Test $r_E$ (%) |
|---|---|---|---|
| $(\mathrm{EA}_1,\mathrm{EA}_2,\mathrm{IP}_1,\mathrm{IP}_2,\ \sqrt{\mu R_e \omega_e^2})$ | $0.56 \pm 0.02$ | $0.43 \pm 0.0004$ | $4.8 \pm 0.1$ |
| $(\chi_1,\ \chi_2,\ \sqrt{\mu R_e \omega_e^2})$ | $0.70 \pm 0.05$ | $0.52 \pm 0.03$ | $6.0 \pm 0.4$ |
| $(\mathrm{EA}_1,\mathrm{EA}_2,\mathrm{IP}_1,\mathrm{IP}_2,\ \chi_1,\ \chi_2)$ | $0.86 \pm 0.006$ | $0.65 \pm 0.02$ | $7.4 \pm 0.05$ |
| $(\mathrm{EA}_1,\mathrm{EA}_2,\mathrm{IP}_1,\mathrm{IP}_2)$ | $0.97 \pm 0.05$ | $0.74 \pm 0.05$ | $8.3 \pm 0.4$ |
| $(\mathrm{EA}_1,\mathrm{EA}_2,\mathrm{IP}_1,\mathrm{IP}_2,\ R_e)$ | $1.04 \pm 0.02$ | $0.81 \pm 0.04$ | $9.1 \pm 0.2$ |
| $(\chi 1,\ \chi_2,\ \alpha_1,\ \alpha_2)$ | $1.29 \pm 0.004$ | $1.01 \pm 0.007$ | $11.2 \pm 0.04$ |
| $(\chi_1,\ \chi_2)$ | $1.35 \pm 0.002$ | $1.05 \pm 0.009$ | $11.7 \pm 0.01$ |
| $(\sqrt{\lvert\chi_1 - \chi_2\rvert},\ \bar{\alpha},\ D_0^{-1})$ | $1.21 \pm 0.03$ | $0.96 \pm 0.03$ | $10.5 \pm 0.3$ |
| $(p_1,p_2,g_1,g_2,\ R_e)$ | $1.25 \pm 0.02$ | $0.94 \pm 0.04$ | $10.8 \pm 0.1$ |

The performance of our GPR models using different sets of features is summarized in Table 3, focusing on 118 out of the 139 molecules in the dataset for which both $R_e$ and $\omega_e$ values are available. To ensure the permutational invariance of the GPR models when swapping the two elements within a molecule (e.g., switching from molecule AB to BA), permutation of the training sets is employed.

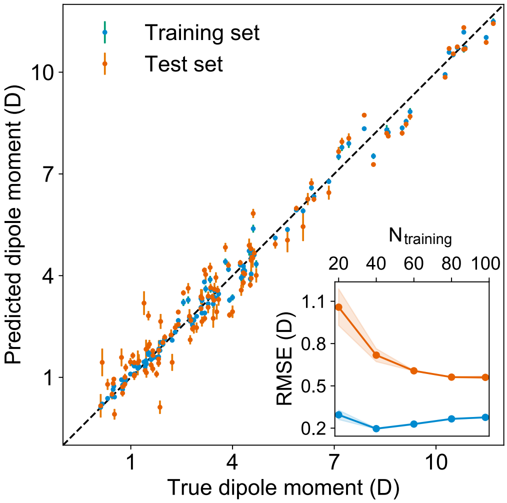


Figure 34: The GPR predictions of the ground-state dipole moments. The values shown in this figure are the average of predictions from 1000 MC sampled training/test splittings. The test set contains 20 molecules, while the training set contains 98 molecules. The mean and standard deviation of the predictions are shown for each molecule when they are used as training data (shown in blue) and test data (shown in orange). The inset shows the learning curve, which shows the training and test RMSE of the model with respect to the number of training data $N_{\text{Training}}$. The shade in the learning curve shows the variance of training/test RMSE, obtained for each point from an MC approach of 500 training/test splittings. Figure reproduced from ref. [21].

After exploring various combinations of atomic and molecular properties, we have determined that the GPR model best predicts the dipole moment of a diatomic molecule when utilizing the following input features: ($\text{EA}_1$, $\text{EA}_2$, $\text{IP}_1$, $\text{IP}_2$, $\sqrt{\mu R_e \omega_e^2}$). The performance of this model is visually represented in Fig. 34. The predicted values closely match the true values, exhibiting only a minor deviation that results in a normalized error $r_E$ of less than 5% (RMSE$=0.56 \pm 0.02$ Debye).

Furthermore, we have computed the learning curve for this GPR model, providing insight into the model's learning and generalization capabilities as a function of the training set size. The results are depicted in the inset of Fig. 34. The training RMSE and test RMSE are shown with respect to the number of training data points, denoted as $N_{\text{Training}}$. The shading in the learning curve illustrates the variance of training/test RMSE, derived from 500 training/test splits using a MC approach. As the number of training data points increases, the mean test error diminishes, demonstrating that the model's performance improves with more training data. Particularly, with 80 training data points, the learning curve approaches convergence, suggesting that further data from the same dataset would not significantly benefit the model's performance. The decreasing variance of the test RMSE as the number of training data points increases indicates the model's robustness and ability to be applied to different subgroups of molecules, with the variance eventually stabilizing at less than 0.02 Debye with 60 training data points.

In Chapter 3, we have demonstrated that it is possible to predict certain properties of diatomic molecules, such as $R_e$, $\omega_e$, and the binding energy, by using features derived from the groups and periods of the constituent atoms. However, it is important to note that the same features prove to be significantly less effective when attempting to predict the dipole moment of diatomic molecules. In this case, we observe test errors with a RMSE of $1.25 \pm 0.02$ Debye and a normalized error $r_E$ of $10.8 \pm 0.1\%$. This discrepancy in predictive performance suggests that the dipole moment is a more intricate property to model compared to the other spectroscopic constants of diatomic molecules. The dipole moment depends on a more complex interplay of factors, making it less amenable to simple feature-based predictions compared to properties like $R_e$ and $\omega_e$.

In Reference [139], it was demonstrated that the dipole moment of diatomic alkali–alkaline earth molecules can be empirically calculated using a formula involving the difference in electronegativity of the constituent atoms, $\sqrt{|\chi_1 - \chi_2|}$, the mean atomic polarizabilities, $\bar{\alpha} = (\alpha_1 + \alpha_2)/2$, and the dissociation energy, $D_e$. Building upon this idea, we have extended it through a Gaussian Process Regression (GPR) model by employing ($\sqrt{|\chi_1 - \chi_2|}$, $\bar{\alpha}$, $D_0^{-1}$) as the input features. It is noteworthy that we substituted the dissociation energy, $D_e$, with the binding energy, $D_0$, in our model as it is more frequently tabulated.

Remarkably, even though the dataset lacks alkali-alkaline earth molecules, our model's performance results in a normalized error of $10.5 \pm 0.3\%$. This outcome suggests that some of the underlying physics governing the dipole moment of alkali-alkaline earth molecules applies to a broader range of diatomic molecules. This unexpected result underscores the underlying universality of the physics governing the dipole moment, extending its applicability beyond specific molecular combinations.

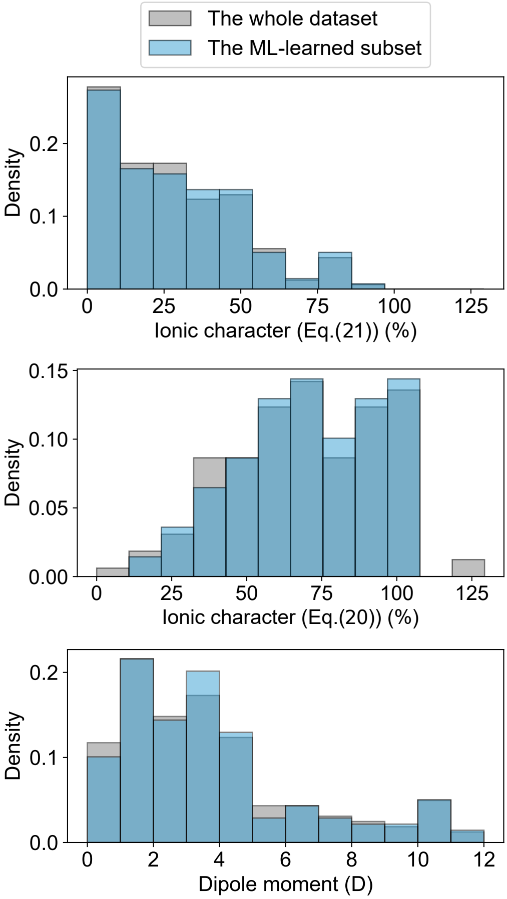


Figure 35: Comparison of the histograms of ionic characters and dipole moments in the whole dataset (shown in grey) and the ML-learned subset of 118 molecules (shown in blue). Panel (a) and (b) show the ionic characters calculated from Eqs. (21) and (20), respectively. Panel (c) plots the histogram of the dipole moment of the molecules. It is worth noticing that the dark blue regions appear in regions where the grey and light-blue bars overlap. Figure reproduced from ref. [21].

### 4.4.2 Interpretation of the input features

The remarkable performance of the feature set ($\mathrm{EA}_1$, $\mathrm{EA}_2$, $\mathrm{IP}_1$, $\mathrm{IP}_2$, $\sqrt{\mu R_e \omega_e^2}$) suggests that the conventional chemical perspective, wherein the difference in electronegativity between atoms in a molecule determines the ionic character of the molecular bond [133, 140, 131], is insufficient to fully characterize a molecule's dipole moment. When we introduce electron affinity and ionization potential as features, the predictive performance improves by 25%. However, it is only when we include $\sqrt{\mu R_e \omega_e^2}$ as a feature that the dipole moment can be predicted with an RMSE below 0.7 Debye. Therefore, our findings highlight the critical importance of including $\sqrt{\mu R_e \omega_e^2}$ as a feature when describing the dipole moment of a diatomic molecule.

It is worth noting that this particular feature is closely related to the derivative of the electronic kinetic energy, denoted as $T(R)$, at the equilibrium bond length, that, as demonstrated by Borkman in 1968 [74], from

$$-\left.\frac{dT(R)}{dR}\right|_{R=R_e} = \mu R_e \omega_e^2, \tag{19}$$

represents a force within the molecule. By equating this force to the pure electrostatic force, one can derive the value of $R_d$. Subsequently, using Eq. (18), it becomes possible to define the ionic character as

$$\mathrm{IC} = 100\left(d\sqrt{\mu R_e \omega_e^2}\right)^{1/2}, \tag{20}$$

where the value of IC is presented as a percentage. It is evident that IC does not have a direct dependence on the electronegativity differences of the atoms, which challenges the conventional understanding of chemistry.

The feature $\sqrt{\mu R_e \omega_e^2}$ was initially introduced by Hou and Bernath [129, 130] as an empirical relationship. In our study, we employ this feature to define the ionic character of a molecular bond.

Alternatively, the ionic character can also be defined based on the electronegativity difference between the two atoms constituting a molecule as

$$\mathrm{IC} = 16|\chi_1 - \chi_2| + 3.5|\chi_1 - \chi_2|^2, \tag{21}$$

following the work of Hannay and Smyth [140]. It is surprising to observe that Eqs. 20 and 21 yield distinct outcomes for the ionic character of the molecules contained in the database, as illustrated in Fig. 35. Indeed, the distribution of ionic character, as predicted by Eq. 21, appears to be the inverse of what is obtained from Eq. 20. This discrepancy can be attributed to the fact that the model proposed by Hou and Bernath (Eq. 20) consistently results in a higher ionic character when compared to the model by Hannay and Smyth.

### 4.4.3 Discussions on the outliers

The GPR model utilizing input features ($EA_1$, $EA_2$, $IP_1$, $IP_2$, $\sqrt{\mu R_e \omega_e^2}$) exhibits a presence of several outliers. In order to assess the significance of these outliers, we have conducted a comparison between the distributions of ionic character and dipole moment for the molecules, as depicted in Fig. 35 (displayed in grey), and the same magnitudes for the subset of 118 molecules that are amenable to learning within this study (displayed in blue). The subset of molecules learned via machine learning exhibits similar overall distributions of dipole moments and ionic characters when compared to the entire dataset. Consequently, it can be concluded that the outliers do not substantially alter the underlying distribution that the molecules follow.

Table 4: Outliers for learning the electric dipole moment of diatomic molecules. These molecules are labeled in Fig. 32 and classified with the types of constituent atoms and the molecular bonds.

| Type of bond | Molecule |
|---|---|
| Nonmetal-nonmetal | IO, CS, SiS, CSe |
| Nonmetal-F | SF, BF, CF, OF |
| Metal-halogen | GaBr |
| Alkaline earth-nonmetal | BaO, SrO, MgO, SrS, BaS |
| Alkaline earth-H | MgD, CaH |
| Metalloid-H | BH, SiH |
| Transition metal-nonmetal | VS, ScS, ThS |
| van der Waals | LiNa, NaCs |

Table 4 presents a categorization of the outliers based on their molecular bonds and constituent atoms. Additionally, we have computed the effective atomic charges for these molecules using a density functional theory (DFT) approach, as detailed in Table 5, employing various charge partitioning methods. These calculations were carried out using the B3LYP functional[141] and the def2-TZVP basis set [142, 143, 144], and were performed with the Gaussian 16 package [145]. Notably, we have observed that the natural bond orbital (NBO) method yields higher effective atomic charges compared to the Mulliken population for these outliers. Furthermore, all the molecules exhibiting an NBO charge exceeding 1.0 also demonstrate an ionic character exceeding 100% according to Eq. 20. Concerning the outliers within the van der Waals molecules, we have identified LiNa and NaCs. LiNa possesses the smallest equilibrium bond length ($R_e$) and dipole moment among the bialkali molecules in this dataset, while NaCs features the largest $R_e$ and dipole moment.

To comprehend the impact of various bonding types on dipole moments, we illustrate in Fig. 36 the relationships between $R_e$ and dipole moments for different categories of molecules in the current dataset, with outliers depicted as red circles. It is evident that the relationships between $R_e$ and dipole moments are contingent upon the type of molecule being considered.

Table 5: The effective atomic charges of the outliers with different charge partitioning methods, calculated with the B3LYP functional[141] and def2-TZVP basis set [142, 143, 144] with the Gaussian 16 package. [145]

| Molecule | Mulliken | Hirschfeld | NBO |
|---|---|---|---|
| MgO | 0.694 | 0.576 | 1.278 |
| SrO | 0.871 | 0.714 | 1.496 |
| BaO | 0.838 | 0.640 | 1.508 |
| BaS | 0.759 | 0.660 | 1.437 |
| BF | 0.099 | 0.073 | 0.549 |
| CF | 0.030 | 0.014 | 0.315 |
| OF | 0.017 | 0.012 | 0.063 |
| SF | 0.198 | 0.108 | 0.431 |
| MgD | 0.187 | 0.241 | 0.657 |
| CaH | 0.276 | 0.318 | 0.738 |
| BH | -0.036 | 0.072 | 0.349 |
| SiH | 0.048 | 0.122 | 0.349 |
| SiS | 0.231 | 0.222 | 0.656 |
| CS | -0.081 | -0.087 | -0.174 |
| SeC | 0.180 | 0.104 | 0.263 |
| IO | 0.412 | 0.214 | 0.625 |
| GaBr | 0.331 | 0.265 | 0.627 |
| ScS | 0.529 | 0.452 | 0.743 |
| VS | 0.425 | 0.247 | 0.343 |
| CsNa | 0.140 | 0.161 | 0.279 |
| NaLi | -0.074 | 0.001 | 0.007 |

As depicted in panel (a) of Fig. 36, there is a linear relationship between $R_e$ and dipole moments for metal-nonmetal molecules, wherein the nonmetal atoms belong to the same group in the periodic table. This linear behavior is akin to what has been observed for group IV/VI diatomic molecules in a previous study (Ref. [146]). In panel (b), focusing on the oxygen halides, we observe that $R_e$ increases nearly linearly with the dipole moment.

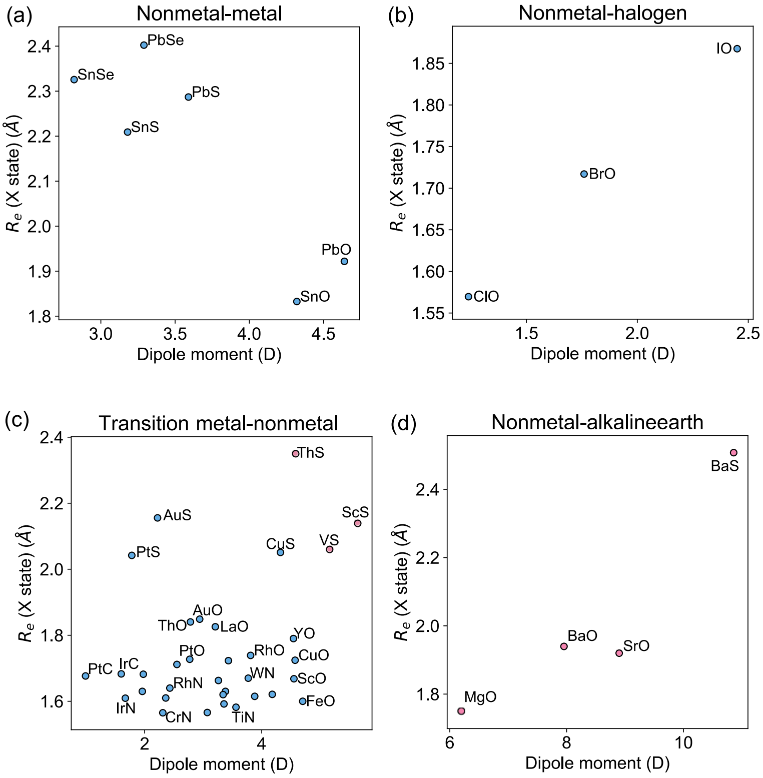


Figure 36: The equilibrium bond lengths $R_e$ as a function of dipole moments, classified by the type of the constituent atoms. The molecules that can be described by the GPR models from ($EA_1$, $EA_2$,$IP_1$,$IP_2$, $\sqrt{\mu R_e \omega_e^2}$) are shown in blue circles, while the outliers are shown in red circles. Figure reproduced from ref. [21].

In panel (c), we notice that molecules containing a transition metal and a nonmetal atom exhibit a different trend in equilibrium distance concerning the dipole moment compared to molecules formed by main-group metal elements and nonmetal atoms, as seen in panel (a). Within this category of molecules, the outliers are characterized by having both the largest dipole moments and the largest $R_e$ values in panel (c).

Remarkably, in panel (d) of Fig. 36, we find that all 4 alkaline earth-nonmetal molecules in the dataset are outliers. This observation aligns with an NBO population exceeding 1.0, as outlined in Table 5. Specifically, SrO, BaO, and BaS have the most substantial atomic charges among all the molecules in the dataset.

## 4.5 Conclusion

In summary, we have demonstrated that a GPR model can establish a relationship between the ground state dipole moments of diatomic molecules and their spectroscopic constants, specifically $R_e$ and $\omega_e$. We achieved accurate predictions of dipole moments, consistently with errors below 5%, all without the need for quantum chemistry calculations. This success is due to utilizing atomic features, encompassing electron affinity and ionic potential, combined with molecular spectroscopic constants, notably $\sqrt{\mu R_e \omega_e^2}$.

Furthermore, our study has revealed a significant departure from the commonly assumed notion in general chemistry, which asserts that differences in electronegativity between constituent atoms sufficiently describe the dipole moments of diatomic molecules. Instead, our data-driven approach has unveiled the intricate nature of dipole moments, highlighting their strong correlation with the fundamental essence of chemical bonding.

It is essential to underscore that the insights gained from our research have been made possible through the development of a comprehensive and unbiased dataset.

# 5

# BENCHMARKING THE ACCURACY OF CCSD(T) ON DIPOLE MOMENT OF DIATOMIC MOLECULES

## 5.1 Background

### 5.1.1 Motivation

Coupled cluster with single, double, and perturbative triple excitations (CCSD(T)) stands out as one of the most widely used methods in electronic structure theory. Notably, it serves as a benchmark reference for the development of various other electronic structure theory approaches, including density functional theory (DFT). CCSD(T) is valued for its size-consistency and its place within the coupled cluster family, making it amenable to systematic improvements. When coupled with specific corrections, CCSD(T) is recognized for its ability to approach sub-chemical accuracy, particularly in properties such as bond energies [14] at the complete basis set (CBS) limit [6].

Traditionally, benchmarking studies in the realm of DFT primarily concentrate on energetic properties [147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 3, 2, 1, 14, 4, 18]. Nevertheless, there has been a growing interest in recent times concerning the assessment of other aspects of the wavefunction, such as the electric dipole moment [157, 158, 11, 12, 19, 9].

Most literature about benchmarking dipole moments relies on assessing the performance of CCSD(T), primarily focusing on molecules composed of light main-group elements. However, in modern applications such as catalysis and materials synthesis, molecules containing elements from the third row and beyond ($Z > 18$), especially transition-metal compounds, play a vital role due to their electronic and magnetic properties [159, 160, 161]. Many of these applications demand an accurate description of energetic properties and electron densities, driving the increasing popularity of CCSD(T) among various quantum chemistry methods. Nevertheless, CCSD(T) primarily relies on single Slater-determinant Hartree-Fock references, which can pose challenges in systems with multi-reference characteristics [162, 1]. Another concern is the accuracy of approximations that can be applied in CCSD(T) calculations. For instance, the frozen-core approximation is a commonly used strategy when dealing with heavy elements in calculations. However, the computational cost associated with core-core and core-valence correlations in CCSD(T) calculations becomes impractical as the number of electrons in systems with effective core potentials (ECPs) increases.

Hence, given the central role of CCSD(T) in modern quantum chemistry and its significance in benchmarking other computational chemistry methods, it becomes essential to scrutinize its performance. In particular, benchmarking CCSD(T) against available experimental data on spectroscopic constants or molecular properties, such as dipole moments, is a necessary step. Fortunately, the availability of reliable experimental data concerning spectroscopic and molecular properties is expanding. In this context, diatomic molecules, though small, prove to be highly effective model systems for benchmarking, as they exhibit a diverse range of bonding and spin configurations that can reflect trends observed in polyatomic systems [151]. Indeed, experimental diatomic datasets have gained prominence as valuable choices for various DFT and wavefunction methods in benchmarking studies concerning equilibrium geometries and bond energies [147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 3, 2, 1, 14, 4, 18]. However, investigations focusing on the performance of CCSD(T) against experimental dipole moments are still relatively scarce [153, 10].

The present study aims to comprehensively examine the accuracy and limitations of CCSD(T) in predicting dipole moments of diatomic molecules. Initially, our investigation involves comparing the performance of CCSD(T) methods using different basis sets against recently collected experimentally measured dipole moments, as reported in Ref. [163]. Additionally, in line with the aforementioned objective, we assess the accuracy of CCSD(T) in predicting equilibrium bond lengths and harmonic frequencies, drawing from experimental data. The dataset comprises 32 diatomic molecules encompassing both main-group and transition metal elements, each characterized by diverse bonding characteristics. The dipole moments in this dataset are derived from well-controlled experimental measurements, such as microwave spectroscopy, with reported uncertainties typically below 0.05 D[1].

### 5.1.2 The basis sets

In practical quantum chemistry implementations, the unknown molecular orbitals are represented as linear combinations of basis functions, making the problem computationally solvable. Various types of functions can be used, such as polynomial or exponential functions, plane waves, wavelets, etc.

[1] This chapter is written based on reference [164]: Xiangyue Liu, Laura McKemmish, and Jesús Pérez-Ríos. The performance of CCSD(T) for the calculation of dipole moments in diatomics. *Physical Chemistry Chemical Physics*, 25(5):4093–4104, 2023, https://doi.org/10.1039/D2CP05060A, with permission from the Royal Society of Chemistry.

Using a *complete* basis set would involve no approximation in the expansion. However, in reality, an infinite number of basis functions would be required, which is not feasible. Therefore, actual calculations use a finite set of basis functions due to the rapid increase in computational complexity with the number of functions. The selection criteria for basis functions include their ability to converge rapidly towards the complete basis set limit as the number of functions increases. This necessitates physically "correct" basis functions tailored to the specific problem. For this reason, Slater-type orbitals (STOs) were once commonly used. STOs consist of spherical harmonic functions with exponential dependence on the nucleus-electron distances, similar to the exact solutions of hydrogen atomic orbitals. Another option is Gaussian-type orbitals (GTOs). However, GTOs offer a less accurate description of the electronic structure compared to STOs. They lack the discontinuous derivative at the nucleus region and converge unrealistically fast to zero at long-range distances from the nucleus, leading to poor representations in both regions. Consequently, achieving the same accuracy as STOs with GTOs requires more basis functions. Despite this limitation, GTOs remain the preferred choice for non-periodic systems due to the easier and more cost-effective calculation of electron integrals.

As mentioned above, using the fewest basis functions possible is desirable to reduce computational costs. The minimum basis set involves only one function for each type of orbital (s-, p-, d-, etc.). For example, a minimum basis set for hydrogen consists of only one *s*-type function for the 1s electron. For sodium (Na), one can use three *s*-type functions to describe the 1s, 2s, and 3s orbitals, and two sets of *p*-type functions for the 2p and 3p orbitals. However, the accuracy of the minimum basis set is quite limited. To improve the accuracy, the functions in the minimum basis set can be doubled, creating a “double zeta ($\zeta$, DZ)” basis. Optimized DZ basis sets usually have different exponential parameters for functions describing each type of orbital to account for broken spatial symmetries in molecules. In the same way, the accuracy of the basis set can be further improved with “triple zeta (TZ)”, “quadruple zeta (QZ)” basis sets, etc. Additionally, higher angular momentum functions might be necessary to describe the polarized electron distribution better. For electron-correlated calculations, additional higher angular momentum functions become important to account for angular dependencies in electron correlations.

In the early stages of quantum chemistry, people used to take the optimized orbitals from atomic calculations as the basis functions for molecular calculations[165, 6]. Later, Almöf and co-workers explored the possibility of systematically improving basis sets to better account for the correlation energy in their seminal work on atomic natural orbital analysis. Encouraged by their findings, Dunning and co-workers developed a series of basis sets [6], which is the correlation consistent polarized cc-pVXZ basis, with X being the "cardinal number" showing the level of the basis set. The primary consideration behind developing correlation-consistent basis sets is properly describing the dynamical electron correlations [6]. It is because the accuracy of quantum chemistry methods on electron correlations, including the near-degeneracy (static) correlation effects, as well as the nondynamical and dynamical correlation effects, is limited by the form of basis functions. Dunning's basis sets are popular for their ability to extrapolate the (post-Hartree Fock) correlation energies from finite basis sets to the CBS limit, as the CBS is not practically applicable. It is achieved via a "consistent" treatment of correlation effects, as it is shown to be imperative to incorporate polarization functions in sets [6]. In the pioneering work in 1989, Dunning advocated the inclusion of all functions within a particular electron shell group, like 1s, 2p, 3d, 4f, as well as functions from higher groups. This approach resulted in sets of basis functions like (1d), (2d1f), (3d2f1g), or (4d3f2g1h) in the correlation-consistent basis sets [6]. This approach ensures that functions within the same set make comparable contributions to the correlation energy. As a result, the correlation energy is improved *systematically* with respect to the increase of the highest angular momentum in the basis set and approaches the CBS limit following an inverse power rule. The basis set convergence behaviour of self-consistent field (Hartree-Fock or DFT) energies has been later discussed by Jensen and co-workers [166, 8, 167], revealing a remarkably faster convergence in comparison to correlation energies.

In a different vein, Ahlrichs and co-workers proposed the Karlsruhe basis sets, particularly the segmented contracted def2-basis sets [168, 144, 169], with the aim of reducing the number of basis functions and, consequently, lowering computational costs. In this approach, the valence basis sets were optimized based on the SCF energy. The low-angular momentum polarization functions were optimized with the nearly degenerate excited states when possible. For higher polarization functions, crucial in correlation calculations, optimization was based on second-order Møller–Plesset perturbation theory (MP2) energy since the atomic HF energy is independent of polarization functions [170]. The optimizations were carried out following the principle of achieving an "energetically balanced" basis [171]. This means that the energy defect should be roughly equivalent for each angular momentum quantum number $l$. As a result, these basis sets accurately reproduce cc-pVQZ atomization energies with chemical accuracy.

## 5.2 Method

### 5.2.1 Computational approach

In this work, two basis set families have been employed to calculate the electric dipole moment: The augmented Dunning's weighted core-valence basis set aug-cc-pwCVT/QZ(-PP) basis [172, 6, 173, 174, 7, 5, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185], and the segmented def2-QZVPP basis set [168, 144, 169]. The aug-cc-pwCV basis includes the core-correlations and, therefore is expected to be very accurate. On the other hand, the def2-QZVPP basis is much cheaper than the aug-cc-pwCV basis. For elements with $Z > 36$, the effective core potentials have been employed.

The core-correlated CCSD(T) method implemented in the CFOUR package [186] has been used to calculate the dipole moment for diatomic molecules. Unrestricted Hartree-Fock (UHF) wavefunctions have been used as references. In the case of closed-shell molecules with def2-QZVPP basis, the results are provided via the Molpro package [187, 188].

### 5.2.2 Zero-point vibrational corrections

Experimental dipole moments often exhibit deviations from theoretical predictions at the equilibrium bond length due to the anharmonic nature of molecular interactions. Vibrational corrections have been established as crucial components in achieving accurate dipole moment calculations that align with experimental values [189, 190, 191, 192, 10]. The extent of their importance hinges on the anharmonicity exhibited by the potential energy curve underlying the molecular system. In the context of dipole moments, which are first-order molecular properties, it becomes crucial to account for the discrepancy between the equilibrium bond length (as determined by the potential energy curve) and the most probable interatomic distance defined by the ground state vibrational wave function.

Consequently, in addition to reporting the dipole moment at the equilibrium bond length, denoted as $\mu_e$, we compute the vibrational average dipole moment $\mu_0$. This can be achieved by numerically averaging the radial-dependent dipole moment in conjunction with the vibrational ground state wavefunction of the molecule under examination. Specifically, we employ a Discrete Variable Representation (DVR) approach to address the single-channel Schrödinger equation associated with the vibrational degrees of freedom over the Born-Oppenheimer potential energy curve obtained.

More precisely, for each molecule, we perform point-wise potential energy computations using a grid of data points spanning from 0.4 times the experimental equilibrium bond length ($R_e$(Exp)) to 3 times $R_e$(Exp) [2]. The spin states of the molecules are determined based on the potential energy curve, selecting the spin state with the lowest energy compared to the other possible states. These determined spin states align with experimental observations.

Subsequently, we extract the equilibrium bond length ($R_e$) and the harmonic vibrational frequency ($\omega_e$) by fitting the potential energy curve. The electric dipole moments ($\mu$) are computed by calculating the analytic gradients at each single-point geometry. This process yields both the dipole moment at the equilibrium bond length ($\mu_e$) and the dipole moment corrected for zero-point vibrational effects ($\mu_0$). The latter incorporates vibrational average corrections using the DVR method for the vibrational wavefunctions [3]. Additionally, the overlap is determined through numerical integration. In this study, we focus solely on the magnitude of the dipole moments and do not delve into discussions regarding their directional aspects.

### 5.2.3 The CBS extrapolation

For Dunning basis sets, the CBS limits are determined utilizing the conventional two-point extrapolation scheme [193, 194, 195]

$$\text{Predicted CBS}(n_1/n_2)^{\text{corr}} = \frac{n_1^3 d_1 - n_2^3 d_2}{n_1^3 - n_2^3}, \tag{22}$$

where $d_i$ represents the correlation contribution of the molecular property (either $\omega_e$ or $\mu$) evaluated at a specific basis set characterized by $n_i$. Specifically, for the aug-cc-pwCVTZ and aug-cc-pwCVQZ basis sets, the values of $n_1$ and $n_2$ are set to 3 and 4, respectively. However, there is no need for extrapolation when calculating $R_e$ since predictions at the quadruple-$\zeta$ level are already very close to convergence.

For elements computed with the aug-cc-pwCVT/QZ-PP basis sets, it is worth mentioning that relativistic effects have been considered by incorporating effective core potentials, which have been found to have minimal

[2] Specifically, we evaluate the potential energy at points -0.4$x$, -0.35$x$, -0.3$x$, -0.24$x$, -0.18$x$, -0.12$x$, -0.08$x$, -0.04$x$, -0.02$x$, 0, 0.02$x$, 0.04$x$, 0.08$x$, 0.12$x$, 0.2$x$, 0.28$x$, 0.36$x$, 0.45$x$, 0.5$x$, 0.6$x$, 0.8$x$, 1.0$x$, 1.5$x$, 2.0$x$, 3.0$x$, where $x = R_e(\text{Exp})$

[3] We employ 200 DVR points to ensure convergence of vibrational energies with an accuracy better than 0.1%.

Table 6: Molecules in the dataset classified by classes of their constituent elements.

| Classes of molecules | Molecules |
|---|---|
| Metal/metalloid-halogen | AlF ($X^1\Sigma^+$), GaF ($X^1\Sigma^+$), InCl ($X^1\Sigma^+$), InF ($X^1\Sigma^+$) |
| Metal/metalloid/nonmetal-metal/metalloid | GeTe ($X^1\Sigma^+$), GeO ($X^1\Sigma^+$), GeS ($X^1\Sigma^+$), PbO ($X^1\Sigma^+$), PbS ($X^1\Sigma^+$), SiO ($X^1\Sigma^+$), SiS ($X^1\Sigma^+$) SnO ($X^1\Sigma^+$), SnS ($X^1\Sigma^+$) |
| Nonmetal/halogen-halogen | BrO ($X^2\Pi_{3/2}$), CF ($X^2\Pi$), IBr ($X^1\Sigma^+$) |
| Nonmetal-nonmetal | CN ($X^2\Sigma^+$), CO ($X^1\Sigma^+$), CS ($X^1\Sigma^+$), CSe ($X^1\Sigma^+$), NO ($X^2\Pi_{1/2}$), PN ($X^1\Sigma^+$), PO ($X^2\Pi$), SO ($X^3\Sigma^-$) |
| Transition metal-halogen | AgBr ($X^1\Sigma^+$), AgF ($X^1\Sigma^+$), AgI ($X^1\Sigma^+$), CuF ($X^1\Sigma^+$), YF ($X^1\Sigma^+$) |
| Transition metal-nonmetal | HfO ($X^1\Sigma^+$), ScO ($X^2\Sigma^+$), ZrO ($X^1\Sigma^+$) |

impact [15]. Nevertheless, a correction for scalar relativistic effects has been applied using the second-order Douglas-Kroll-Hess approximation (DK), implemented in Molpro, for specific molecules when (aug-)cc-p(w)CVT/QZ-DK(3) basis sets are available [196].

### 5.2.4 Dataset

Experimentally, the equilibrium bond length, harmonic frequency, and dipole moment measurements have been conducted in over 100 diatomic molecules, as documented in Ref. [163]. We have carefully selected 32 representative molecules from this extensive pool for an in-depth exploration using high-level quantum chemistry techniques. These molecules were chosen to encompass a wide range of chemical diversity among diatomic species, as listed in Table 1.

To facilitate interpretation, we have categorized these molecules into six distinct classes depicted in Fig. 37. Throughout this chapter, we will employ this graphical representation to present our results consistently. Notably, this dataset encompasses a diverse array of main-group metal and non-metal compounds, showcasing both covalent and ionic bonds. Specifically, the dataset includes 8 transition metal compounds.

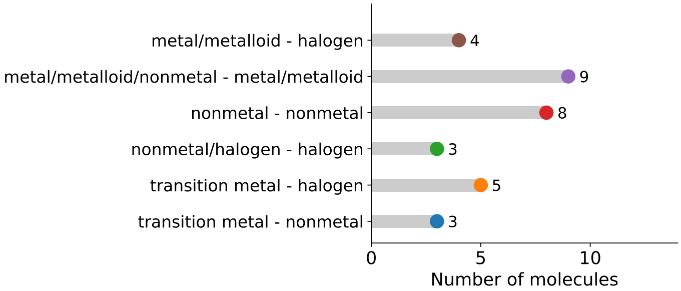


Figure 37: Number of molecules in the present dataset classified by classes of their constituent elements. Figure reproduced from ref. [164].

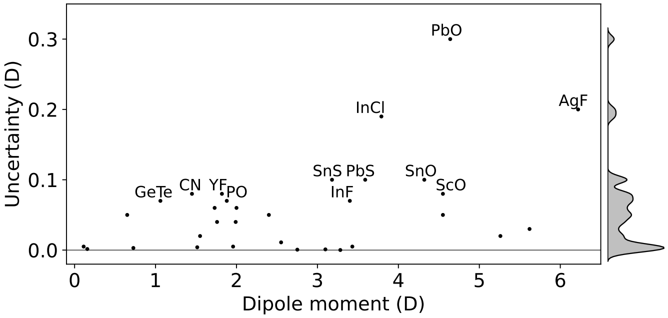


Figure 38: Uncertainties of experimental dipole moments of the molecules included in the dataset. Figure reproduced from ref. [164].

The accuracy of experimentally measured electric dipole moments hinges on precisely determining the magnitude of the applied electric field and its uniformity. In curating the present dataset, we have selected molecules for which dipole moments were determined through high-resolution spectroscopic methods, particularly microwave spectroscopy and molecular beam electric resonance. Molecules with substantial uncertainties in their dipole moments, such as BF and RbI, were excluded from consideration.

As illustrated in Fig. 38, this dataset predominantly features typical error bars ranging from 0.1% to 5%, translating to errors of less than 0.05 D for most of the molecules under investigation. However, a few molecules, namely PbO, AgF, InCl, PbS, SnS, and SnO, are denoted in the figure due to their uncertainties exceeding 0.1 D.

## 5.3 Results and discussions

The performance of CCSD(T) on the equilibrium bond length, $R_e$, and harmonic vibrational frequency, $\omega_e$, of diatomic molecules have been exhaustively investigated in the literature by comparing them with experimental measurements. Not only the accuracy of basis set families [180, 174, 172], but also relativistic pseudopotentials have been discussed and compared with full-electron relativistic treatments [177, 178, 179, 181, 182, 176, 175, 197]. Recently, it has been found that by using non-HF orbitals, the accuracy of CCSD(T) $\omega_e$ can be further improved for diatomic molecules consisting of row 2 and row 3 elements [18]. For transition metal diatomic molecules, the accuracy of CCSD(T) $R_e$ and $\omega_e$, as well as the influence of relativistic effect and multi-reference character have been investigated based on a dataset of 60 molecules [15]. Nevertheless, we first report the calculated $R_e$ and $\omega_e$ for the dataset employed in this work. Then, we present and discuss the theoretical predictions of dipole moments compared to their experimental values for the molecules in the dataset.

The performance of computational predictions in comparison to experimental data is assessed through:

- Residuals: the difference between the experimental value of a molecular property and its computed value, $x_i(\text{Exp.}) - x_i(\text{CCSD(T)})$, where $x_i$ is a property of molecule $i$.

- Root mean squared error (RMSE).

We report the results in two formats for each computational method explored and each molecular property under investigation. Firstly, we provide residual errors for all molecules, with special labeling for molecules exhibiting significant errors. Secondly, we conduct a statistical analysis of the errors by presenting the Root Mean Square Error (RMSE) categorized by molecular class, as outlined in Table 6. Additionally, when necessary, we report the relative error ($r_E$) as

$$r_E = \frac{1}{N}\sum_{i}^{N} \frac{|x_i(\text{Exp.}) - x_i(\text{CCSD(T)})|}{x_i(\text{Exp.})}. \tag{23}$$

Lastly, we present results utilizing both the aug-cc-pwCVQZ and def2-QZVPP basis sets. Notably, the def2-QZVPP basis set contains slightly over half the number of basis functions compared to the aug-cc-pwCVQZ basis set.

### 5.3.1 The accuracy of equilibrium bond lengths

When analyzing the equilibrium bond length ($R_e$), both the aug-cc-pwCVQZ and def2-QZVPP basis sets exhibit remarkably accurate predictions, with an RMSE $\lesssim$ 0.008 Å (and a relative error $r_E$ of around 0.2%), as depicted in Fig. 39. Based on this figure, it is clear that non-metal diatomic molecules, as well as metal/metalloid/non-metal-metal/metalloid compounds, present the smallest RMSE values of approximately 0.002 Å (with relative errors around 0.1-0.2%). On the contrary, molecules containing transition metals exhibit somewhat larger errors, with an RMSE of around 0.01 Å (resulting in a relative error of approximately 0.5%). Additionally, a few outliers are noticeable in this context, including CuF, AgBr, and AgI.

For CuF, Aoto et al. have demonstrated that the error in predicting $R_e$ can be mitigated by incorporating relativistic corrections [15]. Our findings corroborate these observations, as elaborated in Section 5.3.4. However, for certain other molecules containing transition metals, such as ScO or AgF, the correction for scalar relativistic effects appears to have a negligible impact. Another potential source of discrepancy may arise from multi-reference effects. Notably, previous studies have indicated that employing multi-reference coupled-cluster theory for molecules like ScO and AgF yields results similar to those obtained using single-reference methods [15].

There has been relatively limited research in silver-halogen molecules, except for AgF and AgCl, finding that relativistic effects on the equilibrium bond length prediction are negligible. Therefore, given these observations, relativistic effects alone may not be sufficient to account for the substantial errors in the predictions of $R_e$ for AgBr and AgI.

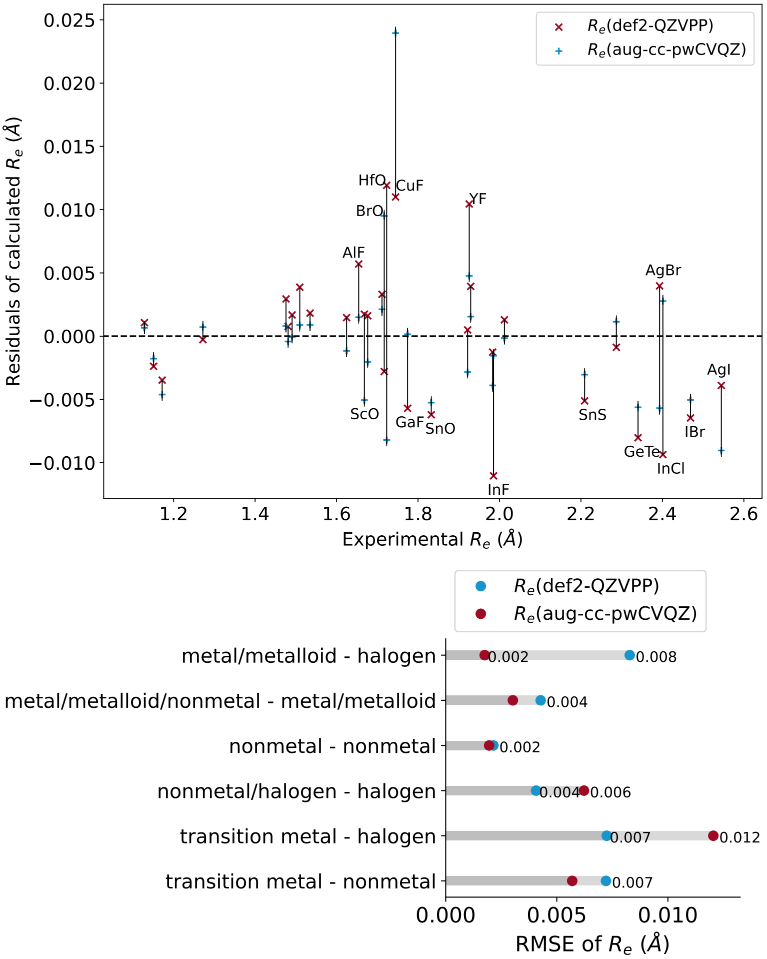


Figure 39: Calculated $R_e$ with def2-QZVPP (red symbols) and aug-cc-pwCVQZ (blue symbols) basis sets. Upper panel: residuals of the calculated $R_e$. Lower panel: RMSE of the computed $R_e$ for different classes of molecules. Figure reproduced from ref. [164].

In our calculations using the def2-QZVPP basis set, we have observed systematically larger equilibrium bond length ($R_e$) values that deviate further from experimental values when compared to the aug-cc-pwCVQZ basis set. Within the aug-cc-pwCV basis family, it is generally observed that increasing the number of basis functions tends to improve the predictions of $R_e$ and brings them closer to experimental values. When moving from aug-cc-pwCVTZ to the quadruple-$\zeta$ level, the Root Mean Square Errors (RMSEs) for most classes of molecules can be significantly reduced, often reaching approximately 50% improvement. Exceptions to this trend are the molecules containing transition metals, for which the aug-cc-pwCVT/QZ basis set yields RMSEs of a similar magnitude.

However, it is noteworthy that the reduction in the number of basis functions in the def2-QZVPP basis sets does not necessarily deteriorate the predictions of $R_e$. In fact, the RMSEs obtained with the def2-QZVPP basis set are generally comparable to those obtained with the aug-cc-pwCVQZ basis set for most molecules studied, with the exception of metal/metalloid halides. Particularly for transition metal halides, the def2-QZVPP basis set can yield predictions much closer to the experimental results than the aug-cc-pwCVQZ basis set.

### 5.3.2 The accuracy of vibrational harmonic frequencies

The performance of CCSD(T) in predicting the vibrational harmonic frequency ($\omega_e$) compared to experimental data is presented in Fig. 40. In this figure, the absolute error of CCSD(T) calculations and their RMSE are depicted for each molecular class. For most molecules, CCSD(T) exhibits $r_E \lesssim 2\%$. The RMSE values with the predicted CBS(aug-cc-pwCVT/QZ) and def2-QZVPP basis sets are 24.7 cm$^{-1}$ and 16.3 cm$^{-1}$, respectively.

CCSD(T) calculations for diatomic molecules involving main-group metal elements yield very accurate results with both basis sets, resulting in an RMSE of less than 10 cm$^{-1}$. These findings are consistent with previous research by Aoto et al., which indicated that diatomic molecules composed of main-group nonmetals exhibit substantial errors [15]. Notably, some of the most significant outliers in these cases include CN and NO, which are correlated with the use of the unrestricted Hartree-Fock (UHF) reference. Recent work has shown that these errors can be mitigated by employing non-HF references [18].

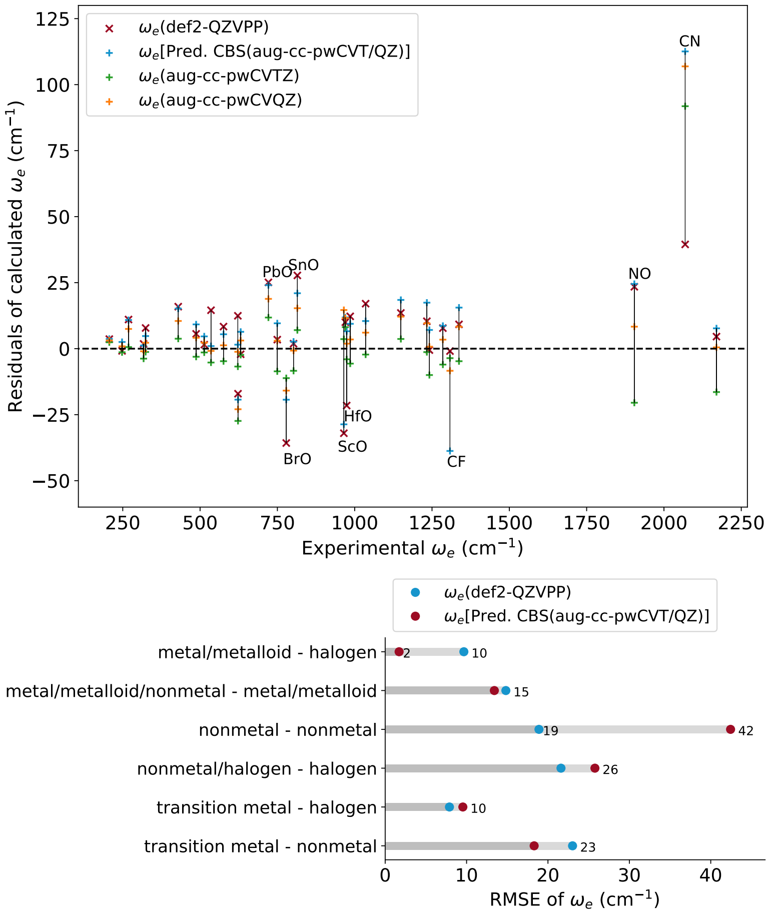


Figure 40: Calculated $\omega_e$ with def2-QZVPP (red symbols) and aug-cc-pwCVQZ (blue symbols) basis sets. Upper panel: residuals of the calculated $R_e$. Lower panel: RMSE of the computed $R_e$ for different classes of molecules. Figure reproduced from ref. [164].

For most molecules, the predictions of $\omega_e$ are already quite close to experimental values at the aug-cc-pwCVTZ level. The difference in RMSE between the aug-cc-pwCVTZ and aug-cc-pwCVQZ basis sets is only around 2.5 cm$^{-1}$. Similarly, the RMSE obtained with the def2-QZVPP basis set is very close to that of the aug-cc-pwCVQZ basis, and in some cases, it even slightly outperforms the latter. However, we notice that for nonmetal diatomic molecules, the use of the aug-cc-pwCV basis significantly improves the results.

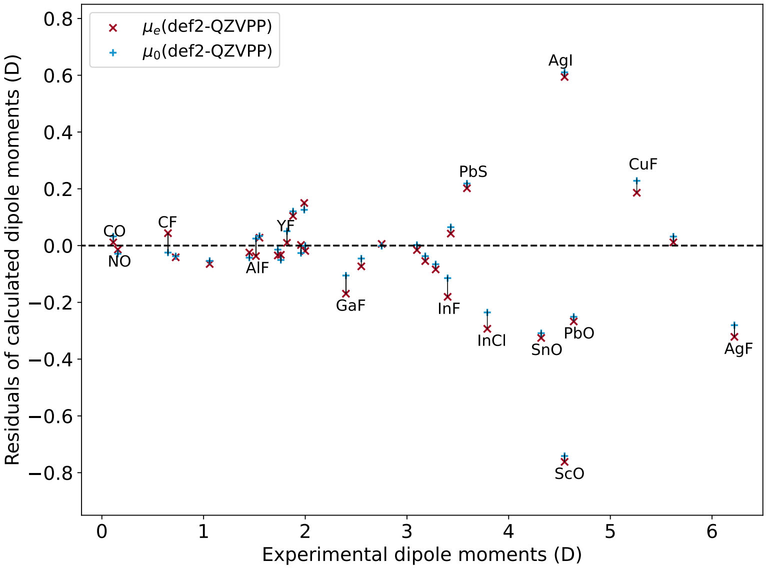


Figure 41: Dipole moment errors calculated with def2-QZVPP basis set versus experimental results, with or without vibrational average corrections. Figure reproduced from ref. [164].

### 5.3.3 The accuracy of dipole moments

#### Effect of methodology

**Vibrational corrections on dipole moment predictions** In Figs. 41 and 42, we observe the residuals of CCSD(T) calculations for the dipole moments of the molecules in our dataset using the def2-QZVPP and CBS(aug-cc-pwCVT/QZ) basis sets. These figures reveal that the vibrational averaging has a negligible impact on the dipole moments, yielding differences of approximately 0.01 D (2%) between the vibrational average and non-vibrational average dipole moments for both basis sets.

However, there are a few outliers, which include diatomic molecules with light elements and short bond lengths, such as CO (approximately 20%), NO (approximately 10%), CF (approximately 10%), AlF (approximately

4%), GaF (approximately 3%), and YF (approximately 2%). Notably, the difference introduced by the vibrational average correction in these outliers appears to correlate with their harmonic vibrational frequency.

Due to the slightly better performance of the vibrational average dipole moment, we will use $\mu_0$ when referring to dipole moments from now on.

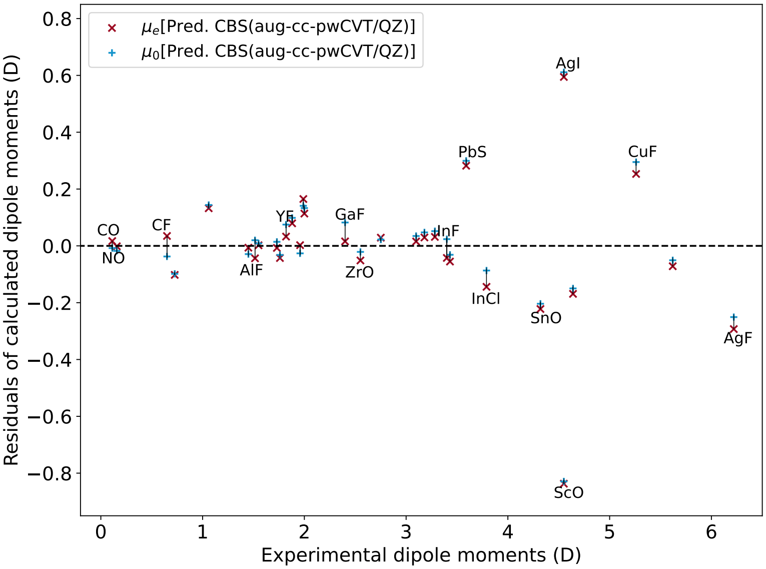


Figure 42: Dipole moment errors calculated with aug-cc-pwCVQZ basis set versus experimental results, with or without vibrational average corrections. Figure reproduced from ref. [164].

**Basis set family and size** A detailed study on the influence of basis sets on the dipole moments is presented in Figs. 43 and 44, where we examine the residuals of the calculated dipole moments in comparison to experimental values. Specifically, we investigate the performance of CCSD(T) dipole moments calculated with cc-pwCV, aug-cc-pwCV, and def2-QZVPP basis sets.

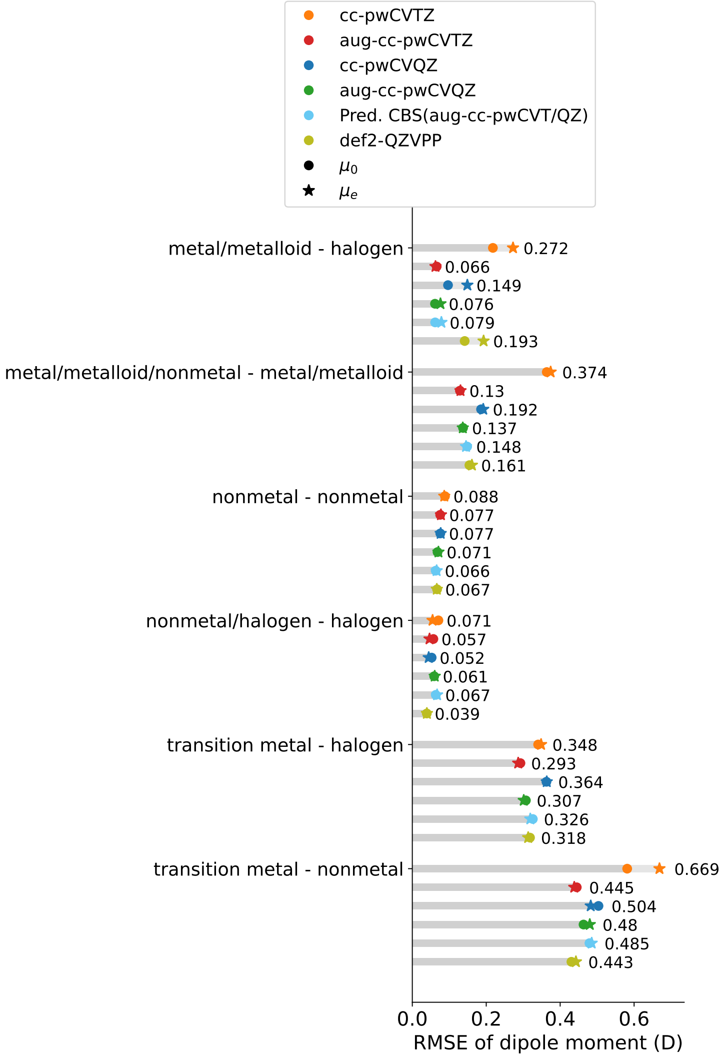


Figure 43: RMSE of CCSD(T) dipole moment calculated with cc-pwCVT/QZ, aug-cc-pwCVT/QZ and def2-QZVPP basis sets. The dipole moment at the equilibrium bond length $\mu_e$ and the zero-point vibrational corrected dipole moment $\mu_0$ are denoted as circle and star, respectively. Figure reproduced from ref. [164].

A significant basis set size effect is observed in the case of the cc-pwCV basis set. For most molecules, increasing the size of the basis set from cc-pwCVTZ to cc-pwCVQZ reduces the underestimation of the dipole moment, leading to an improvement in the RMSE from 0.30 D to 0.24 D. Notably, for molecules containing main-group metal/metalloid elements, increasing the size of the basis set from cc-pwCVTZ to cc-pwCVQZ results in a reduction of the RMSE by a factor of two. However, for nonmetal-nonmetal and nonmetal/halogen-halogen molecules, the improvement in dipole moment accuracy from the triple-$\zeta$ to quadruple-$\zeta$ level is less than 0.02 D. Therefore, for molecules with nonmetal elements, it is generally sufficient to use the cc-pwCVTZ basis.

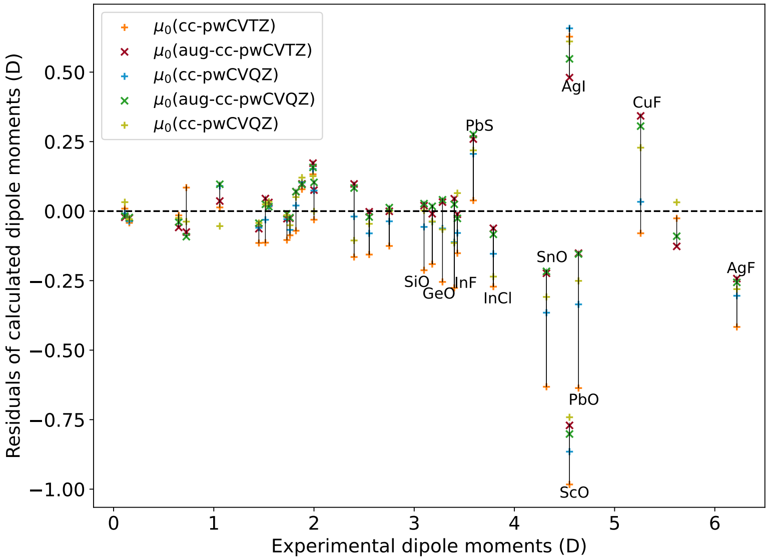


Figure 44: The errors of dipole moment calculated with cc-pwCVT/QZ and aug-cc-pwCVT/QZ basis sets. Figure reproduced from ref. [164].

When considering the aug-cc-pwCV basis, it becomes evident that dipole moments are almost converged at the aug-cc-pwCVTZ level. Consequently, the benefits of using the larger aug-cc-pwCVQZ basis are marginal, typically less than 0.01 D. In some cases, especially for diatomic molecules with metal atoms, the aug-cc-pwCVTZ predictions are slightly more accurate than the aug-cc-pwCVQZ level. The role of augmented functions will be further discussed in Sec. 5.3.3.

Surprisingly, for most molecules, the use of the def2-QZVPP basis set yields a similar level of accuracy to the aug-cc-pwCVQZ basis set, despite the def2-QZVPP basis being much smaller than the aug-cc-pwCVQZ basis. In some instances, the def2-QZVPP basis even provides dipole moments that are closer to experimental values than the CBS(aug-cc-pwCVT/QZ) basis set.

**Influence of diffuse functions** It has been demonstrated that adding diffuse functions to the basis set can play a crucial role in improving dipole moment predictions in hybrid and double-hybrid density functionals, as well as wave function-based methods, to approach CCSD(T) results [19, 198]. To investigate whether the inclusion of diffuse functions enhances CCSD(T) predictions towards experimental values, we used the cc-pwCVT/QZ basis sets to calculate dipole moments and compared the results to those obtained with augmented basis sets, as shown in Fig. 44.

As discussed in Sec. 5.3.3, the prediction of the dipole moment's magnitude increases when transitioning from the cc-pwCVTZ to the quadruple-$\zeta$ level. Adding augmentation functions further amplifies the dipole moment's magnitude, occasionally resulting in an overestimation. Molecules containing metal elements are more sensitive to augmentation due to the longer-range nature of the wave function. On average, the RMSE can be reduced by 0.10 D with the inclusion of augmented functions at the triple-$\zeta$ level. However, at the quadruple-$\zeta$ level, the overall improvement from augmentation diminishes to only 0.03 D, considerably smaller than at the triple-$\zeta$ level. Therefore, for dipole moments of most molecules, the enhancement gained from including augmented functions is negligible at the quadruple-$\zeta$ level.

Notably, exceptions are molecules containing metal/metalloid halides (metal/metalloid-halogen and transition metal-halogen molecules), where the improvement in RMSE by employing diffuse functions at the quadruple-$\zeta$ level is approximately 0.07 D. Consequently, for these molecules, it is advisable to use augmented basis sets.

### Overall performance of CCSD(T) on the dipole moments

The performance of CCSD(T) predictions on the dipole moment $\mu_0$ using the CBS(aug-cc-pwCVT/QZ) and def2-QZVPP basis sets is summarized in

Fig. 45. Overall, the performance of Dunning's and def2- basis sets is very similar, with RMSE values of 0.215 D and 0.209 D, respectively. Significant errors exceeding 0.2 D are primarily observed for molecules with dipole moments greater than 3 D, although there is no clear correlation between the dipole moment error and its absolute value.

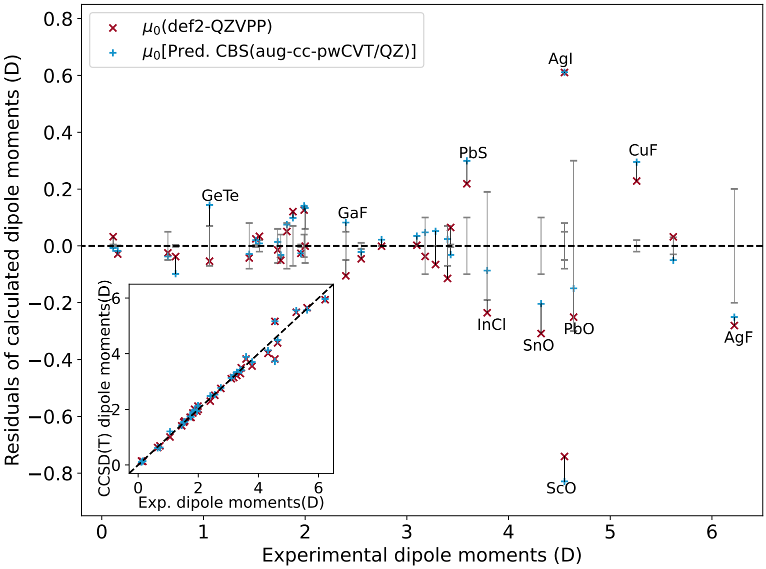


Figure 45: The errors of dipole moments calculated with def2-QZVPP and aug-cc-pwCVQZ basis sets. The error bars of experimental measurements are shown in gray. Figure reproduced from ref. [164].

Specifically, diatomic molecules containing main-group elements, especially non-metal elements, are well-described, with relative errors $r_E$ of less than 5% and RMSE values of less than 0.08 D for non-metal-non-metal molecules and less than 0.5 D for non-metal halides. For molecules containing main-group metals/metalloids, the RMSE values are larger (less than 0.15 D), but they remain close to the experimental uncertainty, except for SnO and PbS.

In contrast, transition metal-containing systems exhibit larger errors. With the predicted CBS(aug-cc-pwCVT/QZ), an RMSE of 0.32 D (relative error $r_E$ of 5.5%) is observed for transition metal halides, and an RMSE of 0.51 D (relative error $r_E$ of 6.9%) is observed for other transition metal-nonmetal diatomic molecules. These errors are notably larger than the experimental uncertainties.

The sources of the discrepancies between CCSD(T) predictions and experimental values can vary. The use of a relatively large basis set for the predicted CBS suggests that the errors are not solely due to basis set size. Another potential source could be the multi-reference character of the molecules. However, in the current dataset, transition metal-containing molecules or their analogs are generally dominated by single-reference character [199, 15]. Furthermore, it is worth noting that the residuals of CCSD(T) predictions do not consistently correlate with the experimental uncertainties. While some molecules, such as PbO and InCl, have residuals closer to the experimental uncertainty, others, like PbS, exhibit residuals much larger than the experimental uncertainty. In the following section, we will explore possible sources of these errors in more detail.

### 5.3.4 Origin of the errors

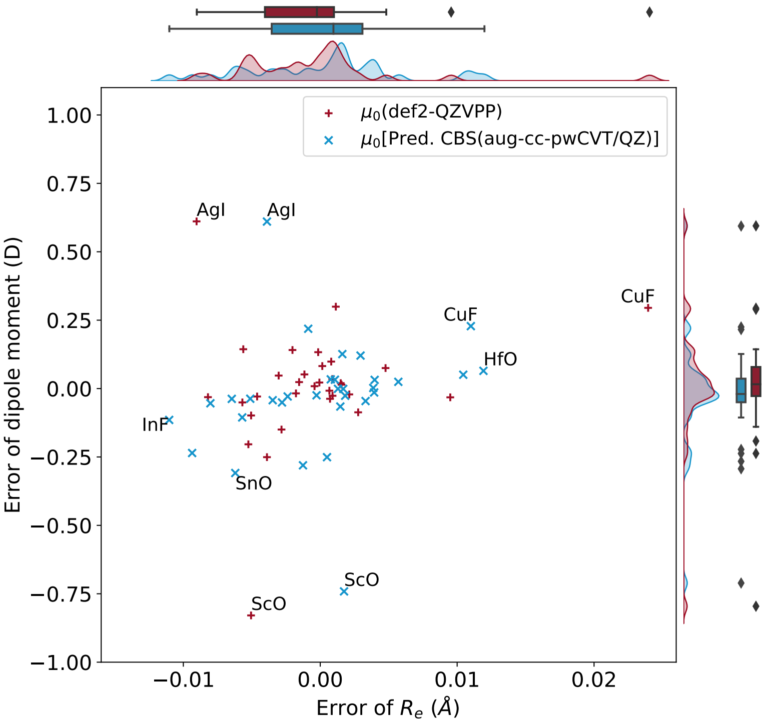


Figure 46: Errors of dipole moments as a function of errors of $R_e$ with def2-QZVPP and predicted CBS(aug-cc-pwCVT/QZ). Figure reproduced from ref. [164].

The analysis of possible sources of error in calculated dipole moments reveals some interesting insights. One potential source could be inaccuracies in predicting bond lengths, as the dipole moment is proportional to the charge separation multiplied by the bond length [4]. However, the analysis in Fig. 46 does not show a clear relationship between errors in equilibrium bond lengths ($R_e$) and dipole moments. This observation aligns with previous studies on the nature of dipole moments in diatomic molecules [163]. For instance, both the aug-cc-pwCV and def2- basis sets overestimate the dipole

[4] For more details, please refer to Chapter 4.

moment of AgI and CuF while simultaneously underestimating the dipole moment of ScO, despite accurate predictions of $R_e$ with aug-cc-pwCVQZ for these molecules. Conversely, several molecules with precise dipole moments exhibit more significant errors in the prediction of $R_e$, such as HfO, IBr, and InF. Similarly, the errors in dipole moments do not correlate with errors in $\omega_e$. These results suggest that benchmark studies based solely on energetic properties may not adequately predict other properties related to electron density.

Furthermore, some outliers in Fig. 45 require additional discussion. While one might expect these errors to be due to the single-reference nature of CCSD(T) calculations, it has been previously demonstrated that molecules with significant dipole moment errors often exhibit a dominant single-reference character [15].

Another potential source of errors is the non-relativistic treatment. To investigate the role of relativistic effects, CCSD(T) calculations were performed, including the scalar relativistic correction, for 9 molecules with significant dipole moment errors. The dipole moments ($\mu_e$) were calculated at the experimental geometry both with and without the relativistic correction, and the results are summarized in Tab. 7. In most cases, including relativistic effects slightly decreases the magnitude of the predicted dipole moment, with differences typically ranging from 0.01 to 0.07 D (0.3% to 1.5%). Notably, for CuF, the difference introduced by relativistic effects is 0.14 D (2.7%). For molecules where the non-relativistic CCSD(T) treatment overestimated the dipole moment (e.g., PbS, AgI, CuF), the relativistic dipole moment becomes closer to the experimental value. Conversely, for other molecules, the underestimation of the dipole moment is further exacerbated by the relativistic correction. Overall, including the relativistic correction, the RMSE for the 9 molecules can be slightly improved from 0.252 D to 0.235 D.

Investigating individual molecules with the most significant errors, such as ScO, CuF, and AgI, provides valuable insights into the origin of these errors.

In the case of CuF, the dipole moment is overestimated by 0.28 D at the CCSD(T)/CBS(aug-cc-pwCVT/QZ) level, compared to the experimental value of 5.26(2) D, as determined by recent supersonic molecular beam high-resolution optical Stark spectroscopy [200]. Additionally, there is a small discrepancy of 0.02 Å between the experimental equilibrium bond length ($R_e$) and the CCSD(T)/aug-cc-pwCVQZ prediction. At the experimental $R_e$, the predicted dipole moment ($\mu_e$) is 5.420 D with CCSD(T)/aug-cc-pwCVQZ,

Table 7: Dipole moment at experimental equilibrium bond length $\mu_e$, calculated with non-relativistic or scalar relativistic CCSD(T).

| Molecule | $\mu_0$(Exp.) (D) | Non-relativistic | | Scalar relativistic | |
|---|---|---|---|---|---|
| | | $\mu_e$ (D) | Basis set | $\mu_e$ (D) | Basis set |
| AgF | 6.22(20) | 5.991 | Ag:cc-pwCVTZ-PP; F:cc-pVQZ | 5.956 | Ag:cc-pwCVTZ-DK; F:cc-pVQZ-DK |
| AgBr | 5.62(3) | 5.789 | Ag:cc-pwCVTZ-PP; Br:cc-pVQZ | 5.716 | Ag:cc-pwCVTZ-DK; Br:cc-pVQZ-DK |
| AgI | 4.55(5) | 5.139 | Ag:cc-pwCVTZ-PP; I:cc-pwCVTZ-PP | 5.087 | Ag:cc-pwCVTZ-DK; I:cc-pwCVTZ-DK3 |
| CuF | 5.26(2) | 5.420 | Cu:aug-cc-pwCVQZ; F:aug-cc-pwCVQZ | 5.278 | Cu:aug-cc-pwCVQZ-DK; F:aug-cc-pwCVQZ-DK |
| InCl | 3.79(19) | 3.629 | In:aug-cc-pwCVQZ-PP; Cl:aug-cc-pwCVQZ | 3.616 | In:aug-cc-pwCVQZ-DK3; Cl:aug-cc-pwCVQZ-DK |
| PbO | 4.64(30) | 4.479 | Pb:aug-cc-pwCVQZ-PP; O:aug-cc-pCVQZ | 4.460 | Pb:aug-cc-pwCVQZ-DK3; O:aug-cc-pCVQZ-DK |
| PbS | 3.59(10) | 3.726 | Pb:aug-cc-pwCVQZ-PP; S:aug-cc-pCVTZ | 3.669 | Pb:aug-cc-pwCVQZ-DK3; S:aug-cc-pCVTZ-DK |
| SnO | 4.32(10) | 4.106 | Sn:aug-cc-pwCVQZ-PP; O:aug-cc-pCVQZ | 4.074 | Sn:aug-cc-pwCVQZ-DK3; O:aug-cc-pCVQZ-DK |
| SnS | 3.18(10) | 3.190 | Sn:aug-cc-pwCVQZ-PP; S:aug-cc-pCVQZ | 3.147 | Sn:aug-cc-pwCVQZ-DK3; S:aug-cc-pCVQZ-DK |

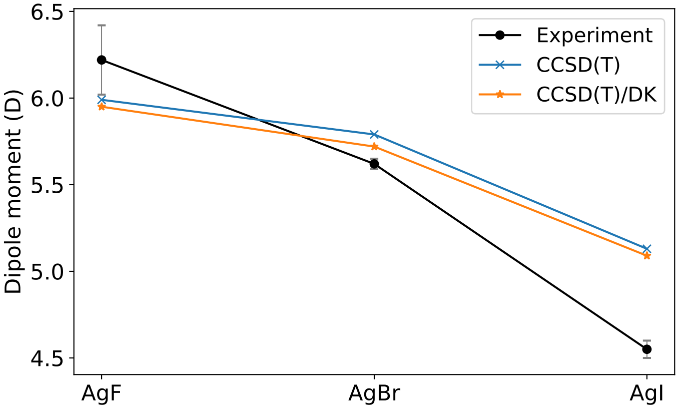


Figure 47: Dipole moments of AgF, AgBr, and AgI. The theoretical predictions are calculated at experimental $R_e$ with CCSD(T)/(Ag:cc-pwCVTZ-PP; F/Br: cc-pVQZ; I:cc-pwCVTZ-PP), and CCSD(T)/(Ag:cc-pwCVTZ-DK; F/Br:cc-pVQZ-DK; I:cc-pwCVTZ-DK) with scalar DK relativistic corrections. The experimental uncertainties are shown in gray. Figure reproduced from ref. [164].

still 0.16 D away from the experimental dipole moment. Interestingly, when scalar relativistic corrections (Douglas-Kroll method) are applied to CCSD(T) calculations with corresponding relativistic-contracted basis sets (Cu:aug-cc-pwCVQZ-DK; F:aug-cc-pwCVQZ-DK), the experimental $R_e$ is perfectly reproduced with an error of only 0.001 Å, and the dipole moment obtained is 5.28 D. This result is consistent with previous reports [201] and aligns with studies on other diatomic molecules like CuH, AgH, and AuH, where small changes in predicted $R_e$ due to the inclusion of relativistic effects have been observed to lead to significant changes in $\mu_e(R_e)$ [202].

For the Stark effect in the rotational spectrum of AgI, experimental measurements yield a dipole moment of 4.55(5) D [203]. However, our calculated value using the CBS(aug-cc-pwCVT/QZ) method, which includes the vibrational average correction, stands at 5.15 D, indicating an overestimation compared to the experimental result. Interestingly, our theoretical prediction aligns closely with previous computational estimates [204].

On the other hand, for AgF, an isovalent analogue of AgI, CCSD(T) calculations tend to underestimate the dipole moment. Additionally, when examining silver-halogen molecules, it becomes evident that experimentally, there is a significant reduction in dipole moment as the halogen atom becomes heavier. However, theoretical predictions suggest a less pronounced decrease, as illustrated in Fig. 47. The disagreement between CCSD(T) calculations and experimental observations persists even when considering scalar relativistic effects. The difference in dipole moment obtained by applying relativistic corrections (CCSD(T)/(Ag:cc-pwCVTZ-DK; I:cc-pwCVTZ-DK)) and non-relativistic calculations (CCSD(T)/Ag:cc-pwCVTZ-PP; I:cc-pwCVTZ-PP) amounts to only 0.05 D at the experimental equilibrium bond length ($R_e$).

Additionally, we notice that a previous measurement by the same research group reported a dipole moment of $\mu_0 = 5.10(15)$ D [205], which closely aligns with our CCSD(T)/(aug-cc-pwCVT/QZ) predictions. This discrepancy between different experimental measurements and theoretical calculations calls for further revision and investigation.

## 5.4 Conclusion

In this study, we have comprehensively benchmarked CCSD(T) calculations for predicting dipole moments by comparing computational results, particularly those obtained using large basis sets, with precise experimental measurements. Our investigation focused on 32 diatomic molecules, encompassing a wide range of bonding characteristics and elemental compositions, for which accurate experimental dipole moment data were available. This approach allowed us to assess the computational accuracy against experimental reference values directly. Additionally, we have examined the accuracy of equilibrium bond lengths and vibrational harmonic frequencies and compared them to the precision of dipole moment predictions.

Our findings indicate that single-reference CCSD(T) calculations, utilizing basis sets from the def2- and aug-cc-pwCVX families (with X representing T and Q), generally provide satisfactory descriptions of the dipole moments for most molecules within the dataset. The errors typically remain below 0.15 D, especially for molecules composed solely of main-group elements. However, the choice of the most suitable basis sets depends on the specific molecule under investigation. For instance, our study revealed that for the cc-pwCV basis at the triple-$\zeta$ level, the inclusion of diffuse functions plays a crucial role in improving accuracy, especially in molecules containing metal elements. Nonetheless, the augmentation with diffuse functions has a minor impact on dipole moment predictions at the quadruple-$\zeta$ level, where the basis set already provides sufficient accuracy. Similarly, we observed that basis set incompleteness errors are apparent at the triple-$\zeta$ level when employing the cc-pwCV basis, while its augmented counterpart yields nearly converged results. Notably, the def2-QZVPP basis exhibits comparable performance to the much larger aug-cc-pwCVQZ basis set, making it a preferable choice for dipole moment predictions in larger systems.

Our results indicate that non-relativistic predictions are generally accurate enough for most molecules. While scalar relativistic corrections may be essential in dipole moment calculations for certain molecules, such as CuF, they do not appear to be the primary source of error in most cases. For molecules like ScO and AgI, which exhibit significant discrepancies with experimental data that cannot be satisfactorily explained by multi-reference or relativistic effects, we recommend a comprehensive evaluation involving both experimental and theoretical approaches, potentially including multi-reference coupled-cluster calculations, to gain a deeper understanding of dipole moments in these systems. Furthermore, expanding the dataset to include molecules involving alkali metals and other relevant systems would be valuable for future benchmark studies.

Finally, our study highlights an important insight: errors in predicting dipole moments do not necessarily correlate with inaccuracies in equilibrium bond lengths. This observation underscores that errors in dipole moment predictions are primarily associated with deviations in the electron distribution rather than differences in bond lengths. It reinforces the notion that different properties are predicted with varying degrees of accuracy within computational approximations, highlighting the need for a comprehensive assessment of methods beyond energetic and geometric properties.

# 6

# THE HYPERFINE CONSTANTS OF ALUMINUM MONOFLUORIDE

## 6.1 The hyperfine constants of diatomic molecules

The hyperfine structure of molecules stems from nuclear magnetic moments and higher-order multipolar moments. Specifically, the hyperfine structure arises from different magnetic interactions of the nuclear magnetic moments with the electronic orbital angular momentum, the spin angular momentum, the molecular rotation-induced magnetic field, and the interaction between the magnetic moments of two nuclei, along with quadrupole interaction [206].

In this study, we compute the magnetic hyperfine coupling constants $b_F, c, d$, along with the electric hyperfine coupling constant (the nuclear quadrupole coupling constant) $eq_0Q$, following the definitions of $c, d$ introduced by Frosch and Foley to approximate the magnetic interactions in open-shell electronic states of diatomic molecules [207], and the notations $eq_0Q$, $b_F$ defined by Brown and Carrington[208]. A comprehensive comparison between these hyperfine constants can be found in [208] [1].

$b_F$ is called the Fermi contact parameter, which relies solely on the electron density of unpaired electrons at the nuclei as

$$b_F \sim \int \psi^2(\boldsymbol{r})\delta(\boldsymbol{r})d\boldsymbol{r}. \tag{24}$$

$b_F$ encompasses the non-dipolar isotropic component of the electron-nuclear spin interaction, given that it involves only the integration over the probability density of the wave function of the unpaired electron, which solely depends on the electron-nuclear distance $\boldsymbol{r}$. Indeed, a non-zero $b_F$ implies the existence of an s-type component in a molecule's molecular orbitals. Therefore, $b_F$ can be calculated as the polarization of these molecular orbitals, which is the difference between the electronic densities of up and down spins at the nucleus [210].

The hyperfine constants $c$ and $d$ characterize the axial and perpendicular anisotropic components, respectively, and can be obtained from the Cartesian components of the dipolar hyperfine tensor calculated by quantum

[1] This chapter is written based on reference [209]: Nicole Walter, Maximilian Doppelbauer, Silvio Marx, Johannes Seifert, Xiangyue Liu, Jesús Pérez-Ríos, Boris G. Sartakov, Stefan Truppe, and Gerard Meijer. Spectroscopic characterization of the a$^3\Pi$ state of aluminum monofluoride. The Journal of Chemical Physics, 156(12), 2022, https://doi.org/10.1063/5.0082601.

chemistry methods [211, 212]. The nuclear quadrupole coupling constant $eQ_0Q$ (in MHz) can be determined by establishing its connection with the electric field gradients (EFGs, in atomic units a.u.) and the nuclear quadrupole moment Q (in MBarn), as[213, 214]

$$eq_0Q = \frac{Q\,\langle V_{zz}|V_{zz}\rangle_v}{4.255958}, \tag{25}$$

where $Q = 146.6$ MBarn for the $^{27}$Al atom [215],
and $\langle V_{zz}|V_{zz}\rangle_v = \langle \psi_v|V_{zz}|\psi_v|\psi_v|V_{zz}|\psi_v\rangle$ stands for the expectation value of $V_{zz}$ based on a given vibrational state $|\psi_v\rangle$.

This correlation is derived from the multipole expansion of the hyperfine contribution to the Hamiltonian. For each vibrational state, it is necessary to compute the expectation value of the EFGs.

## 6.2 Computational details

In this work, we have determined the hyperfine constants for the a$^3\Pi$ state of AlF, which is the lowest state of the triplet manifold of AlF. We have employed density functional theory (DFT) implemented in Gaussian 2016 [145], utilizing the CAM-B3LYP functional[216], which is a range-separated hybrid density functional with a significant percentage of Hartree-Fock– included in the long-range interactions. The basis set employed is the aug-cc-pV5Z basis set[217, 218, 219].

The potential energy curve of AlF's a$^3\Pi$ state has been derived from 23 DFT-calculated potential energies spanning a range of interatomic distances from 1 to 6 Å. To ascertain the expectation values of the hyperfine constants at various vibrational states, we have solved the vibrational Schrödinger equation utilizing the discrete variable representation (DVR) approach [48, 49]. It has been shown that employing 200 grid points yields a converged result [2].

## 6.3 Results

The spectroscopic constants, including the kinetic energy $T_e$, equilibrium internuclear distance $R_e$, harmonic frequency $\omega_e$, are fitted from CAM-

[2] The difference between the $eq_0Q$ results obtained from 200 and 300 DVR points is less than $10^{-9}$ MHz.

B3LYP/aug-cc-pV5Z potential energy curves and summarized in Table 8. Both the ground states of the singlet and triplet states have been obtained, and the fitted harmonic vibrational frequency aligns well with experimental observations.

Table 8: Spectroscopic parameters of $^{27}$Al$^{19}$F, including the kinetic energy $T_e$, equilibrium internuclear distance $R_e$, harmonic frequency $\omega_e$ and at the CAM-B3LYP/aug-cc-pV5Z level. The experimental values are taken from [30].

| | $T_e$ (cm$^{-1}$) | $R_e$ (Å) | $\omega_e$ (cm$^{-1}$) |
|---|---|---|---|
| Exp. | 27239.4529(53) | 1.64708 | 830.280 7(11) |
| DFT | 25587.86 | 1.655 | 818.72 |

The nuclear quadrupole coupling constants $eq_0Q$, the isotropic Fermi contact coupling constants $b_F$, the anisotropic spin-dipole coupling constants $c$ and $d$ of $^{27}$Al$^{19}$F $a^3\Pi$ state are shown in Fig. 48, 49, 50 and 51, respectively, for vibrational states $v = 0$ to $v = 5$, and summarized in Table 9.

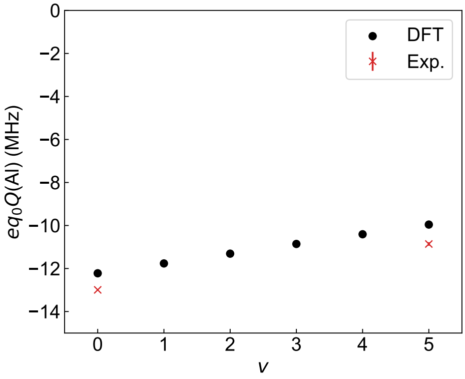


Figure 48: Black: The DFT-calculated $eQq(^{27}\mathrm{Al})$ of the a$^3\Pi$ state of $^{27}$Al$^{19}$F for different vibrational states, at CAM-B3LYP/aug-cc-pV5Z level. Red: The experimental values taken from [5].

As shown in Fig. 48, the nuclear quadrupole coupling constant $eq_0Q(Al)$ decreases for higher vibrational states, where $eq_0Q(Al)(v=5)/eq_0Q(Al)(v=0) = 0.81$, showing the same trend as the experimental measurements. The calculated values of $eq_0Q$ agree with the experiment with deviations $\lesssim 5\%$.

The change of hyperfine parameters from $v = 0$ to $v = 5$ can be attributed to the increase of the internuclear distance. The discrepancies could arise from inaccuracies in the electron correlation treatment within the DFT calculations.

Table 9: The DFT-calculated hyperfine constants $eQq(\mathrm{Al})$, $b_F(\mathrm{Al})$, $b_F(\mathrm{F})$, $c(\mathrm{Al})$, $c(\mathrm{F})$, $d(\mathrm{Al})$ and $d(\mathrm{F})$ (in MHz) of the $^{27}\mathrm{Al}^{19}\mathrm{F}$ $a^3\Pi$ state for different vibrational states, at CAM-B3LYP/aug-cc-pV5Z level.

| $v$ | $eq_0Q(\mathrm{Al})$ | $b_F(\mathrm{Al})$ | $b_F(\mathrm{F})$ | $c(\mathrm{Al})$ | $c(\mathrm{F})$ | $d(\mathrm{Al})$ | $d(\mathrm{F})$ |
|---|---|---|---|---|---|---|---|
| 0 | -12.221 | 1186.2 | 150.1 | -24.6 | 154.3 | 126.5 | 128.6 |
| 1 | -11.763 | 1189.8 | 147.0 | -25.3 | 160.8 | 126.9 | 127.1 |
| 2 | -11.309 | 1193.1 | 144.0 | -25.9 | 167.6 | 127.4 | 125.5 |
| 3 | -10.857 | 1196.2 | 141.0 | -26.6 | 174.7 | 127.8 | 124.0 |
| 4 | -10.407 | 1199.1 | 138.0 | -27.2 | 182.0 | 128.3 | 122.4 |
| 5 | -9.957 | 1201.7 | 135.1 | -27.9 | 189.7 | 128.7 | 120.8 |

Likewise, we observe a dependency on vibrational frequency in the calculated isotropic Fermi contact coupling constant $b_F$ (Fig. 49), as well as the anisotropic coupling constants $c$ (Fig. 50) and $d$ (Fig. 51). The disparities between the experimental and computed $d$ values are quite small, measuring less than 5%. However, the deviations become more significant in the case of $c$ and $b_F$. This discrepancy may arise from the necessity of maintaining the $c$ parameter constant in the fit, while $b_F$ is determined experimentally through its relationship with the Frosch-Foley hyperfine constants, with $b_F = b + c/3$ [5, 208].

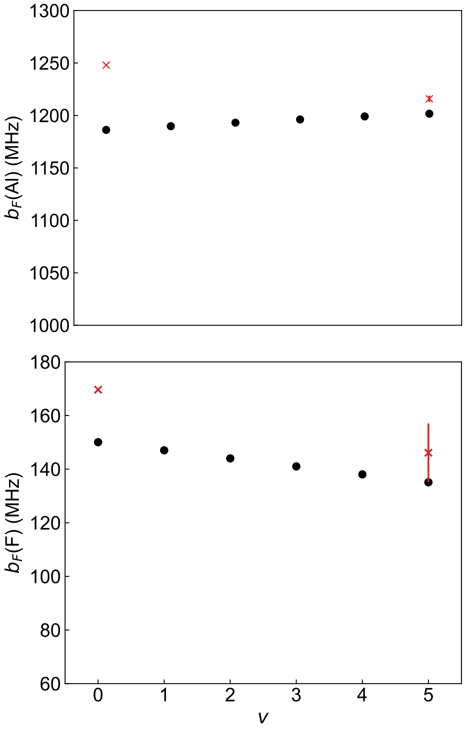


Figure 49: Black: The DFT-calculated Fermi contact $b_F$(Al) and $b_F$(F) of the $^{27}$Al$^{19}$F $a^3\Pi$ state for different vibrational states, at CAM-B3LYP/aug-cc-pV5Z level. Red: The experimental values taken from [5].

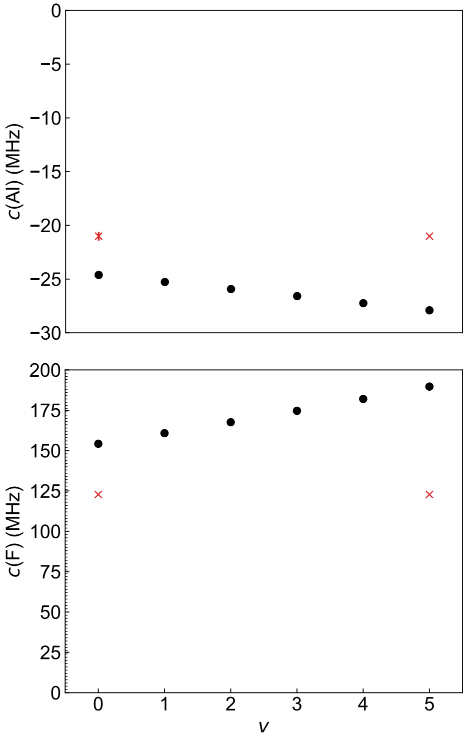


Figure 50: Black: The DFT-calculated anisotropic spin-dipole coupling constants $c$(Al) and $c$(F) of the $^{27}$Al$^{19}$F $a^3\Pi$ state for different vibrational states, at CAM-B3LYP/aug-cc-pV5Z level. Red: The experimental values taken from [5].

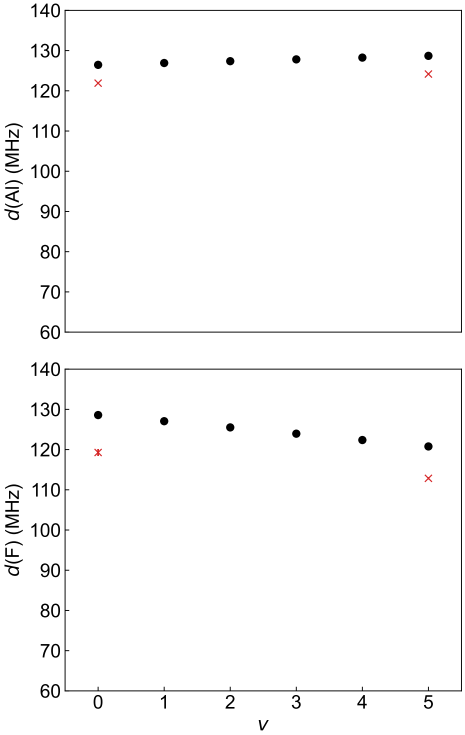


Figure 51: Black: The DFT-calculated anisotropic spin-dipole coupling constants $d$(Al) and $d$(F) of the $^{27}$Al$^{19}$F $a^3\Pi$ state for different vibrational states, at CAM-B3LYP/aug-cc-pV5Z level. Red: The experimental values taken from [5].

## 6.4 Conclusion

Comparing experimental and calculated hyperfine constants is of paramount importance in benchmarking studies. This is because hyperfine constants are more sensitive to the accuracy of the electron-correlation treatment than energetic properties. In this work, we have computed the hyperfine constants, including the nuclear quadrupole coupling constants $eq_0Q$, isotropic Fermi contact coupling constants $b_F$, and anisotropic spin-dipole coupling constants $c$ and $d$. We have noted that both the sign and magnitude of the calculated hyperfine parameters align well with the experimentally determined values. These comparisons serve as a critical validation step, assessing the accuracy and reliability of quantum chemistry methods. In addition, the comparison necessitates meticulous consideration of the assumptions made during the fitting process when deriving hyperfine constants from experimental measurements.

Part II

# CHEMISTRY OF DIATOMIC FLUORIDES

# 7

# CHEMISTRY OF ALF AND CAF PRODUCTION IN BUFFER GAS SOURCES

## 7.1 Background and method

### 7.1.1 Introduction

One way to obtain molecules in the ultracold regime is via direct cooling techniques, where molecules are produced in a source and then extracted into a molecular beam, which is subsequently slowed and captured in a trap [220, 221, 222, 223, 30]. As a first step toward ultracold molecules, it is necessary to bring molecules in the cold regime (T$\sim$ 1K). One possibility is cryogenic buffer gas cooling, exploiting the heat dissipation via collisions between the molecules and a cold reservoir, often a buffer gas such as helium gas. Cryogenic buffer gas cooling has become common in this field due to its ability to efficiently cool the rotational and vibrational degrees of freedom in molecules [224].

Metal-fluorine diatomic molecules, among the species amenable to cryogenic buffer gas cooling, have been extensively employed in precision spectroscopy and laser cooling applications [220, 221, 222, 223]. For instance, AlF has been recently theoretically and experimentally found to be feasible for laser-cooling because of its highly-diagonal Franck-Condon factor in its $A^1\Pi, v = 0 \leftarrow X^1\Sigma^+, v' = 0$ transition [29, 30]. Similarly, CaF, with its ground state of $^2\Sigma^+$, has been identified for years as a promising candidate for direct laser cooling [221].

This Chapter focuses on the production of molecules in buffer gas sources. In many of these sources, molecules are created through a chemical reaction involving laser-ablated atoms with an initial temperature of several thousand Kelvin and a fluorine-donor reactant gas. This process spans a wide range of energy scales, involves complex reaction kinetics, and is not well understood. Recently, there have been successful simulations of molecular beams emerging from buffer gas sources based on general physical properties [225, 226, 227]. However, the primary focus of these numerical simulations is to investigate how the density and velocity of the buffer gas, along with the design of the buffer gas cell, influence the cooling process and dynamic characteristics of the molecular beam. Indeed, the impact of the reactive gas has not been explored yet, neither experimentally nor theoretically. Additionally, properties of the molecular beam, such as its overall yield, short- and long-term stability, and phase-space distribution, play a crucial role in determining which downstream experiments are feasible. Consequently, a thorough understanding of the chemistry in buffer gas cells is essential to design molecular beams that are both brighter and colder, thereby facilitating the implementation of subsequent cooling techniques.

Our study investigates explicitly the formation efficiency of AlF and CaF molecules after laser-ablating Al or Ca atoms in two fluorine-donor gases: $NF_3$ and $SF_6$. The primary objective is to gain a deeper understanding of the reaction dynamics and analyze how the specific fluorine donor gas influences the properties of the molecular beam. Through molecular dynamics simulations, we demonstrate that the number and kinetic energy of the product molecules depends not only on the intensity of the ablating laser, which sets the reaction temperature but also on the nature of the fluorine donor gas. Here, we investigate $NF_3$ and $SF_6$. These results lay the foundation for a more detailed comprehension of the properties of buffer gas molecular beams, enabling improved design and optimization in the future. [1] .

### 7.1.2 Computational details and methodology

First, we must select an appropriate ensemble for simulating the Al/Ca+$NF_3$/$SF_6$ reactions in the buffer gas cell. AlF and CaF molecules form when highly energetic ablated metal atoms collide with lower-energy fluorine-containing molecules. A grand canonical ensemble should be used during the simulation to describe this non-equilibrium system adequately. However, this approach is computationally expensive.

Nonetheless, we can make a reasonable approximation by considering that the temperature of the ablated atoms is much higher than the buffer gas temperature, making the atom temperature a suitable proxy for the reaction temperature. This approximation is valid since the reaction takes less time ($\sim 1$ ps) compared to the collision time between He atoms and F-containing molecules, estimated to be around $\sim 500$ ns, assuming a He-molecule elastic cross-section of approximately $\sim 10^{-14}$ cm$^{-2}$ and a He density of $10^{21}$ m$^{-3}$. Consequently, we can treat the process as energy-conserved and simulate the reactions using the microcanonical ensemble.

In a reaction chamber with constant temperature, the thermodynamic states of the molecules follow a Maxwell-Boltzmann distribution. Therefore, in our molecular dynamics (MD) simulations, we initiate $N_t$ trajectories, with their initial states determined by the appropriate Maxwell-Boltzmann distribution. The interaction potential has been calculated *ab initio* using

[1] This chapter is written based on reference [228]: Xiangyue Liu, Weiqi Wang, Sidney C. Wright, Maximilian Doppelbauer, Gerard Meijer, Stefan Truppe, and Jesús Pérez-Ríos. The chemistry of AlF and CaF production in buffer gas sources. *The Journal of Chemical Physics*, 157(7):074305, 08 2022.https://doi.org/10.1063/5.0098378

the hybrid BHLYP functional [229, 230], previously shown to reproduce the experimental rate constants for the Al + $SF_6$ reaction [231]. Additionally, we include the D3 dispersion correction [232] to account for long-range interactions. All calculations are performed using the def2-TZVP basis set [142, 143, 144], as implemented in the Gaussian 16 package [145].

### 7.1.3 Initial conditions

As depicted in Fig. 52, the metal atoms' initial positions for the MD simulations are randomly chosen from a sphere with a radius of 7 Å, centered at either atom N or S of the target molecules $NF_3$ or $SF_6$. Here, we use the molecules' symmetry to restrict the range of angular degrees of freedom. The speed of the metal atoms satisfies a Maxwell-Boltzmann distribution at corresponding temperatures, and its direction is randomly sampled following the polar and azimuthal angles in spherical coordinates to ensure a collision with the target molecule as sketched by the pink surface in Fig.52, which represents the surface that connects all the fluoride atoms in the molecules ($NF_3$ or $SF_6$).

### 7.1.4 Reaction probability

We compare AlF/CaF production reactions based on the reaction probability. The reaction probability is defined as the probability that the reactants result in a specific product state. In particular, we examine the probability of obtaining the desired product AlF/CaF, as well as the probabilities of forming by-products $AlF_n/CaF_n$ (n = 2, 3). The reaction productivity can be expressed as

$$\mathscr{P}_r = \frac{N_r}{N_t}, \tag{26}$$

where $N_r$ denotes the number of trajectories leading to a given reaction product. In particular, we run $N_t = 1000$ trajectories for each temperature and colliding species, tracking the metal atom until it reaches a final state beyond a distance of 7 Å from the N or S atom.

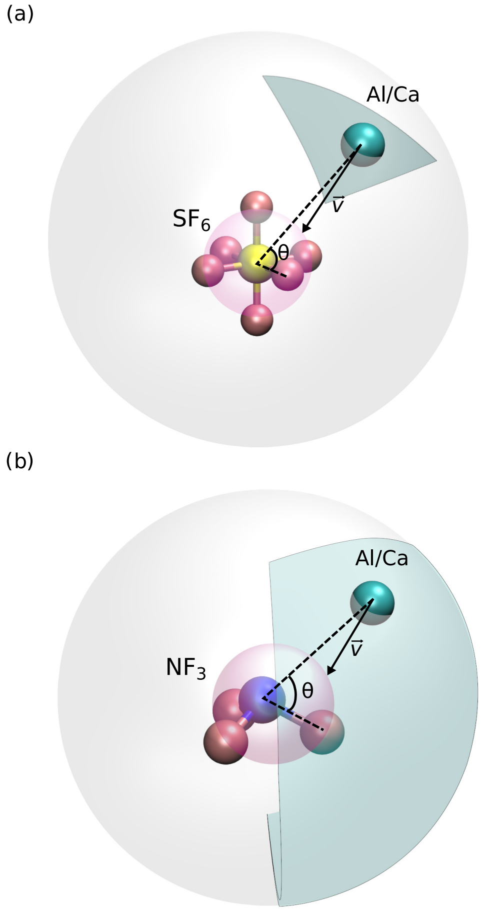


Figure 52: Initial positions of Al/Ca atoms in the AIMD simulations, randomly sampled from a symmetry-reduced sphere of radius 7 Å, centered at the S or N atoms. The velocity vector direction is randomly sampled to ensure an atom-molecule collision as the pink surface sketches it. Figure reproduced from ref. [228].

## 7.2 Results and discussions

### 7.2.1 Reaction probability

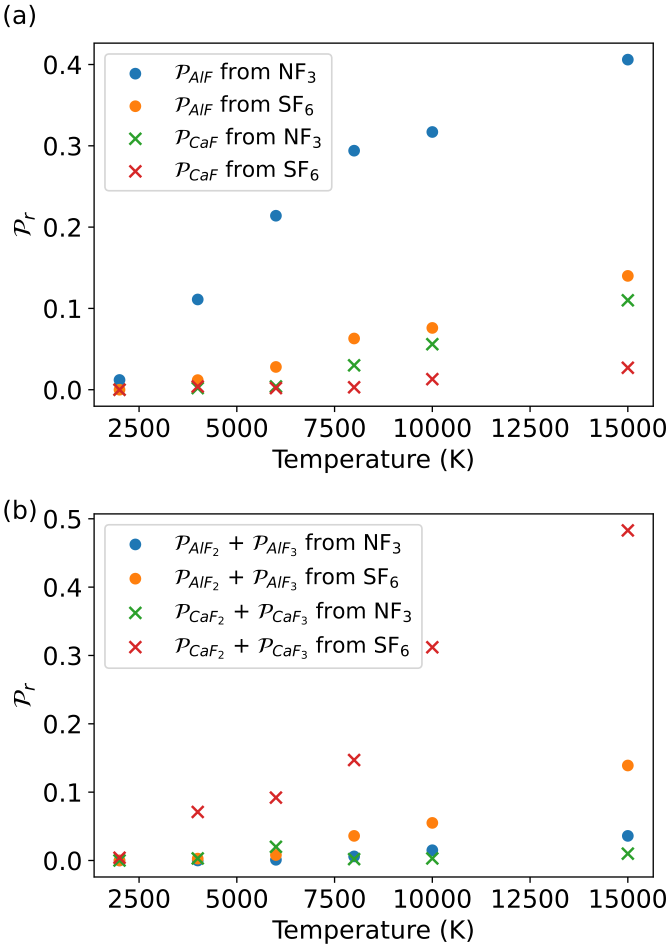


Figure 53: Reaction probability of (a) AlF/CaF and (b) $AlF_n$/$CaF_n$ by-products for hot collisions of Al/Ca with $SF_6$ and $NF_3$ gases as a function of the temperature. Figure reproduced from ref. [228].

The efficiency of AlF and CaF molecule formation in Al/Ca + $SF_6$/$NF_3$ collisions is presented in Fig. 53, displaying the reaction probabilities for different products as a function of temperature. Across all temperatures, AlF production is consistently more efficient than CaF, regardless of the type of reactant gas used in the reaction. This observation aligns with experimental findings where the brightness of CaF and AlF beams emerging from a buffer gas cell was compared[224, 223, 233]. Indeed, at higher temperatures, the reaction probability for AlF and CaF production becomes more pronounced, as expected for reactions with a barrier. On the contrary, regarding by-products, $CaF_2$ is produced more efficiently than $AlF_2$ and $AlF_3$ in both gases, as demonstrated in panel (b) of Fig. 53.

Across the entire temperature range studied, the reaction probability for AlF and CaF formation via $NF_3$ remains higher than that via $SF_6$, as indicated in panel (b) of Fig. 53. Notably, this difference can be as large as an order of magnitude for specific temperatures. Therefore, Al/Ca + $NF_3 \rightarrow NF_2$ + AlF/CaF reactions exhibit a lower reaction barrier compared to Al/Ca + $SF_6 \rightarrow SF_5$ + AlF/CaF and Al/Ca reactions. Experimental determination of the activation energy for Al + $SF_6$ and Al + $NF_3$ reactions yields values of 9.5 kcal/mol (4781 K) and 5.99 kcal/mol (2990 K), respectively [231]. Consequently, using $NF_3$ in the buffer gas cell is expected to produce a brighter beam than using $SF_6$. Additionally, employing $NF_3$ reduces the number of by-products and potential contamination in the buffer gas cell.

The higher reaction probability of producing AlF or CaF molecules when using $NF_3$ can be attributed to the difference in the bond energy of F-atoms in $NF_3$ and $SF_6$. Specifically, the bond energy of F-atoms in $NF_3$ is 2.9 eV [234], which is 1.1 eV lower than the bond energy of F-atoms in $SF_6$ (4.0 eV) [235]. As a result, removing a fluorine atom from $NF_3$ is easier than from $SF_6$, leading to a higher reaction probability at a given temperature. Additionally, the bond energy of AlF (6.9 eV) is 1.4 eV larger than that of CaF (5.5 eV), leading to a higher probability of forming AlF molecules than CaF. This simple picture is further supported in Section 7.2.2, which delves into the role of stereochemistry.

The temperature dependence is analyzed in more detail in panels (a-b) and (d-e) of Fig. 54. These panels display the reaction probability as a function of collision energy and the initial angle of the metal atom, as introduced in Fig. 52. For the $SF_6$ reactions, the production of either AlF or by-products generally occurs at relatively high collision energies ($\gtrsim 0.8$ eV $\sim 9300$ K). In contrast, AlF is produced at much lower collision energies of $\gtrsim 0.4$ eV $\sim 4600$ K when $NF_3$ is used. The formation of reaction by-products is only observed at very high energies $\gtrsim 1.6$ eV ($\sim 18600$ K) in both $SF_6$ and $NF_3$. This suggests that such reactions require the activation of F atoms promoted by the collision with an Al atom. However, the production of by-products necessitates the activation of more F atoms, indicating a more complex reaction mechanism.

In the case of reactions involving Ca atoms, the production of CaF or $CaF_2$ occurs at similar collision energies ($\gtrsim 0.8$ eV) when reacting with $SF_6$. Interestingly, the reaction preferentially produces $CaF_2$ rather than CaF. Conversely, in the reaction of Ca + $NF_3$, CaF is produced at higher collision energies $\gtrsim 1.2$ eV ($\sim 13900$ K), while the production of $CaF_2$ occurs over a broader range of collision energies. These observations indicate that Al and Ca undergo distinctly different reaction mechanisms when reacting with $SF_6$ and $NF_3$.

### 7.2.2 Stereochemistry

The orientation of the reactants affects the efficiency of a given chemical reaction. These effects are due to subtleties in the underlying energy landscape of every molecular interaction, such as geometry effects or local equilibrium states. Here, by looking into the reaction probability as a function of the atom's angle of incidence, we can evaluate selectivity effects based on the orientation of the interacting partners, as presented in panels (a), (b), (d), and (e) of Fig. 54. Concretely, it informs us about the anisotropy of the interaction and the geometry effects on the interaction energy.

First, one notices that the production of AlF/CaF via $NF_3$ occurs over a wider range of angles than in the case of $SF_6$, indicating a more isotropic interaction. Similarly, when comparing the production of CaF and AlF in $SF_6$, we notice that CaF is formed only at incident angles close to 30°, whereas AlF is formed over a range of angles between 10° and 45°. The same effect, although not as pronounced, is observed in the case of $NF_3$. AlF is produced at angles between 0 and 90°, and CaF for angles between 20° and 90°. Therefore, AlF formation is less selective than CaF, which may impact the reaction probability.

### 7.2.3 Velocity distribution of the products

When AlF and CaF are produced through the reaction of either reactant gas, a substantial amount of energy, which can be several thousand kelvins, is released. This energy release is quite significant, as it is comparable to or even exceeds the collision energy involved in the reaction process. This energy can be carried away through the translational motion of the resulting products, or dissipated through the internal energy of the produced molecules. This internal energy includes electronic, vibrational, and rotational energies of the AlF and CaF molecules.

Panels (c) and (f) of Fig. 54 depict the velocity distribution of AlF and CaF. Notably, it is observed that independent of the colliding atom, $SF_6$ produces a narrower velocity distribution compared to $NF_3$. This phenomenon may arises from the higher bond energy of F-atoms in $SF_6$, which is 1.1eV greater than in $NF_3$, resulting in lower exothermicity. Furthermore, the presence of $NF_3$ as a reactant leads to a higher number of molecules with lower velocities. This behavior seems to be correlated with the reaction's stereochemistry: lower incident angles in Al/Ca + $NF_3$ reactions increase the reaction probability significantly, while Al/Ca + $SF_6$ reactions exhibit the highest reaction probability at larger incident angles. The lower incident angle facilitates a more efficient energy transfer between the metal atom and the F atom, as opposed to larger incident angles, where different F atoms and internal excitation of the F-containing molecule play a role in the dynamics. Therefore, employing $NF_3$ is preferable when aiming to obtain colder beams from a buffer gas cell. Finally, it is worth noting that the temperature has only a subtle influence on the most probable velocity (or distribution mode). This is expected, as the temperature exerts minimal impact on the reaction channels.

### 7.2.4 By-products

In this section, we delve into the investigation of by-products generated during the Al + $SF_6$/$NF_3$ and Ca + $SF_6$/$NF_3$ reactions. Specifically, the former reaction may produce $AlF_2$ and $AlF_3$ molecules as by-products, while the latter leads to $CaF_2$. The summarized results are presented in Fig. 55, which reveals distinct behaviors in the reaction probabilities for the formation of by-products between AlF and CaF.

When reactions involve $SF_6$, a higher reaction probability is observed for the formation of by-products compared to those involving $NF_3$. This difference can be attributed to the fact that $SF_6$ provides more available F atoms compared to $NF_3$, leading to increased opportunities for by-product formation.

Moreover, it is observed that reactions involving Ca exhibit a higher probability of generating by-products compared to reactions involving Al. This can be attributed to the fact that $CaF_2$ takes precedence as the predominant product in reactions involving Ca, with CaF being produced as a dissociation product of $CaF_2$. For a more comprehensive understanding and detailed information, please refer to Section III.E of [6].

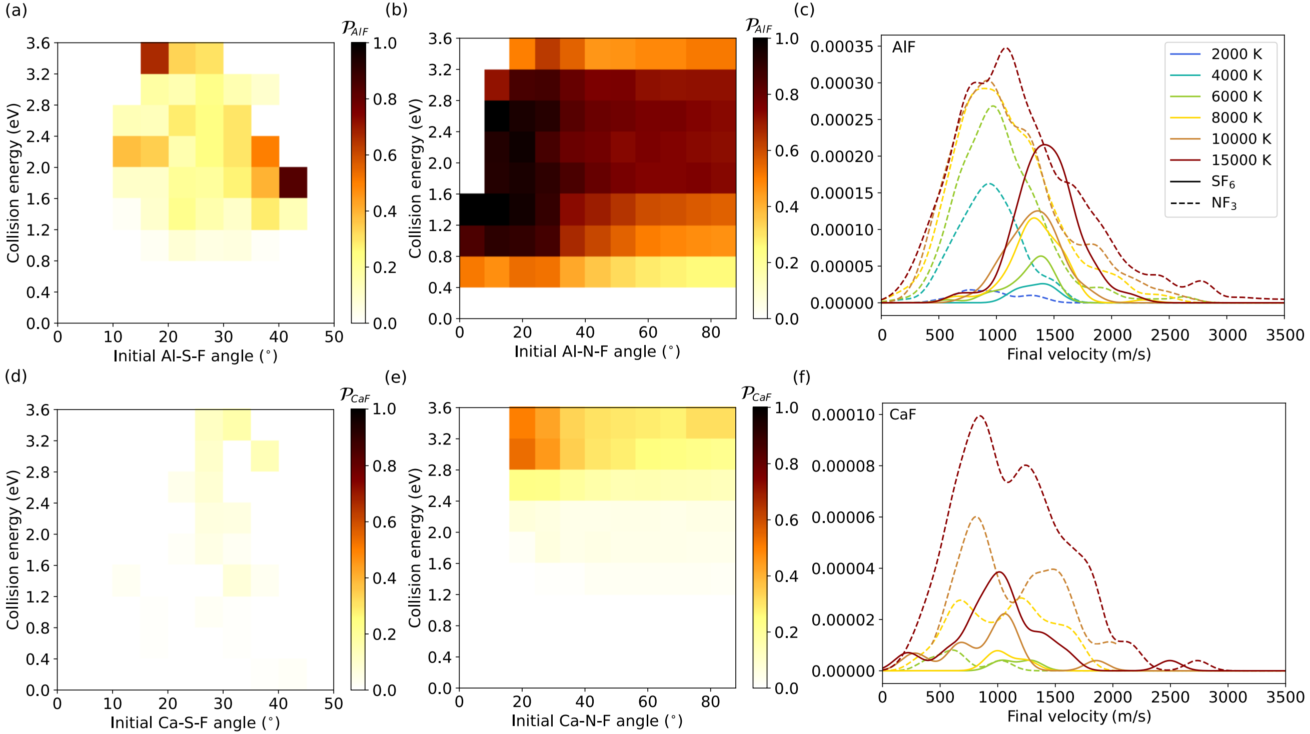


Figure 54: (a) Reaction probability to produce AlF by using $SF_6$. (b) Reaction probability of AlF using $NF_3$. (c) Velocity distribution of AlF, normalized to the corresponding reaction probabilities at different temperatures. (d) Reaction probability of CaF using $SF_6$.(e) Reaction probability of CaF using $NF_3$. (f) Velocity distribution of CaF, normalized to the corresponding reaction probabilities at different temperatures. Figure reproduced from ref. [228].

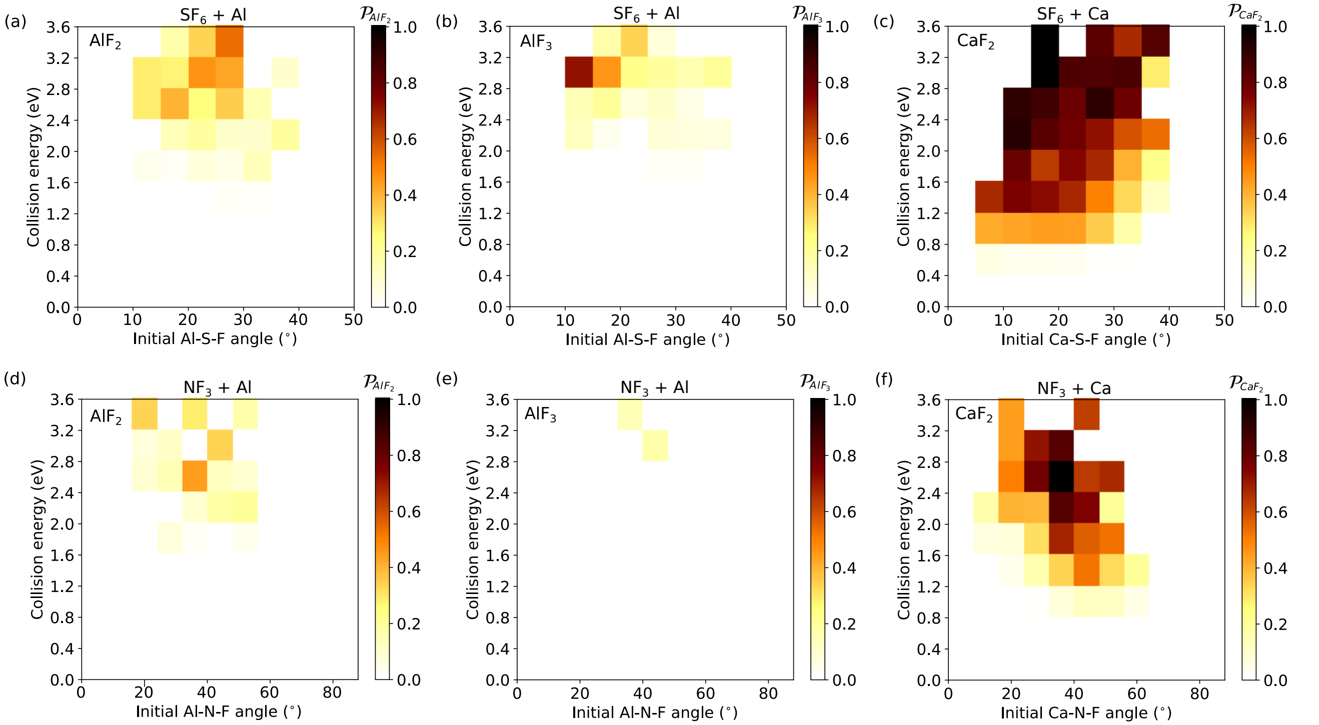


Figure 55: Reaction probability producing different by-products as a function of initial Al-S-F angle and collision energy in the SF$_6$/NF$_3$ + Al/Ca reactions. (a) Reaction probability of AlF$_2$ via SF$_6$ + Al. (b) Reaction probability of AlF$_3$ via SF$_6$ + Al. (c) Reaction probability of CaF$_2$ via SF$_6$ + Ca. (d) Reaction probability of AlF$_2$ via NF$_3$ + Al. (e) Reaction probability of AlF$_3$ via NF$_3$ + Al. (f) Reaction probability of CaF$_2$ via NF$_3$ + Ca. Figure reproduced from ref. [228].

## 7.3 Conclusion and outlook

In this work, we have shown that the probability of forming AlF and CaF via ablation of metal atoms in an F-containing reactant gas is higher when using $NF_3$ as a reactant gas than in the case of $SF_6$. Indeed, this effect seems to relate to the reaction's exothermicity: the more significant the difference between the binding energy of the product (AlF/CaF) and the bond energy of fluorine into the reactant molecule, the larger the reaction probability is. In particular, the exothermicity for AlF from $NF_3$, AlF from $SF_6$, CaF from $NF_3$ and CaF from $SF_6$ is given by 4.0 eV, 2.9 eV, 2.6 eV and 1.5 eV, respectively, in line with the observed probabilities shown in panel (a) of Fig. 66. These are further depicted in Fig. 56, in which the exothermic energy corresponds to the length of the arrows. Therefore, it is easier for a hot metal atom to take a fluorine atom from an $NF_3$ than from an $SF_6$ molecule. Moreover, Fig. 56 shows a detailed list of XF molecules, taken from Ref. [20] that can be formed via exothermic processes utilizing $NF_3$ and $SF_6$ gases. As a result, and concerning the lower binding energy of the N-F bond in $NF_3$, many XF molecules could be formed in a buffer gas source. In other words, $NF_3$, as the fluorine-donor gas, will help to explore more fluorine-containing diatomic molecules. Given this, $XeF_2$ is anticipated to be a good candidate as an F-atom donor based on its very low bond energy in comparison with AlF and CaF molecules.

We have shown that different fluorine-donor molecules in a buffer gas cell environment have an important impact on the target molecules' production efficiency and velocity distribution. In particular, we have demonstrated that the Ca/Al + $NF_3$ reaction is a more efficient route towards the production of CaF and AlF molecules than the Ca/Al + $SF_6$ one. Indeed, the difference in reaction probability can be as large as one order of magnitude. In addition, we have identified the main reaction mechanisms for those reactions using a tree-shaped reaction model. Our results indicate that the buffer gas cell's amount of by-products and possible contamination depends on the fluorine-donor molecule. The velocity distribution of the products depends drastically on the reactants. For instance, we notice that Ca/Al + $NF_3$ leads to a broader velocity distribution than Ca/Al + $SF_6$. The higher reaction efficiency with $NF_3$ means that a significantly lower flow rate of reactant gas can be used. This reduces contamination and the build-up of ice in the cell, thereby securing a more efficient thermalization of the molecules with the cryogenic helium. In addition, the lower velocity of the reactants reduced the number of collisions required to cool the molecules. A lower helium flow reduces the overall gas load in the system and allows for a faster extraction from the cell, which is highly advantageous for experiments

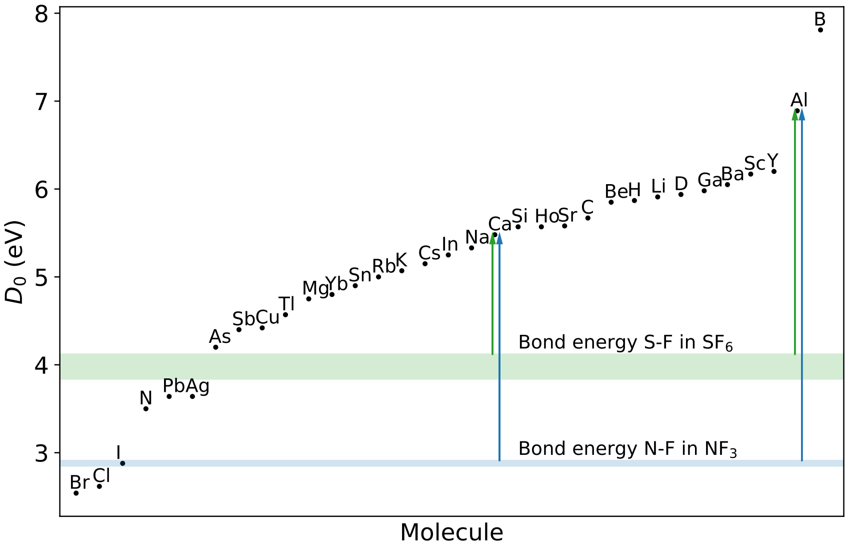


Figure 56: The dissociation energy of diatomic monofluoride molecules XF, for different atoms X, taken from Ref. [53]. The two red lines denote the limits for the bond energy of S-F in $SF_6$ whereas the black lines denote the same magnitude for $NF_3$. The arrows indicate the exothermic energy of the reactions to form CaF and AlF. Figure reproduced from ref. [228].

that are sensitive to collisions with helium or benefit from short molecular pulses.

Finally, we can conclude that, in general, it would be better to use $NF_3$ as a fluorine-donor molecule than $SF_6$ for forming metal-fluorine molecules. It is worth emphasizing that a better understanding of the chemistry in a buffer gas helps design better buffer gas cells, thereby achieving brighter and colder molecular beams, which are required to exploit the full potential of ultracold molecules.

# 8

# MOLECULAR DYNAMICS-DRIVEN GLOBAL POTENTIAL ENERGY SURFACES: APPLICATION TO THE ALF-ALF COMPLEX

## 8.1 Background

### 8.1.1 Motivation

The *ab initio* dynamics simulations of a system usually require information about the system's energy at certain atomic arrangements, which is the potential energy surface (PES) of the system. The concept of PES is critical for understanding atomic and molecular systems' spectroscopic and scattering properties. As a result, the development of accurate PESs is one of the most active research areas in chemical physics. For instance, in cold and ultracold molecule physics, it is necessary to develop efficient and general fitting techniques to calculate PESs for tetra-atomic systems relevant to calculating sticky collision lifetimes. Specifically, ultracold molecule-molecule collisions may lead to the formation of long-lived complexes, as observed in bi-alkali molecules [236, 237, 238, 239, 240, 241], resulting in a significant unexpected molecular loss [236, 242] and compromising the utility of ultracold molecules in many applications. So far, most theoretical efforts have been dedicated to bi-alkali systems. However, other molecules like CaF [243] or AlF [126] are getting more attention these days. Indeed, in the case of AlF-AlF, there is no available PES despite being one of the most prospective candidates for direct laser cooling. [1]

### 8.1.2 Potential energy surface construction via machine-learning methods

While the *ab initio* computation of energies for PES is of great importance, it is usually computationally demanding. An emerging alternative, which has gained popularity in recent years, involves applying machine learning techniques to mathematically model the relationship between the system's structure and its energy. Precisely, the energy of the system is computed using electronic structure methods for a defined collection of configurations utilized to train and validate a specific machine learning algorithm. Then, the trained model is harnessed to forecast the energy of novel configurations. Therefore, this approach avoids dealing with the Schrödinger equation and leverages the existence of reference data. It capitalizes on the notion that, in principle, when coupled with advanced machine learning methodolo-

[1] This chapter is written based on reference [244]: Xiangyue Liu, Weiqi Wang, and Jesús Pérez-Ríos. Molecular dynamics-driven global potential energy surfaces: Application to the AlF dimer. *The Journal of Chemical Physics*, 159(14):144103, 10 2023. https://doi.org/10.1063/5.0169080

gies featuring appropriate forms and parameters, virtually any real-valued function can be accurately fitted [92].

Various machine-learning methods have been developed to construct PESs, finding widespread applications in the fields of chemistry, physics, and materials science [245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267]. In the case of small systems comprising fewer than a few dozen atoms, machine-learning methods have been applied to fit high-level electronic structure reference data. Representative work includes the permutationally invariant polynomials (PIP) [253, 254, 255, 256], which aims to replicate the system's permutational symmetry. This approach has subsequently been expanded to fundamental invariants, where the dimension of representations is further reduced by excluding redundancies and incorporating only the irreducible secondary and primary invariants, enabling the investigation of larger systems [268]. For large systems, localized structural representations [258, 259, 260, 261, 262] have been developed. These representations capture local atomic environments while preserving symmetry invariance, rendering them suitable for investigating extensive systems. For comprehensive reviews on this topic, please refer to [269, 270, 271, 272, 270, 271, 273].

The construction of machine-learning PESs generally involves the following stages [272]:

- Preparing the reference data: The initial phase entails sampling the configurational space of interest utilizing specific sampling methods depending on the applications.
- Model establishment: In this step, the system's structures are firstly translated into inputs suitable for machine-learning regressors. Then, the models can be constructed with certain regressors using the reference datasets.
- Testing: The models are subjected to validation within validation sets and the optimal models are subsequently applied to various test set scenarios.

### 8.1.3 Structural representations

Machine-learning PESs map system geometries to their corresponding energy values, relying on structural information presented to the regression method in certain forms. A crucial aspect of a machine-learning potential is the preservation of energy invariance with respect to translation, rotation, and permutation [262]. However, these invariances are challenging to reproduce solely through the regression method, necessitating their accurate description in the input data for regression. Consequently, utilizing machine-learning regressors directly on Cartesian coordinates is not optimal due to the lack of translational and rotational invariance. This naturally leads to considering internal coordinates, such as interatomic distances, bending angles, and dihedral angles, as they remain unaffected by global rotation and translation. Nonetheless, permutational invariance must be reproduced in the structural representation.

For large systems such as extended systems, the success of machine learning methods is greatly indebted to the innovative strategy devised by Behler and Parrinello [258]. Their approach involves utilizing a structural representation named “symmetry functions” to summarize only the local environment of each atom within a cutoff sphere instead of considering the entire environment. This approach effectively addresses challenges posed in extended systems, where the replication of lattice cells can lead to innumerable dimensions. Meanwhile, this approach offers the flexibility of altering the number of atoms in the system from reference data to application, presenting a notable advantage.

In recent years, various forms of representation have been proposed. The atomic environment can be expanded by a series of basis sets of different forms to restore rotational and permutational invariance. For example, in the Smooth Overlap of Atomic Positions (SOAP) [260], the spherical harmonics basis is employed to characterize atomic environments. In a different vein, invariance can be explicitly incorporated into a hierarchy of $k$-body geometry functions for atom pairs, angles, and similar structural aspects, as seen in many representations, such as the many-body tensor representation [262]. A comprehensive assessment of the functional forms and the performance of several representations can be found in [274].

The situation is different for systems comprising only a small number of atoms. The overall dimensions of the system are sufficiently low to employ a structural representation encompassing all the atoms within the system. This offers the advantage of capturing structural information without any loss of information due to the truncation of long-range interactions, as is the case when using a local representation. In this work, we adopt a simple representation based on the geometry functions derived from internal coordinates.

### 8.1.4 Regression methods in machine-learning potentials

The structural representations can be seamlessly integrated with various regression methodologies to establish connections between structural information and energy. For instance, in the early applications of PIP, it was directly combined with linear least squares regression [253]. Subsequently, PIP was incorporated as input into neural networks [256, 257], leading to improved prediction capabilities. Notably, neural networks have emerged as potent tools for constructing PESs for both small molecules and materials, after the pioneering work of Behler and Parrinello [252, 258, 259, 268]. Gaussian Process Regression (GPR) has also found extensive applications across various systems [250, 251, 252, 255, 275, 260, 261]. In this study, GPR has been utilized to construct the PES of AlF-AlF. This approach offers the advantage of not only predicting energies but also providing associated uncertainties.

## 8.2 Method and computational details

### 8.2.1 Structural representation of AlF-AlF complex

When considering a system consisting of two diatomic molecules, it is a natural choice to use Jacobi coordinates to describe its geometry. As illustrated in Fig. 57, Jacobi coordinates consist of two interatomic distances $r_{\text{Al-F}}$, two azimuthal angles $\theta_1$ and $\theta_2$, one torsion angle $\phi$, and the intermolecular distance $R$ associated to the vector joining the center of mass of the two molecules.

However, Jacobi coordinates are not ideal for the structural representation to be fed to the machine-learning regressors. As introduced in 8.1.3, a representation should adhere to the symmetries that preserve or alter the energy. In this context, Jacobi coordinates maintain translational and rotational symmetry but do not maintain the system's permutational invariance.

For the AlF-AlF complex, the configuration can be described by

$$f(\mathbf{x}) = \hat{S}\{G(\mathbf{x},\mathbf{i})\}, \tag{27}$$

wherein $\mathbf{x}$ represents any geometry variable, and $i$ labels atoms within the molecular system. Here, to account for Coulomb and long-range interactions, we incorporate two-body interactions characterized by the inverse interatomic distances $\bar{r}ij^{-1}$ between atoms $i$ and $j$, as well as the Morse-like exponential terms $e^{-\bar{r}ij}$. In order to distinguish the chemical differences between different pairs of atoms, the interatomic distance $\bar{r}ij$ has been normalized by the equilibrium interatomic distance $r^*ij$, resulting in the expression $\bar{r}_{ij} = r_{ij}/r^*_{ij}$. The permutational invariance can simply be restored with a symmetrization operator $\hat{S}$ that sorts the two-body terms of the same atom pairs, i.e., the pair-wise distances are grouped by the chemical species.

Since the system consists of only four atoms and is non-periodic, reducing the dimensionality of the structural representation is unnecessary. Furthermore, we have assessed the effectiveness of higher-order representations, such as incorporating three-body polar angles and four-body dihedral angles. However, in the case of the AlF-AlF system, we observed that the improvement achieved through these additional representations is negligible.

### 8.2.2 The regressor

In this work, we utilize the GPR[92] as the regressor, to fit the relationship between the structural representation and the energy of the AlF-AlF complex.

We have tested different combinations of the kernels by analyzing the mean absolute error (MAE) and median absolute error of the test-set predic-

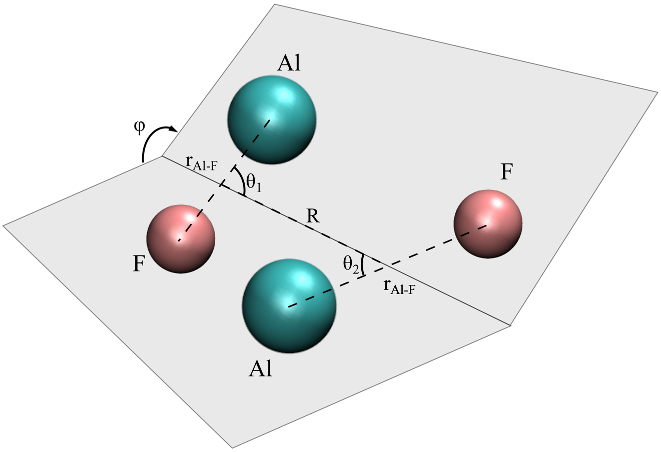


Figure 57: Jacobi coordinates for the AlF dimer. Figure reproduced from ref. [244].

tions. As a result, the Matérn kernel with $\nu = 5/2$ yields the most accurate results. A dot-product kernel can slightly improve the fitting. Furthermore, a white noise kernel has been applied to indicate the noise level of the training set, which is typically very small for quantum chemistry predictions. In our particular case, we have set the white noise level to be no smaller than $10^{-7}$.

### 8.2.3 The datasets

To generate the reference *ab initio* datasets, eight *ab initio* MD trajectories have been run within the canonical ensemble at 200 and 800 K using different initial configurations. At 200 K, the AlF-AlF system adopts stable configurations characterized by a dimer complex structure with short interatomic distances. However, at higher temperatures, such as 800 K, the MD simulation allows for sampling configurations with higher energies. This includes configurations like dissociated states where the molecules/atoms are no longer in close proximity to each other. The sampling of these unstable or metastable structures is vital for constructing the training set, as they capture the long-range interactions present in the system. Including these configurations in the training set helps to accurately capture the system's behavior in the long-range regions. Consequently, we have sampled 18,732 configurations relevant to short- and long-range interactions, with the intermolecular distance $R$ ranging from approximately 1.5 to 17.5 Å. The test set, on the other hand, comes from a different MD trajectory with 3633 steps, simulated at 800 K, to sufficiently encompass important configurations.

To test the performance of the machine-learning PES with different sizes, we have generated training sets from 2,000, 5,000, and 10,000 configurations randomly selected from the 18732 MD sampled configurations. We have also created two additional training sets by selecting 2,271 and 5,026 "landmarks". These landmarks correspond to configurations with the highest high-dimensional distances from the rest of the data points in the representation space. Ideally, they should be representative of the configurational space explored by MD simulations and, therefore, more efficient than the randomly sampled configurations to be used as the training set.

In earlier studies involving tetra-atomic systems such as NaK-NaK and CaF-CaF complexes, the training set was generated using a Latin hypercube sampling, which employed randomly selected equidistant grid points within the range of Jacobi and spherical coordinates [276, 243]. In our work, to assess the suitability of grid-based training sets for MD simulations, we have also generated an additional training set through the random hypercube sampling of the Jacobi coordinates of the dimer, as depicted in Fig. 57. Specifically, we have varied the Al-F bond length $r_{Al-F}$ and the intermolecular distance $R$ over a range from 2.5 to 25 and 1.5 to 75 Å, respectively. Additionally, the angular degrees of freedom $\theta_1$, $\theta_2$, and $\phi$ have been varied over a range of 0 to $\pi$.

### 8.2.4 *Ab initio* calculations

The *ab initio* energies have been calculated with the coupled-cluster theory with single, double, and perturbative triples [CCSD(T)] implemented in the Molpro package [187, 188]. The forces have been calculated with the second-order Møller– Plesset perturbation theory (MP2) level. The calculations were performed with the aug-cc-pVQZ basis set [6, 173, 5].

### 8.2.5 Active learning

The accuracy of a fitted PES is contingent upon several factors, including not only the chosen representation, regression methods, but also the composition of the training set. Although training sets can be generated through hypercube sampling [276, 243], there is a distinct advantage when the training and test sets sample similar regions within the configurational space. This congruence facilitates a seamless generalization of the trained model to specific applications.

However, certain scenarios, particularly when investigating dynamic properties at finite temperatures, necessitate the use of *ab initio* molecular dynamics (AIMD) simulations with thermostats. In such cases, the simulations can venture beyond the interpolation ranges, exploring regions not adequately represented by the training set. To address this challenge, active learning strategies have been harnessed to assimilate these "outliers" of the fitted PES [272, 277, 264, 265, 266, 267, 263, 275]. A noteworthy illustration of this concept is the real-time incorporation of additional *ab initio* training points during machine-learning-accelerated AIMD simulations [278, 279, 263, 267, 280].

On the other hand, certain regions of the PES pose greater challenges than others for the machine learning model. To reduce prediction uncertainty and enhance accuracy, the usual approach involves incorporating additional training data specifically in the challenging regions, so that the geometry-energy relationship can be constrained by the newly included data. To minimize additional computational expenses, one can benefit from an active learning scheme during MD simulations. Specifically, one only calculates the *ab initio* energies for configurations that cannot be accurately described by the PES model.

Various methodologies exist for implementing an active learning scheme in molecular dynamics simulations [272]. Among these, the most direct approach involves actively incorporating new *ab initio* points into the training set and subsequently retraining the model. This process is iterated until the model fulfills a specified convergence criterion. At that juncture, the model can be deemed to have attained a satisfactory level of accuracy, rendering it suitable for application to test sets or other tasks. Convergence criteria can take the form of metrics such as the average error (MAE, RMSE) on the test set or the maximum variance of predictions, which gauges the predictive uncertainty associated with the test set. However, in the context of this study, we adopt an alternative approach. In our case, the repulsive wall represents the most difficult region to handle. In the meantime, certain regions of the high-dimensional PES in the training sets may be sparsely sampled, leading to lower accuracy in those areas.

Fig. 58 depicts the active learning scheme implemented in this work. In the first step, an initial training set has been used to construct an initial PES model, which is then employed in the MD simulation. Then, during each MD step, the PES model predicts the energy of the current AlF-AlF configuration.[2] To decide if the energy of the current configuration requires *ab initio* calculation, we rely on a specific criterion. Typically, in GPR, inaccurate predictions are accompanied by relatively large prediction uncertainties, reflecting its lack of confidence in those specific regions of the PES. Hence, the prediction uncertainty can serve as a reliable criterion to determine whether a specific configuration can be accurately predicted. If the uncertainty exceeds a given threshold, the *ab initio* energy is calculated and added to the training set, thereby improving the overall PES model, which will be used for the next MD step. This iterative process continues until the MD sampling reaches convergence.

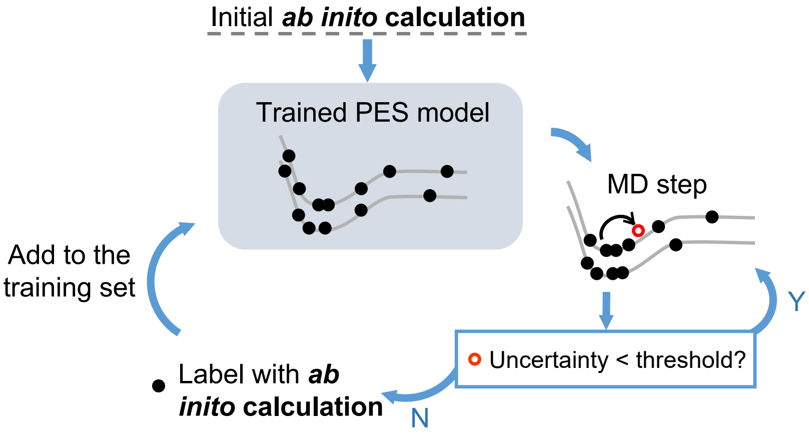


Figure 58: Schematic diagram of the active learning approach in this work. Starting from an initial trained model, the MD simulation is performed with the force calculated by the finite difference of energies predicted by the model. If the energy prediction uncertainty of the new MD step is larger than the threshold, then the configuration will be calculated with *ab initio* method and added to the training set. Figure reproduced from ref. [244].

[2] The energies used for finite difference calculation of the force, are also predicted by the PES model.

## 8.3 Results and discussions

### 8.3.1 Accuracy of the initial PES models

Before utilizing the active learning approach, we want to evaluate the accuracy of our approach in reproducing the CCSD(T)/aug-cc-pVQZ energies. This assessment will guide us in selecting an appropriate initial training set. Specifically, we are interested in the performance of the models as a function of the training set size, generated with the strategies described in Sec. 8.2.3.

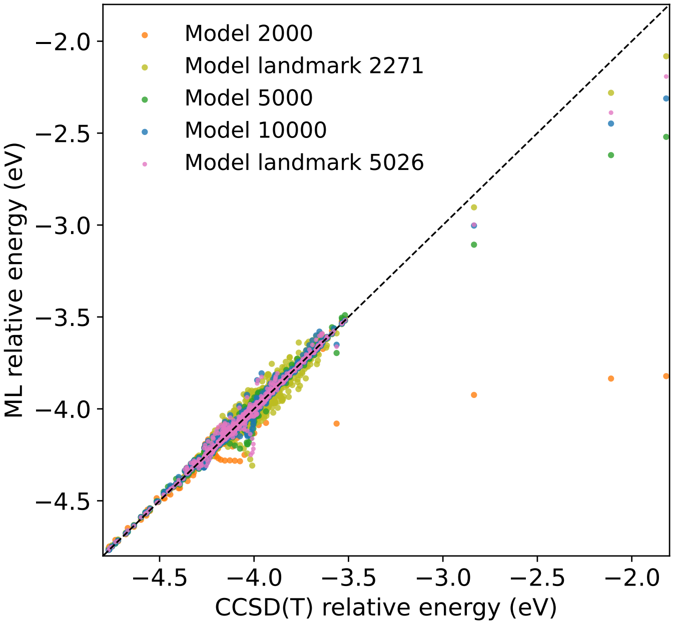


Figure 59: The GPR-predicted energy versus the CCSD(T)/aug-cc-pVQZ energy. Figure reproduced from ref. [244].

The comparison of model performance is presented in Fig. 59, tested on the test set obtained from an MD trajectory consisting of 3633 steps. As expected, when the training sets are composed of randomly selected configurations from the MD trajectories, increasing the size of the training set enhances the model's performance. By increasing the training set size from 2000 to 5000 configurations, the error is reduced by approximately 25%, resulting in mean and median absolute errors of 0.78 meV/atom and 0.018 meV/atom, respectively. This demonstrates the high accuracy achieved by the models with more training data. When using 10,000 training configurations, the mean and median absolute errors no longer exhibit further improvement. However, predictions on the outliers with the highest energies get closer to the reference energy, indicating the model's ability to better capture extreme cases.

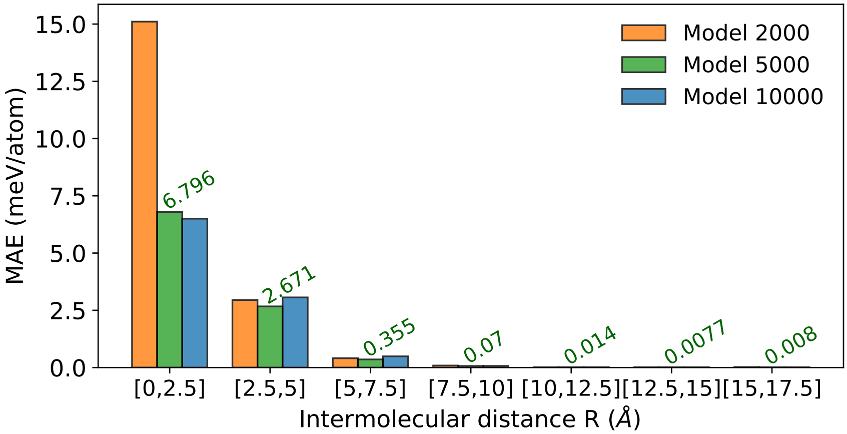


Figure 60: MAE in different regions of intermolecular distance $R$, tested on the test set with 3633 MD steps. The MAE from the model with 5000 training data is labeled in green. Figure reproduced from ref. [244].

Despite the impressive overall accuracy of the models, they exhibit lower accuracy for configurations resulting in high interaction energies. These configurations correspond to the repulsive region at short intermolecular distances $R$. In this region, the model encounters challenges in accurately predicting the interactions due to the extreme nature of the repulsive forces and the small number of configurations within the training data in these specific high-energy regions. To gain further insights into the accuracy of our model with respect to $R$, we calculated the mean absolute error for different $R$ regions, and the results are displayed in Fig.60. This Figure reveals that the region $R \in [0, 2.5]$ Å exhibits the highest mean absolute error across all models. On the contrary, for $R > 7.5$ Å, the errors are below 0.1 meV/atom for all models. These effects are better shown in the more comprehensive analysis presented in Table 10, which reports the mean and median absolute errors, respectively, both of which imply a scarcity of

Table 10: The mean absolute error and median absolute error for the test set with 3633 MD steps, presented in meV/atom, with the errors reported for the entire test set, as well as for configurations within different intermolecular ranges of $R$ (in Å).

| | Mean absolute error (meV/atom) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Range (Å) | all | [0,2.5] | [2.5,5] | [5,7.5] | [7.5,10] | [10,12.5] | [12.5,15] | [15,17.5] |
| Model 2000 | 1.18 | 15.1 | 2.9 | 0.41 | 0.088 | 0.013 | 0.0088 | 0.022 |
| Model 5000 | 0.78 | 6.80 | 2.67 | 0.36 | 0.070 | 0.014 | 0.0077 | 0.0080 |
| Model 10000 | 0.85 | 6.50 | 3.06 | 0.49 | 0.071 | 0.014 | 0.0077 | 0.0071 |
| Landmark 2271 | 0.81 | 4.70 | 3.16 | 0.52 | 0.074 | 0.032 | 0.023 | 0.027 |
| Landmark 5026 | 0.85 | 5.99 | 3.18 | 0.51 | 0.065 | 0.016 | 0.012 | 0.012 |
| Hypercube 5000 | 15.81 | 180.61 | 41.10 | 9.70 | 0.865 | 0.300 | 0.058 | 0.110 |
| | Median absolute error (meV/atom) | | | | | | | |
| Range (Å) | all | [0,2.5] | [2.5,5] | [5,7.5] | [7.5,10] | [10,12.5] | [12.5,15] | [15,17.5] |
| Model 2000 | 0.024 | 4.49 | 1.51 | 0.25 | 0.048 | 0.009 | 0.0068 | 0.0069 |
| Model 5000 | 0.018 | 3.05 | 1.37 | 0.19 | 0.031 | 0.010 | 0.0055 | 0.0047 |
| Model 10000 | 0.019 | 2.74 | 1.65 | 0.19 | 0.071 | 0.010 | 0.0052 | 0.0039 |
| Landmark 2271 | 0.044 | 2.24 | 1.57 | 0.27 | 0.047 | 0.021 | 0.019 | 0.021 |
| Landmark 5026 | 0.024 | 2.14 | 1.63 | 0.22 | 0.034 | 0.010 | 0.0093 | 0.0089 |
| Hypercube 5000 | 0.13 | 27.69 | 8.64 | 1.20 | 0.184 | 0.062 | 0.030 | 0.040 |

configurations within the $R \in [0, 2.5]$ Å range. This scarcity of data points in the $R \in [0, 2.5]$ Å region leads to higher errors. This observation leads us to adopt an active learning approach since it requires additional *ab initio* training data in these regions.

As summarized in Table 10, we noted a substantial disparity in performance between the model trained on 5000 hypercubic points and the models generated from MD data. The errors observed in the former case are more than tenfold higher than those observed in the latter. This contrast can be attributed to the inherent nature of molecular dynamics, which samples the configurational space in a manner significantly distinct from a random distribution.

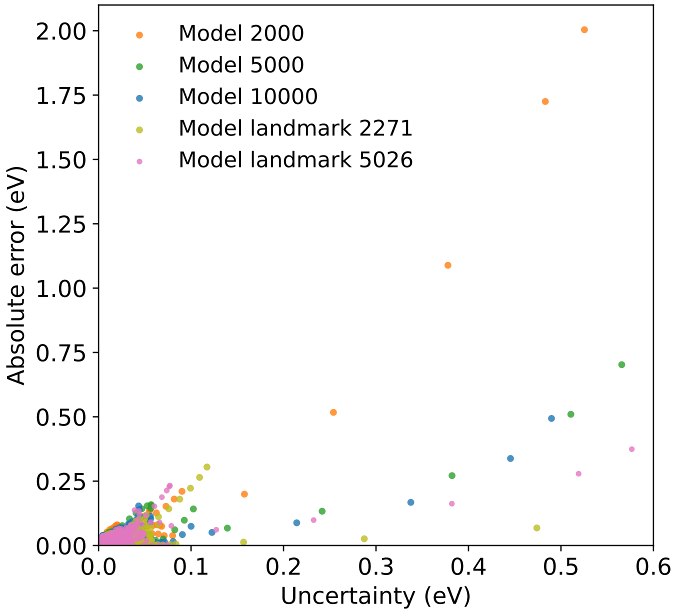


Figure 61: The relationship between the absolute error and the uncertainty of GPR prediction. Figure reproduced from ref. [244].

In addition, as introduced in Sec. 8.2.3, two training sets have been constructed with 2,271 and 5,026 landmarks. We aim to use minimum training data in this approach, assuming that the landmarks are representative of the sampled configurational space. Indeed, we observe that the training sets with 2,271 and 5,026 landmarks achieve comparable overall accuracy to the models constructed from the training set with 10,000 randomly selected configurations. However, in the long-range tail of the PES, models with landmarks show larger errors than the ones trained with randomly selected configurations. This effect could be related to the possibility that configurations in the long-range of the PES exhibit greater similarity in the representation space, resulting in a smaller percentage of configurations being selected from the long-range compared to the short-range. Consequently, the models might not allocate sufficient weight to the long-range data during training. In fact, in machine-learning applications in other fields, such as image processing, it has been noted that incorporating similar training data with slight variations can enhance model accuracy and reduce overfitting [281]. Furthermore, considering that MD simulations extensively sample configurations associated with repulsive short-range interactions at high temperatures, it raises doubts about the suitability of landmark models for simulation purposes. Hence, it is more precise to utilize the most pertinent configurations visited by the MD simulations as training points.

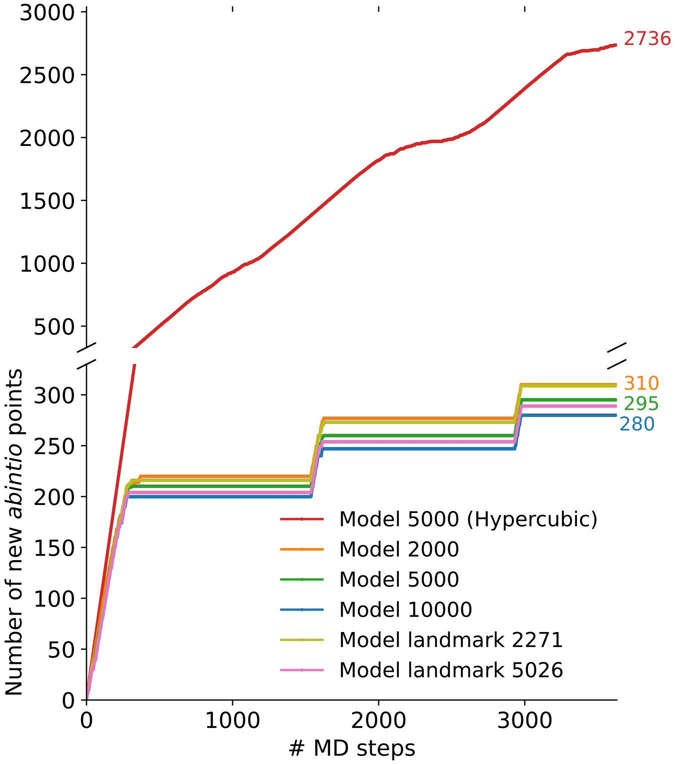


Figure 62: Number of new *ab initio* points additionally required in active-learning during the MD simulation, tested on the MD trajectory with 3633 steps. The uncertainty $> 0.01$ eV has been used as the criterion for additional *ab initio* calculations. Figure reproduced from ref. [244].

So far, we have compared the accuracy of different models using randomly selected training data and landmarks. However, we still need to determine the optimal model for the initial PES model for active learning. Our interest lies in the accuracy of various models and their efficiency, which can be quantified by the number of additional *ab initio* calculations required during MD simulations. To address this, we have implemented our active learning approach with different initial training sets and exposed the models to the MD trajectory with 3,633 MD steps. The results are shown in Fig. 62, with the uncertainty threshold set to be 0.01 eV.

The models trained on configurations randomly selected from eight MD trajectories necessitate approximately 10% of the configurations to be calculated via *ab initio* calculations. As anticipated, models with more extensive training data require fewer additional *ab initio* calculations. We observed a convergence pattern of new *ab initio* points concerning the number of MD steps. This suggests that in longer simulations, it is likely that fewer than 10% of the configurations will be required to be calculated *ab initio*. Similarly, training sets with landmark configurations exhibit comparable efficiency. However, they require slightly more *ab initio* calculations. In contrast, the initial model trained on hypercubic grids demands the *ab initio* evaluation of almost 65% of the configurations, indicating the poor efficiency of Latin hypercube sampled training sets particularly in the context of MD simulations. Indeed, the dissimilarity in the distribution of sampled configurations in the configurational space between MD simulations and hypercubic sampling may account for this variation. Consequently, employing Latin hypercube sampling would lead to prolonged computational times, while adopting the PES models from MD training sets would yield a more efficient approach.

### 8.3.2 Implementation in Realistic MD Simulations

The previous tests were conducted on a test set consisting of 3,633 configurations from a short MD trajectory. In practical applications, MD simulations can take much longer to sufficiently sample the configurational space and achieve equilibrium and achieve equilibrium. For example, a typical MD simulation for estimating the AlF-AlF complex lifetime should be at least 10 times longer than its lifetime, which can be several nanoseconds. To further evaluate the feasibility of our approach in realistic simulations, we have performed a simulation using the replica-exchange molecular dynamics (REMD) method [282, 283]. REMD is an enhanced sampling technique that efficiently facilitates the attainment of dissociation equilibrium at low temperatures, effectively reducing the simulation time required. In this study, we have simultaneously simulated 10 trajectories at temperatures ranging from 200 K to 1000 K. Each of the ten replicas has been run for a total of 5.4 ns, resulting in a comprehensive simulation time. For the REMD simulation, we have employed the active learning approach described above, using an initial PES model trained with 22,365 configurations obtained from all nine MD trajectories.

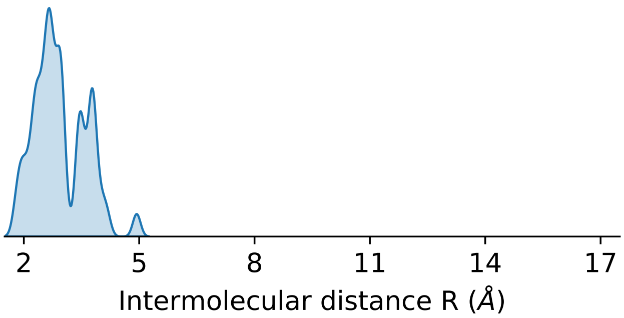


Figure 63: Distribution of intermolecular distance *R* of the configurations requiring additional *ab initio* calculations in the MD simulation of AlF-AlF complex. Figure reproduced from ref. [244].

During the REMD simulation, only a small subset of 2,038 configurations (approx. 0.008%) have been selected out of 26,891,350 REMD steps for further CCSD(T) calculations. These selections were based on a prediction uncertainty criterion of 0.05 eV. The chosen configurations are mainly concentrated in the regions with intermolecular distances (*R*) below 5 Å, as shown in Fig. 63. As a result, we have a PES model training on 24,403 configurations.

To assess the accuracy of the model, a comparison has been made between the CCSD(T) reference energies and the model predictions along a one-dimensional profile of the PES, as shown in Fig. 64. Even though the uncertainty is more significant in the repulsive region than in the long-range interactions, the model's overall precision remains remarkably high. This underscores the model's reliability and ability to capture the system's behavior across the entire PES accurately.

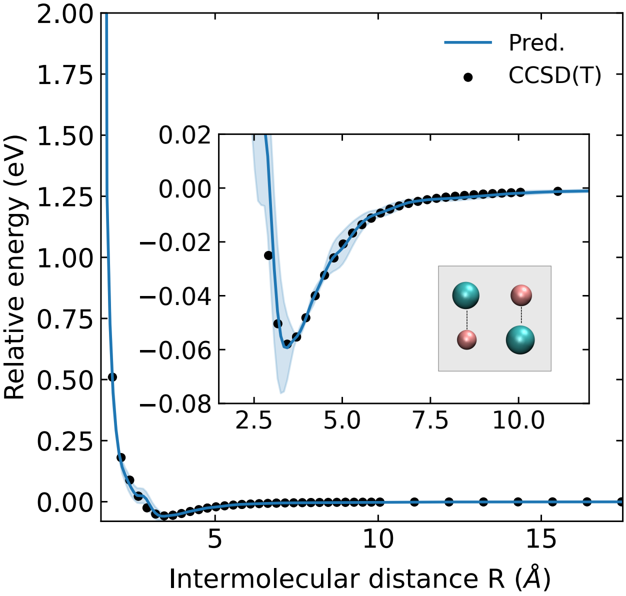


Figure 64: An one-dimensional profile of the PES, showing the *ab initio* and predicted energies with the prediction uncertainty indicated by shaded regions. The model is trained with a training set of 24,403 configurations. The total energy is referenced to the energy of the dissociated AlF-AlF complex. Figure reproduced from ref. [244].

### 8.3.3 Properties of the PES

Table 11 provides an overview of the general properties of the *ab initio* PES. It is particularly noteworthy that the reactants must overcome 11.56 eV of energy to undergo the transformation: AlF + AlF → $Al_2$ + $F_2$. As a result, this reaction is highly endothermic.

Table 11: The CCSD(T)/aug-cc-pVQZ relative energies of various configurations. The energies are referenced to the potential energy of two AlF molecules with the intermolecular distance $R = 20$ Å. In this case, the geometry of AlF molecule has been fixed to its CCSD(T)-optimized geometry with $d$(AlF) = 1.669 Å.

| Configuration | Relative energy (eV) |
|---|---|
| AlF + AlF ($d$(AlF-AlF) = 20 Å) | 0.0 |
| AlF-AlF complex | -0.696 |
| $Al_2$ + $F_2$ ($d$($Al_2$-$F_2$) = 20 Å) | 11.561 |
| Dissociated 4 atoms ($d$(Al-F) = 20 Å) | 18.167 |

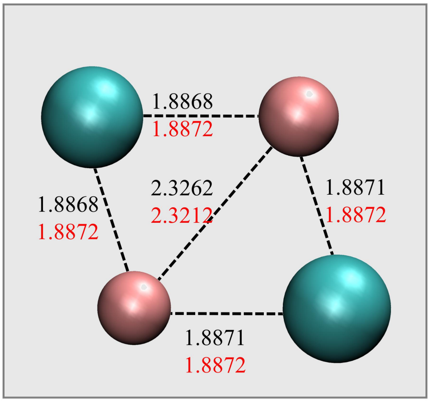


Figure 65: Configurations of AlF-AlF complex, optimized from the ML-fitted PES (black), and compared with CCSD(T)/aug-cc-pVQZ optimized geometry (red). The ML-predicted relative energy of the complex is -0.695 eV, referenced to the potential energy of two AlF molecules with the intermolecular distance $R = 20$ Å.

Based on the PES model, the stable configuration of the AlF-AlF complex has been optimized using the Fast Inertial Relaxation Engine (FIRE) method [284]. The resulting geometry exhibits a $D_{2h}$ symmetry, with all Al-F bond lengths being almost identical, as shown in Fig. 65. This geometry closely matches the CCSD(T)-optimized structure, with very small differences in Al-F bond lengths less than 0.0004 Å, while the predicted energy is only

0.0025 eV higher than the CCSD(T) result. As a result, the binding energy of the AlF-AlF complex is 0.69 eV.

Notably, there is no barrier for the atom-exchange reaction $Al^{(1)}F^{(2)} + Al^{(3)}F^{(4)} = Al^{(1)}F^{(4)} + Al^{(3)}F^{(2)}$. Indeed, this reaction has been frequently observed during the reported MD simulations of the AlF dimer.

An intriguing observation is that the $D_{2h}$ configuration stands as the sole stable configuration in the PES, which is in contrast to the behavior observed in CaF[243] or NaK [276]. Irrespective of the initial configurations employed, the geometry optimization of AlF-AlF consistently converges to the stable $D_{2h}$ configuration. In contrast, other configurations, such as the local minimum configuration found for the CaF-CaF complex with $C_s$ symmetry, are unstable for the AlF-AlF complex.

### 8.3.4 Discussions

While there are ongoing discussions about how many training points are needed to create a PES, this necessity varies greatly based on the system being studied. When performing molecular dynamics simulations for the AlF-AlF complex, we have found it crucial to introduce several thousand training points to represent the behavior of the PES. This is especially important in the repulsive region. In this study, active learning throughout the molecular dynamics simulations guarantees the integration of essential *ab initio* points into the training set. As depicted in Fig. 63, the molecular dynamics simulation process (with 26,891,350 REMD steps) has involved the inclusion of more than 3,000 additional *ab initio* points within the repulsive region, even if the initial training set already contains more than 20,000 configurations. Nevertheless, it is worth noting that in some instances, like NaK-NaK, a smaller number of *ab initio* training points can still lead to a favorable fitting of the PES [276]. In fact, for the NaK+NaK system, a mere 2,000 hypercube training points prove adequate to yield errors spanning from 4 to 140 $cm^{-1}$ (equivalent to 0.1-4 meV/atom). In contrast, when employing the same approach for the CaF+CaF system, the errors span a broader range, reaching between 500 and 1,500 $cm^{-1}$ [243].

## 8.4 Conclusion and outlook

In this research, we have relied on absolute average errors as a metric to establish the accuracy of the PES. Nonetheless, arguments could suggest that these errors hold significance primarily at extremely short distances, where the potential exhibits strong repulsion. In other words, these errors reflect a relatively minor deviation in the position of the repulsive potential. However, utilizing absolute errors is more appropriate than relying on relative errors. Unlike quasi-classical trajectory simulations, molecular dynamics simulations frequently explore the repulsive regions. Consequently, even a small error in these regions can lead to a substantial discrepancy within the sampled distribution.

In summary, in this study, we have developed an ML method for fitting tetra-atomic PESs. The approach utilizes the most relevant configurations from MD simulations at certain temperatures as the training set, which is then used to train a GPR model for predicting the energy of new configurations. This adaptable process allows the method to meet specific threshold criteria and achieve the required accuracy for a given system. Consequently, highly accurate tetra-atomic PESs can be constructed while calculating only a small fraction (less than 0.1%) of the configurations *ab initio*, resulting in precise outcomes with minimal computational effort. Combining this approach with a representation that considers the system's symmetry and locality makes it feasible to extend the method to larger systems.

To demonstrate the viability of the method, it was applied to the AlF-AlF system as a proof of concept. Surprisingly, a single minimum was found, in contrast to previous findings in bi-alkali molecules or for the CaF dimer. The computed binding energy for the four-body complex was determined to be 0.69 eV. This result suggests that the AlF-AlF complex may exhibit a significantly different lifetime compared to CaF or bi-alkali systems.

Further extensions of this work encompass several promising avenues:

- Expanding the approach to other tetra-atomic systems: The current methodology can be applied to investigate and construct PESs for various tetra-atomic systems, such as AlCl, CaCl, etc. This extension would provide valuable insights into the behavior of different molecular complexes and enable a comprehensive understanding of their properties.

- Comparing the present approach with different structural representations and regressors: It would be valuable to compare the performance of the proposed method with alternative structural representations and regression techniques, including neural networks.

Specifically, exploring the use of Deep Gaussian Processes, which are deep neural networks based on Gaussian process regressions, could yield valuable insights into the predictive capabilities of such models.

- Utilizing multi-output/multi-task Gaussian processes to predict both energy and force: By employing multi-output/multi-task Gaussian processes, it becomes possible to predict both the energy and force simultaneously. This approach leverages the physical relationship between energy and force, leading to a more constrained model that offers higher accuracy and better adherence to physical behavior.
- Incorporating derivative observations for force prediction: Integrating derivative observations into the Gaussian process model for force prediction can lead to enhanced accuracy and reduced prediction uncertainty. This addition can improve the model's ability to capture fine-grained details in the PES and provide more accurate force predictions.
- Exploring the possibility of transfer-learning, i.e. to acquire CCSD(T) energy information from Hartree-Fock or density functional theory energies with cheaper basis sets.

By pursuing these extensions, the method's capabilities can be further refined and its applicability to a broader range of molecular systems with various properties can be explored.

9

# CONCLUSION

In this thesis, we investigate the spectroscopic characteristics and chemical behavior of diatomic molecules, showcasing their potential for applications in various fields such as quantum information sciences, precise measurement of physics constants, and cold and ultracold chemistry.

This thesis consists of two main parts. The first part (**Chapter 2** to **Chapter 6**) focuses on diatomic molecule spectroscopic constants, introducing the construction of databases of diatomic molecule spectroscopic constants and electric dipole moments. Based on these databases, we have revealed relationships between the spectroscopic constants and benchmarked the accuracy of quantum chemistry methods. The second part (**Chapter 7** and **Chapter 8**) delves into the chemistry of diatomic fluorides, particularly AlF and CaF, exploring their production efficiency and presenting a machine-learning method for fitting the AlF-AlF system's potential energy surface. Specifically:

- **Chapter 2** of the thesis implements the diatomic molecule spectroscopic database, which provides accessible spectroscopic constants and Franck-Condon factors for polar diatomic molecules in both ground and excited states. This dynamic database allows registered users to contribute to the collection of spectroscopic data, leading to its growth since its development in 2020.
- **Chapter 3** demonstrates the relationships between key spectroscopic constants of diatomic molecules, uncovered by machine-learning models. These relationships appear to be largely independent of the chemical bond nature. This study challenges the perception of machine-learning methods as mere black-box fitting techniques. It extracts valuable insights, demonstrating that accurate quantitative predictions of spectroscopic constants can be made based on the group and period of constituent atoms. Moreover, the research reveals dependencies of spectroscopic constants on the number of valence electrons and electron shells in the molecule's atoms. This knowledge opens avenues for predicting Franck-Condon factors crucial for transitions influencing ultracold molecule cooling processes.
- **Chapter 4** extends this work by constructing a dataset of contemporary experimental electric dipole moments. The machine-learning model effectively establishes a relationship between the ground state dipole moments of diatomic molecules and their spectroscopic constants. Based on these relationships, predictions of dipole moments are achieved without the need for quantum chemistry calculations.

This success is attributed to the incorporation of atomic features, including electron affinity and ionic potential, in conjunction with molecular spectroscopic constants, notably the equilibrium internuclear distance and harmonic vibrational frequency. Moreover, this approach challenges the conventional notion that electronegativity differences alone can describe dipole moments, emphasizing their intricate correlation with chemical bonding. The valuable insights gained from this research are made possible through the development of a comprehensive and unbiased dataset, highlighting the significance of robust data in advancing our understanding of diatomic molecules.

- Based on the accurate experimental electric dipole moment dataset, **Chapter 5** conducts a rigorous assessment of advanced quantum chemistry methods, with a particular focus on the accuracy of coupled-cluster with single, double, and perturbative triple excitations [CCSD(T)]. This method has long served as a reference for the implementation of quantum chemistry methods. The study compares computational results, particularly those obtained using large augmented correlation-consistent basis sets, and the segmented basis sets, focusing on diatomic molecules with diverse bonding characteristics and elemental compositions. The findings demonstrate that single-reference CCSD(T) calculations, coupled with specific basis sets, generally provide satisfactory descriptions of dipole moments, especially for molecules composed solely of main-group elements. Additionally, the study highlights the significance of considering deviations in electron distribution, rather than bond lengths, in understanding errors in dipole moment predictions. Furthermore, **Chapter 6** involves the calculation and comparison of hyperfine constants for the $a^3\Pi$ state of aluminum monofluoride (AlF) with experimental values. Our research highlights the importance of conducting a comprehensive assessment that incorporates both experimental and theoretical approaches.

- **Chapter 7** delves into the chemical reactions responsible for producing AlF and CaF molecules. These species have gained significant interest in experiments involving laser cooling and trapping of cold molecules. The study compares the effectiveness of various fluorine-donor molecules in generating AlF and CaF through metal atom ablation within a buffer gas cell. The results demonstrate that using $NF_3$ as a reactant gas leads to a higher reaction probability in forming AlF and CaF through metal atom ablation than using $SF_6$. This effect is attributed to the reaction's exothermicity, which is influenced by the difference between product binding energy and reactant molecule bond energy. Additionally, the velocity distribution of products varies depending on the reactants: $NF_3$ leads to a broader distribution. Overall, the results advocate for the use of $NF_3$ over $SF_6$ as a valuable fluorine-donor gas for exploring a wider range of fluorine-containing diatomic molecules. The study highlights the importance of understanding buffer gas chemistry for optimizing molecular beam production.

- **Chapter 8** introduces a machine learning (ML) method for fitting tetra-atomic potential energy surfaces (PESs), specifically applied to the AlF-AlF system. The approach employs relevant configurations obtained from molecular dynamics simulations at specific temperatures at the CCSD(T) level as the training set, which is then utilized to train a machine learning model. This PES model accurately predicts the energy of new configurations, enabling the construction of highly precise PESs. Aided by an active learning scheme, this method achieves exceptional accuracy for the system while only requiring calculations for a very small fraction (less than 0.1%) of the configurations from scratch, thereby minimizing computational effort. Interestingly, the PES reveals the presence of a single minimum, contrary to previous findings in bi-alkali molecules or the CaF dimer. The computed binding energy for the four-body complex is 0.69 eV. This result suggests that the AlF-AlF complex may have a substantially different lifetime than systems involving CaF or bi-alkali molecules, indicating potentially unique chemical behavior. By incorporating considerations for the system's symmetry and locality, this approach can potentially be extended to larger systems.

# APPENDIX

# Appendix: Gaussian process regression

In this Appendix, we briefly review the Gaussian process regression (GPR) method, mainly following Ref. [92].

The core principle of GPR is its probabilistic nature. Rather than offering a fixed functional form that precisely matches the data, GPR provides a range of functions within a distribution that effectively fits the data, assuming that the observation is generally not perfect in reality. The resultant fitted model provides not only the prediction of target values but also measures of uncertainty. Here, it is important to notice that, even when the data points are "absolutely exact" in the absence of noise, still the probabilistic framework provides a probability distribution, or confidence intervals associated with the predictions. This concept is important because generally the underlying true function is unknown, and we can only get models that fit the observed data.

The probabilistic feature of GPR is achieved within the Bayesian framework. Initially, a prior distribution of functions is postulated, incorporating any prior knowledge or assumptions about the underlying function. After observing the data, this prior belief is revised, leading to the derivation of a posterior distribution that encapsulates the updated understanding of the function.

Under the Bayesian linear regression framework, the relationship between a set of inputs $\boldsymbol{X} \in R^D$ and the noisy observation $y \in R$ can be modeled with

$$y = f((x)) = \boldsymbol{w}^T \boldsymbol{x} + \varepsilon, \tag{28}$$

where $\varepsilon \sim N(0, \sigma)$, $\boldsymbol{w} \sim N(0, \boldsymbol{\Sigma})$, both obey the Gaussian distribution. After exposing this model to a known dataset $D = (X \in R^D, y \in R)$, it is possible to predict the value of target $\hat{y}$ with a new input $\hat{x}$ according to the "marginal likelihood" $p(y|X)$, which is defined as "the integral of the likelihood times the prior" [92]

$$p(\hat{y}|\hat{\boldsymbol{x}}, D) = \int p(\hat{y}|\hat{\boldsymbol{x}}, \boldsymbol{w}) p(\boldsymbol{w}|D) d\boldsymbol{w} \tag{29}$$

It has been shown that the prediction of $\hat{y}$ with a new input feature $\hat{\boldsymbol{x}}$ also follows a Gaussian distribution, that

$$\hat{y} \sim N(\hat{\mu}, \hat{\sigma}) \tag{30}$$

with the prediction mean $\hat{\mu} = \sigma^{-2}\hat{\boldsymbol{x}}^T(\sigma^{-2}\boldsymbol{X}\boldsymbol{X}^T + \Sigma^{-1})^{-1}\boldsymbol{X}y$ and the variance $\hat{\sigma} = \sigma^2 + \hat{\boldsymbol{x}}^T(\sigma^{-2}\boldsymbol{X}\boldsymbol{X}^T + \Sigma^{-1})^{-1}\hat{\boldsymbol{x}}$. They can be rewritten in a form that contains only inner products between data points as

$$\hat{\mu} = k(\hat{\boldsymbol{x}},\boldsymbol{X})(k(\boldsymbol{X},\boldsymbol{X}) + \sigma^2\boldsymbol{I}) \tag{31}$$

and

$$\hat{\sigma} = \sigma^2 + k(\hat{\boldsymbol{x}},\hat{\boldsymbol{x}}) - k(\hat{\boldsymbol{x}},\boldsymbol{X})(\sigma^2\boldsymbol{I} + k(\boldsymbol{X},\boldsymbol{X}))^{-1}k(\boldsymbol{X},\hat{\boldsymbol{x}}) \tag{32}$$

This process is called “kernelization”. Indeed, in GPR and other kernel-based approaches, the kernel function (or covariance function) $k(x_i,x_j)$ is a pivotal component. It actually measures the “similarity” in the input space by definition. The significance of the similarity between inputs in kernel methods lies in the expectation that as the inputs $x_i$ and $x_j$ draw closer, the functions describing the input-output relationship, $y_i \sim f_i((x)_i)$ and $y_j \sim f_j((x)_j)$, will exhibit greater similarity, and vice versa. The form of the kernel function can be pre-defined when establishing the prior distribution. Subsequently, during the optimization process of GPR models, the kernel’s parameters can be adjusted to better align with the observed data. Typically, there is a parameter named “characteristic length scale”, which controls how rapidly the similarity between $\boldsymbol{x} - y$ relationship decays as the distance between inputs becomes larger. When the length scale is too small, the model can suffer from an overfitting problem.

Specifically, in Chapter 3, when learning the equilibrium interatomic distance $R_e$, the employed covariance function is the exponential kernel, defined as

$$k(\boldsymbol{x}_i,\boldsymbol{x}_j|\theta) = \sigma_f^2 \exp\left(-\frac{r}{l}\right), \tag{33}$$

where $\sigma_f$ is the signal variance, $l$ is the characteristic length scale, and $r$ is the Euclidean distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

When learning the harmonic vibrational frequency $\omega_e$, we utilize the Matérn class of covariance functions [92],

$$k_{\text{Matern}}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)}\frac{\sqrt{2}r}{l}^{\nu} K_\nu \frac{\sqrt{2}r}{l}, \tag{34}$$

with $\nu = 5/2$. $K_\nu$ is modified Bessel function in D dimensions, $r$ is the Euclidean distance between $x$ and $x'$, then the Matern 5/2 kernel function is

$$k_{\nu=5/2}(r) = \left(1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2}\right) \exp\left(-\frac{\sqrt{5}r}{l}\right). \tag{35}$$

In learning $R_e$, we employ linear basis functions. However, when dealing with $\omega_e$ and $\log\left(\frac{D_e}{R_e^3 Z_1 Z_2}\right)$, the basis functions are set to be constant.

When dealing with dipole moment in Chapter 4, the kernel function with the best performance is the rational quadratic kernel [97] defined by

$$k(x_i, x_j|\theta) = \sigma_f^2 \left(1 + \frac{r^2}{2\alpha\sigma_l^2}\right)^{-a}, \tag{36}$$

where $\sigma_l$ is the length scale, and $\alpha$ is a scale-mixture parameter, $r$ is the Euclidean distance between $x_i$ and $x_j$ defined as

$$r = \sqrt{(x_i - x_j)^T (x_i - x_j)}. \tag{37}$$

## Appendix to Chapter 4: The dataset for dipole moments of diatomic molecules

The dipole moment of the diatomics dataset is summarized in Table 12, which consists of dipole moments $d$ of 162 polar diatomic molecules, 156 of which have information about equilibrium bond length $R_e$ while 139 also have harmonic vibrational frequency $\omega_e$. The references to the dipole moments are also listed in the table.

Table 12: The dipole moments, $d$, equilibrium bond length, $R_e$, and harmonic vibrational frequency, $\omega_e$, employed in this work. The references to the dipole moments are also listed in the table. $R_e$ and $\omega_e$ are taken from Ref. [17], [285] or the same reference of the dipole moment of the corresponding molecules, except when indicated.

| Molecule | $d$ (D) | $R_e$ (Å) | $\omega_e$ (cm$^{-1}$) | Ref. |
|---|---|---|---|---|
| AgBr | 5.62 | 2.393 | 247.7 | [135] |
| AgCl | 6.08 | 2.281 | 343.5 | [135] |
| AgF | 6.22 | 1.983 | 513.5 | [286] |
| AgH | 2.86 | 1.618 | 1759.9 | [287] |
| AgI | 4.55 | 2.545 | 206.5 | [135] |
| AlF | 1.515 | 1.654 | 802.3 | [126] |
| AuF | 4.32 | 1.918 | 539.4 [1] | [288] |
| AuO | 2.94 | 1.849 | 624.59 [2] | [289] |
| AuS | 2.22 | 2.156 | 410.19 [3] | [289] |
| BaF | 3.17 | 2.163 | 468.9 | [290] |
| BaO | 7.955 | 1.94 | 669.8 | [291] |
| BaS | 10.86 | 2.507 | 379.4 | [292] |
| BF | 0.5 | 1.263 | 1402.1 | [293] |
| BH | 1.27 | 1.232 | 2366.9 | [294] |
| BrCl | 0.519 | 2.136 | 444.3 | [135] |
| BrF | 1.422 | 1.759 | 670.8 | [135] |
| BrO | 1.76 | 1.717 | 778.7 | [295] |
| CaBr | 4.36 | 2.594 | 285.3 [4] | [296] |
| CaCl | 4.257 | 2.439 | 367.5 | [296] |
| CaD | 2.51 | 2.01 |  | [297] |
| CaF | 3.07 | 1.967 | 581.1 | [298] |

1 From Ref.[299]. 2 From Ref. [300]. 3 From Ref. [301]. 4 From Ref. [302].

Table 12 Continued.

| Molecule | $d$ (D) | $R_e$ (Å) | $\omega_e$ (cm$^{-1}$) | Ref. |
|---|---|---|---|---|
| CaH | 2.53 | 2.003 | 1298.3 | [297] |
| CaI | 4.5968 | 2.829 | 238.7 | [303] |
| CF | 0.65 | 1.272 | 1308.1 | [304] |
| CH | 1.46 | 1.12 | 2858.5 | [135] |
| ClD | 1.1033 | 1.275 | 2145.2 | [305] |
| ClF | 0.85 | 1.628 | 786.2 | [306] |
| ClH | 1.1085 | 1.275 | 2990.9 | [305] |
| ClO | 1.239 | 1.57 | 853.8 | [307] |
| CN | 1.45 | 1.172 | 2068.6 | [308] |
| CO | 0.112 | 1.128 | 2169.8 | [309] |
| CoF | 2.82 | | | [310] |
| CoH | 1.88 | | | [310] |
| CoO | 4.18 | 1.621 | | [311] |
| CrD | 3.51 | 1.663 | 1182 | [312] |
| CrN | 2.31 | 1.5652 [5] | 854.0 [6] | [313] |
| CrO | 3.88 | 1.615 | 898.4 | [314] |
| CS | 1.958 | 1.535 | 1285.1 | [315] |
| CsBr | 10.82 | 3.072 | 149.7 | [316] |
| CsCl | 10.387 | 2.906 | 214.2 | [317] |
| CSe | 1.99 | 1.676 | 1035.4 | [318] |
| CsF | 7.8839 | 2.345 | 352.6 | [317] |
| CsI | 11.69 | 3.315 | 119.2 | [316] |
| CuF | 5.26 | 1.745 | 622.7 | [319] |
| CuO | 4.57 | 1.724 | 640.2 | [320] |
| CuS | 4.31 | 2.051 | 415 | [321] |
| DBr | 0.823 | 1.415 | 1884.8 | [295] |
| DF | 1.819 | 0.917 | 2998.2 | [295] |
| FeC | 2.36 | 1.61 | | [322] |
| FeH | 2.63 | | | [323] |
| FeO | 4.7 | 1.6 | 970 | [324] |
| GaBr | 2.45 | 2.352 | 263 | [146] |
| GaF | 2.4 | 1.774 | 622.2 | [295] |
| GeO | 3.2824 | 1.625 | 985.5 | [325] |
| GeS | 2 | 2.012 | 575.8 | [146] |
| GeSe | 1.648 | 2.135 | 408.7 | [326] |
| GeTe | 1.06 | 2.34 | 323.9 | [326] |
| HBr | 0.8272 | 1.414 | 2649 | [135] |
| HF | 1.826526 | 0.917 | 4138.3 | [327] |
| HfF | 1.66 | 1.85 | | [328] |
| HfO | 3.431 | 1.723 | 974.1 | [329] |
| HI | 0.448 | 1.609 | 2309 | [135] |

[5] From Ref. [330]. [6] From Ref. [331].

Table 12 Continued.

| Molecule | $d$ (D) | $R_e$ (Å) | $\omega_e$ (cm$^{-1}$) | Ref. |
|---|---|---|---|---|
| IBr | 0.726 | 2.469 | 268.6 | [135] |
| ICl | 1.207 | 2.321 | 384.3 | [332] |
| ID | 0.316 | 1.609 | 1639.7 | [309] |
| IF | 1.948 | 1.91 | 610.2 | [135] |
| InCl | 3.79 | 2.401 | 317.4 | [135] |
| InF | 3.4 | 1.985 | 535.4 | [333] |
| IO | 2.45 | 1.868 | 681.5 | [304] |
| IrC | 1.6 | 1.683 | 1060.1 | [334] |
| IrF | 2.82 | 1.851 |  | [335] |
| IrN | 1.67 | 1.609 |  | [334] |
| KBr | 10.6281 | 2.821 | 213 | [336] |
| KCl | 10.2688 | 2.667 | 281 | [317] |
| KF | 8.59255 | 2.171 | 428 | [336] |
| KI | 10.82 | 3.048 | 186.5 | [316] |
| LaO | 3.207 | 1.826 | 812.8 | [329] |
| LiBr | 7.2262 | 2.17 | 563.2 | [337] |
| LiCl | 7.1289 | 2.021 | 643.3 | [317] |
| LiF | 6.32736 | 1.564 | 910.3 | [317] |
| LiH | 5.882 | 1.596 | 1405.7 | [338] |
| LiI | 7.4285 | 2.392 | 498.2 | [339] |
| LiK | 3.45 | 3.27 | 207 | [135] |
| LiNa | 0.47 | 2.81 | 256.8 | [340] |
| LiO | 6.84 | 1.695 | 851.5 | [135] |
| LiRb | 4 | 3.466 | 195.2 | [135] |
| MgD | 1.318 | 1.73 | 1077.9 | [341] |
| MgO | 6.2 | 1.749 | 785.1 | [135] |
| MoC | 6.07 |  |  | [342] |
| MoN | 3.38 | 1.63 |  | [343] |
| NaBr | 9.1183 | 2.502 | 302.1 | [317] |
| NaCl | 9.002 | 2.361 | 366 | [317] |
| NaCs | 4.7 | 3.851 | 98.9 | [295] |
| NaF | 8.1558 | 1.926 | 536 | [344] |
| NaH | 6.4 | 1.889 | 1176 | [345] |
| NaI | 9.2357 | 2.711 | 258 | [317] |
| NaK | 2.693 | 3.589 | 124.1 | [135] |
| NaRb | 3.1 | 3.644 | 106.9 | [135] |
| NbN | 3.26 | 1.663 |  | [346] |
| NH | 1.39 | 1.036 | 3282.3 | [135] |
| NiH | 2.4 | 1.476 | 1926.6 | [347] |
| NO | 0.157 | 1.151 | 1904.2 | [348] |
| NS | 1.86 | 1.494 | 1218.7 | [304] |

Table 12 Continued.

| Molecule | $d$ (D) | $R_e$ (Å) | $\omega_e$ (cm$^{-1}$) | Ref. |
|---|---|---|---|---|
| OD | 1.653 | 0.97 | 2720.2 | [295] |
| OF | 0.0043 | 1.354 | 1028.7 | [135] |
| OH | 1.6498 | 0.97 | 3737.8 | [349] |
| PbO | 4.64 | 1.922 | 721 | [350] |
| PbS | 3.59 | 2.287 | 429.4 | [350] |
| PbSe | 3.29 | 2.402 | 277.6 | [326] |
| PbTe | 2.73 | 2.595 | 212 | [326] |
| PN | 2.7514 | 1.491 | 1337.2 | [351] |
| PO | 1.88 | 1.476 | 1233.3 | [352] |
| PtC | 0.99 | 1.677 | 1051.1 | [353] |
| PtF | 3.42 | 1.868 | | [354] |
| PtN | 1.977 | 1.682 | | [355] |
| PtO | 2.77 | 1.727 | 851.1 | [353] |
| PtS | 1.78 | 2.042 | | [353] |
| RbBr | 10.86 | 2.945 | 169.5 | [316] |
| RbCl | 10.51 | 2.787 | 228 | [317] |
| RbF | 8.5465 | 2.27 | 376 | [317] |
| RbI | 11.48 | 3.177 | 138.5 | [316] |
| ReN | 1.96 | 0.61 | | [356] |
| RhN | 2.43 | 1.64 | | [357] |
| RhO | 3.81 | 1.739 | | [358] |
| RuF | 5.34 | 1.916 | | [359] |
| ScO | 4.55 | 1.668 | 965 | [360] |
| ScS | 5.64 | 2.139 | 565.2 | [361] |
| SD | 0.7571 | 1.341 | 1885.5 | [362] |
| SeD | 0.48 | 1.47 | 1708 | [295] |
| SeF | 1.52 | 1.741 | 757 | [304] |
| SeH | 0.5 | 1.47 | 2400 | [295] |
| SF | 0.87 | 1.596 | 837.6 | [304] |
| SH | 0.758 | 1.341 | 2711.6 | [363] |
| SiH | 5.9 | 1.52 | 2041.8 | [295] |
| SiO | 3.0982 | 1.51 | 1241.6 | [325] |
| SiS | 1.73 | 1.73 | 749.6 | [364] |
| SiSe | 1.1 | 2.058 | 580 | [146] |
| SnO | 4.32 | 1.833 | 814.6 | [146] |
| SnS | 3.18 | 2.209 | 487.3 | [146] |
| SnSe | 2.82 | 2.326 | 331.2 | [146] |
| SnTe | 2.19 | 2.523 | 259 | [146] |
| SO | 1.55 | 1.481 | 1149.2 | [365] |
| SrF | 3.4676 | 2.075 | 502.4 | [366] |

Table 12 Continued.

| Molecule | $d$ (D) | $R_e$ (Å) | $\omega_e$ (cm$^{-1}$) | Ref. |
|---|---|---|---|---|
| SrO | 8.9 | 1.92 | 653.5 | [295] |
| ThO | 3.534 | 1.84 | 895.8 | [367] |
| ThS | 4.58 | 2.35 | 477 [7] | [368] |
| TiH | 2.455 | | | [369] |
| TiN | 3.56 | 1.582[8] | 1039[9] | [370] |
| TiO | 3.34 | 1.62 | 1009 | [371] |
| TlBr | 4.49 | 2.618 | 192.1 | [295] |
| TlCl | 4.5429 | 2.485 | 283.8 | [146] |
| TlF | 4.2282 | 2.084 | 477.3 | [372] |
| TlI | 4.61 | 2.814 | 143 | [135] |
| VN | 3.07 | 1.566[10] | 1033[11] | [313] |
| VO | 3.355 | 1.592[12] | 1011.3 | [373] |
| VS | 5.16 | 2.06 | | [374] |
| WC | 3.9 | | | [375] |
| WN | 3.77 | 1.67[13] | | [356] |
| YbF | 3.91 | 2.016 | 501.9 | [376] |
| YF | 1.82 | 1.926 | 631.3 | [377] |
| YO | 4.524 | 1.79 | 861 | [329] |
| ZrO | 2.551 | 1.712 | 969.8 | [329] |

[7] From Ref. [378]. [8] From Ref. [379]. [9] From Ref. [380]. [10] From Ref. [381].
[11] From Ref. [382]. [12] From Ref. [383]. [13] From Ref. [384].

## Appendix to Chapter 5: Experimental versus CCSD(T) calculated equilibrium bond length $R_e$, harmonic vibrational frequency $\omega_e$ and electric dipole moment $\mu$

The experimental and calculated equilibrium bond length $R_e$ and harmonic vibrational frequency $\omega_e$, employing different basis sets, are listed in Table 13 and Table 14, respectively.

Table 15 displays the experimental electric dipole moments employed in this work, including the pertinent references. Similarly, it includes the calculated dipole moments employing different basis sets.

Table 13: The experimental and calculated equilibrium bond length $R_e$ (in Å). The experimental values are taken from Refs. [17] and [285], or the same reference of the experimental dipole moment of the corresponding molecule.

| Molecule | State | $R_e$(Exp.) | $R_e$ (aug-cc-pwCVTZ) | $R_e$ (aug-cc-pwCVQZ) | $R_e$ (def2-QZVPP) |
|---|---|---|---|---|---|
| AgBr | X $^1\Sigma^+$ | 2.393 | 2.394 | 2.387 | 2.397 |
| AgF | X $^1\Sigma^+$ | 1.983 | 1.982 | 1.979 | 1.982 |
| AgI | X $^1\Sigma^+$ | 2.545 | 2.542 | 2.536 | 2.541 |
| AlF | X $^1\Sigma^+$ | 1.654 | 1.660 | 1.656 | 1.660 |
| BrO | X $^2\Pi_{3/2}$ | 1.717 | 1.728 | 1.726 | 1.714 |
| CF | X $^2\Pi$ | 1.272 | 1.277 | 1.273 | 1.272 |
| CN | X $^2\Sigma^+$ | 1.172 | 1.170 | 1.167 | 1.168 |
| CO | X $^1\Sigma^+$ | 1.128 | 1.132 | 1.129 | 1.129 |
| CS | X $^1\Sigma^+$ | 1.535 | 1.539 | 1.536 | 1.537 |
| CSe | X $^1\Sigma^+$ | 1.676 | 1.678 | 1.674 | 1.678 |
| CuF | X $^1\Sigma^+$ | 1.745 | 1.773 | 1.769 | 1.756 |
| GaF | X $^1\Sigma^+$ | 1.774 | 1.778 | 1.775 | 1.769 |
| GeO | X $^1\Sigma^+$ | 1.625 | 1.627 | 1.623 | 1.626 |
| GeS | X $^1\Sigma^+$ | 2.012 | 2.019 | 2.012 | 2.013 |
| GeTe | X $^1\Sigma^+$ | 2.340 | 2.343 | 2.335 | 2.332 |
| HfO | X $^1\Sigma^+$ | 1.723 | 1.719 | 1.715 | 1.735 |
| IBr | X $^1\Sigma^+$ | 2.469 | 2.478 | 2.464 | 2.463 |
| InCl | X $^1\Sigma^+$ | 2.401 | 2.413 | 2.404 | 2.392 |
| InF | X $^1\Sigma^+$ | 1.985 | 1.988 | 1.984 | 1.974 |
| NO | X $^2\Pi_{1/2}$ | 1.151 | 1.152 | 1.149 | 1.148 |
| PN | X $^1\Sigma^+$ | 1.491 | 1.496 | 1.491 | 1.493 |
| PO | X $^2\Pi$ | 1.476 | 1.482 | 1.477 | 1.479 |
| PbO | X $^1\Sigma^+$ | 1.922 | 1.924 | 1.919 | 1.922 |
| PbS | X $^1\Sigma^+$ | 2.287 | 2.297 | 2.288 | 2.286 |
| SO | X $^3\Sigma^-$ | 1.481 | 1.486 | 1.481 | 1.482 |
| ScO | X $^2\Sigma^+$ | 1.668 | 1.674 | 1.663 | 1.670 |
| SiO | X $^1\Sigma^+$ | 1.510 | 1.515 | 1.511 | 1.514 |
| SiS | X $^1\Sigma^+$ | 1.929 | 1.939 | 1.931 | 1.933 |
| SnO | X $^1\Sigma^+$ | 1.833 | 1.832 | 1.827 | 1.826 |
| SnS | X $^1\Sigma^+$ | 2.209 | 2.215 | 2.206 | 2.204 |
| YF | X $^1\Sigma^+$ | 1.926 | 1.935 | 1.930 | 1.936 |
| ZrO | X $^1\Sigma^+$ | 1.712 | 1.719 | 1.714 | 1.715 |

Table 14: The experimental and calculated harmonic vibrational frequency $\omega_e$ (in cm$^{-1}$). The experimental values are taken from Refs. [17] and [285], or the same reference of the experimental dipole moment of the corresponding molecule.

| Molecule | State | $\omega_e$ (Exp.) | $\omega_e$ (aug-cc-pwCVTZ) | $\omega_e$ (aug-cc-pwCVQZ) | $\omega_e$ [CBS(aug-cc-pwCVT/QZ)] | $\omega_e$ (def2-QZVPP) |
|---|---|---|---|---|---|---|
| AgBr | $X\,^1\Sigma^+$ | 247.7 | 246.4 | 248.6 | 250.2 | 246.9 |
| AgF | $X\,^1\Sigma^+$ | 513.5 | 512.0 | 515.7 | 518.1 | 515.0 |
| AgI | $X\,^1\Sigma^+$ | 206.5 | 208.9 | 209.7 | 210.2 | 210.1 |
| AlF | $X\,^1\Sigma^+$ | 802.3 | 793.9 | 801.5 | 804.9 | 804.1 |
| BrO | $X\,^2\Pi_{3/2}$ | 778.7 | 767.5 | 762.8 | 631.2 | 743.0 |
| CF | $X\,^2\Pi$ | 1308.1 | 1304.5 | 1299.7 | 1269.4 | 1307.2 |
| CN | $X\,^2\Sigma^+$ | 2068.6 | 2160.4 | 2175.5 | 2181.2 | 2108.1 |
| CO | $X\,^1\Sigma^+$ | 2169.8 | 2153.4 | 2170.3 | 2177.5 | 2174.4 |
| CS | $X\,^1\Sigma^+$ | 1285.1 | 1279.1 | 1288.5 | 1293.7 | 1292.9 |
| CSe | $X\,^1\Sigma^+$ | 1035.4 | 1033.1 | 1041.4 | 1045.8 | 1052.4 |
| CuF | $X\,^1\Sigma^+$ | 622.7 | 595.3 | 599.7 | 603.3 | 605.6 |
| GaF | $X\,^1\Sigma^+$ | 622.2 | 615.4 | 621.0 | 623.7 | 634.7 |
| GeO | $X\,^1\Sigma^+$ | 985.5 | 979.8 | 989.0 | 994.9 | 997.7 |
| GeS | $X\,^1\Sigma^+$ | 575.8 | 571.1 | 577.1 | 581.3 | 584.2 |
| GeTe | $X\,^1\Sigma^+$ | 323.9 | 322.7 | 326.1 | 328.7 | 331.7 |
| HfO | $X\,^1\Sigma^+$ | 974.1 | 970.1 | 976.0 | 980.7 | 952.5 |
| IBr | $X\,^1\Sigma^+$ | 268.6 | 269.2 | 276.1 | 279.5 | 279.7 |
| InCl | $X\,^1\Sigma^+$ | 317.4 | 313.6 | 316.4 | 318.6 | 319.2 |
| InF | $X\,^1\Sigma^+$ | 535.4 | 530.1 | 534.5 | 536.3 | 549.9 |
| NO | $X\,^2\Pi_{1/2}$ | 1904.2 | 1883.7 | 1912.5 | 1928.7 | 1927.6 |
| PN | $X\,^1\Sigma^+$ | 1337.2 | 1332.5 | 1345.6 | 1352.8 | 1346.4 |
| PO | $X\,^2\Pi$ | 1233.3 | 1232.0 | 1242.9 | 1250.8 | 1243.7 |
| PbO | $X\,^1\Sigma^+$ | 721.0 | 732.8 | 739.9 | 745.1 | 746.1 |
| PbS | $X\,^1\Sigma^+$ | 429.4 | 433.2 | 439.9 | 444.6 | 445.3 |
| SO | $X\,^3\Sigma^-$ | 1149.2 | 1152.9 | 1161.3 | 1167.7 | 1162.7 |
| ScO | $X\,^2\Sigma^+$ | 965.0 | 968.6 | 979.6 | 936.3 | 932.9 |
| SiO | $X\,^1\Sigma^+$ | 1241.6 | 1231.6 | 1242.3 | 1248.6 | 1241.0 |
| SiS | $X\,^1\Sigma^+$ | 749.6 | 741.1 | 752.6 | 759.2 | 753.1 |
| SnO | $X\,^1\Sigma^+$ | 814.6 | 821.6 | 829.9 | 835.6 | 842.3 |
| SnS | $X\,^1\Sigma^+$ | 487.3 | 484.2 | 491.4 | 496.5 | 492.9 |
| YF | $X\,^1\Sigma^+$ | 631.3 | 628.9 | 634.4 | 637.7 | 629.2 |
| ZrO | $X\,^1\Sigma^+$ | 969.8 | 977.9 | 981.4 | 981.6 | 979.8 |

Table 15: The experimental and calculated dipole moments in Debye.

| Mol. | State | Exp. | $\mu_e$ (cc-pwCVTZ) | $\mu_0$ (cc-pwCVTZ) | $\mu_e$ (cc-pwCVQZ) | $\mu_0$ (cc-pwCVQZ) | $\mu_e$ (aug-cc-pwCVTZ) | $\mu_0$ (aug-cc-pwCVTZ) | $\mu_e$ (aug-cc-pwCVQZ) | $\mu_0$ (aug-cc-pwCVQZ) | $\mu_e$ [CBS (aug-cc-pwCVT/QZ)] | $\mu_0$ [CBS (aug-cc-pwCVT/QZ)] | $\mu_e$ (def2-QZVPP) | $\mu_0$ (def2-QZVPP) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AgBr | X $^1\Sigma^+$ | 5.62(3) [385] | 5.574 | 5.594 | | | 5.473 | 5.493 | 5.509 | 5.530 | 5.548 | 5.569 | 5.631 | 5.652 |
| AgF | X $^1\Sigma^+$ | 6.22 (20) [286] | 5.767 | 5.803 | 5.877 | 5.916 | 5.935 | 5.977 | 5.922 | 5.964 | 5.927 | 5.969 | 5.898 | 5.940 |
| AgI | X $^1\Sigma^+$ | 4.55(5) [203] | 5.161 | 5.178 | 5.191 | 5.208 | 5.015 | 5.030 | 5.082 | 5.098 | 5.145 | 5.161 | 5.144 | 5.160 |
| AlF | X $^1\Sigma^+$ | 1.515 (4) [30] | 1.343 | 1.401 | 1.423 | 1.484 | 1.496 | 1.560 | 1.476 | 1.540 | 1.471 | 1.535 | 1.478 | 1.540 |
| BrO | X $^2\Pi_{3/2}$ | 1.76(4) [304] | 1.690 | 1.673 | 1.710 | 1.693 | 1.747 | 1.734 | 1.729 | 1.737 | 1.717 | 1.728 | 1.727 | 1.709 |
| CF | X $^2\Pi$ | 0.65(5) [304] | 0.700 | 0.635 | 0.687 | 0.619 | 0.662 | 0.591 | 0.680 | 0.611 | 0.685 | 0.613 | 0.694 | 0.625 |
| CN | X $^2\Sigma^+$ | 1.45(8) [308] | 1.362 | 1.335 | 1.415 | 1.390 | 1.413 | 1.387 | 1.431 | 1.407 | 1.444 | 1.421 | 1.426 | 1.407 |
| CO | X $^1\Sigma^+$ | 0.112 (5) [309] | 0.143 | 0.121 | 0.127 | 0.104 | 0.114 | 0.090 | 0.119 | 0.094 | 0.129 | 0.104 | 0.123 | 0.144 |
| CS | X $^1\Sigma^+$ | 1.958 (5) [315] | | | | | 1.983 | 1.954 | 1.972 | 1.943 | 1.961 | 1.932 | 1.960 | 1.932 |

Table 16: Table 15 continued: The experimental and calculated dipole moments in Debye.

| Mol. | State | Exp. | $\mu_e$ (cc-pwCVTZ) | $\mu_0$ (cc-pwCVTZ) | $\mu_e$ (cc-pwCVQZ) | $\mu_0$ (cc-pwCVQZ) | $\mu_e$ (aug-cc-pwCVTZ) | $\mu_0$ (aug-cc-pwCVTZ) | $\mu_e$ (aug-cc-pwCVQZ) | $\mu_0$ (aug-cc-pwCVQZ) | $\mu_e$ [CBS (aug-cc-pwCVT/QZ)] | $\mu_0$ [CBS (aug-cc-pwCVT/QZ)] | $\mu_e$ (def2-QZVPP) | $\mu_0$ (def2-QZVPP) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CSe | X $^1\Sigma^+$ | 1.99(4) [318] | 2.146 | 2.123 | 2.165 | 2.142 | 2.186 | 2.163 | 2.171 | 2.147 | 2.155 | 2.131 | 2.140 | 2.116 |
| CuF | X $^1\Sigma^+$ | 5.26(2) [200] | 5.144 | 5.180 | 5.256 | 5.294 | 5.561 | 5.603 | 5.524 | 5.566 | 5.513 | 5.555 | 5.446 | 5.488 |
| GaF | X $^1\Sigma^+$ | 2.45(5) [333] | 2.176 | 2.235 | 2.318 | 2.381 | 2.431 | 2.498 | 2.417 | 2.483 | 2.416 | 2.482 | 2.230 | 2.295 |
| GeO | X $^1\Sigma^+$ | 3.2824 (1) [325] | 3.012 | 3.028 | 3.202 | 3.220 | 3.295 | 3.315 | 3.303 | 3.323 | 3.314 | 3.334 | 3.198 | 3.217 |
| GeS | X $^1\Sigma^+$ | 2.00(6) [350] | 1.952 | 1.969 | 2.056 | 2.074 | 2.057 | 2.076 | 2.084 | 2.103 | 2.114 | 2.133 | 1.981 | 1.999 |
| GeTe | X $^1\Sigma^+$ | 1.06(7) [326] | 1.063 | 1.073 | 1.140 | 1.151 | 1.086 | 1.096 | 1.146 | 1.157 | 1.192 | 1.204 | 0.995 | 1.006 |
| HfO | X $^1\Sigma^+$ | 3.431 (5) [329] | 3.258 | 3.280 | 3.330 | 3.352 | 3.396 | 3.420 | 3.381 | 3.405 | 3.376 | 3.400 | 3.473 | 3.496 |
| IBr | X $^1\Sigma^+$ | 0.726 (3) [135] | 0.687 | 0.811 |  |  | 0.647 | 0.651 | 0.630 | 0.634 | 0.624 | 0.628 | 0.685 | 0.688 |
| InCl | X $^1\Sigma^+$ | 3.79 (19) [135] | 3.465 | 3.518 | 3.581 | 3.637 | 3.672 | 3.728 | 3.650 | 3.707 | 3.646 | 3.703 | 3.497 | 3.555 |

Table 17: Table 15 continued: The experimental and calculated dipole moments in Debye.

| Mol. | State | Exp. | $\mu_e$ (cc-pwCVTZ) | $\mu_0$ (cc-pwCVTZ) | $\mu_e$ (cc-pwCVQZ) | $\mu_0$ (cc-pwCVQZ) | $\mu_e$ (aug-cc-pwCVTZ) | $\mu_0$ (aug-cc-pwCVTZ) | $\mu_e$ (aug-cc-pwCVQZ) | $\mu_0$ (aug-cc-pwCVQZ) | $\mu_e$ [CBS (aug-cc-pwCVT/QZ)] | $\mu_0$ [CBS (aug-cc-pwCVT/QZ)] | $\mu_e$ (def2-QZVPP) | $\mu_0$ (def2-QZVPP) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| InF | X $^1\Sigma^+$ | 3.40(7) [333] | 3.066 | 3.124 | 3.227 | 3.288 | 3.378 | 3.445 | 3.358 | 3.425 | 3.358 | 3.424 | 3.219 | 3.285 |
| NO | X $^2\Pi_{1/2}$ | 0.1595 (15) [386] | 0.126 | 0.115 | 0.135 | 0.121 | 0.144 | 0.132 | 0.149 | 0.134 | 0.156 | 0.139 | 0.143 | 0.128 |
| PN | X $^1\Sigma^+$ | 2.7514 (6) [351] | 2.634 | 2.626 | 2.722 | 2.715 | 2.758 | 2.751 | 2.771 | 2.764 | 2.780 | 2.773 | 2.757 | 2.750 |
| PO | X $^2\Pi$ | 1.88(7) [352] | 1.946 | 1.959 | 1.966 | 1.983 | 1.962 | 1.978 | 1.958 | 1.976 | 1.959 | 1.979 | 1.984 | 2.001 |
| PbO | X $^1\Sigma^+$ | 4.64 (30) [350] | 3.990 | 4.004 | 4.289 | 4.305 | 4.470 | 4.489 | 4.467 | 4.486 | 4.471 | 4.490 | 4.373 | 4.389 |
| PbS | X $^1\Sigma^+$ | 3.59 (10) [350] | 3.614 | 3.629 | 3.780 | 3.796 | 3.833 | 3.849 | 3.849 | 3.866 | 3.872 | 3.889 | 3.792 | 3.809 |
| SO | X $^3\Sigma^-$ | 1.55(2) [365] | 1.566 | 1.566 | 1.564 | 1.568 | 1.574 | 1.581 | 1.559 | 1.566 | 1.552 | 1.559 | 1.578 | 1.584 |
| ScO | X $^2\Sigma^+$ | 4.55(8) [387] | 3.419 | 3.567 | 3.727 | 3.685 | 3.793 | 3.779 | 3.721 | 3.748 | 3.713 | 3.721 | 3.788 | 3.809 |

Table 18: Table 15 continued: The experimental and calculated dipole moments in Debye.

| Mol. | State | Exp. | $\mu_e$ (cc-pwCVTZ) | $\mu_0$ (cc-pwCVTZ) | $\mu_e$ (cc-pwCVQZ) | $\mu_0$ (cc-pwCVQZ) | $\mu_e$ (aug-cc-pwCVTZ) | $\mu_0$ (aug-cc-pwCVTZ) | $\mu_e$ (aug-cc-pwCVQZ) | $\mu_0$ (aug-cc-pwCVQZ) | $\mu_e$ [CBS (aug-cc-pwCVT/QZ)] | $\mu_0$ [CBS (aug-cc-pwCVT/QZ)] | $\mu_e$ (def2-QZVPP) | $\mu_0$ (def2-QZVPP) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SiO | $X\ ^1\Sigma^+$ | 3.0982 (10) [325] | 2.871 | 2.886 | 3.025 | 3.041 | 3.099 | 3.118 | 3.106 | 3.125 | 3.114 | 3.133 | 3.082 | 3.100 |
| SiS | $X\ ^1\Sigma^+$ | 1.73(6) [388] | 1.608 | 1.626 | 1.681 | 1.701 | 1.683 | 1.703 | 1.699 | 1.720 | 1.723 | 1.744 | 1.695 | 1.716 |
| SnO | $X\ ^1\Sigma^+$ | 4.32(10) [350] | 3.675 | 3.688 | 3.939 | 3.955 | 4.079 | 4.097 | 4.085 | 4.103 | 4.098 | 4.116 | 3.995 | 4.011 |
| SnS | $X\ ^1\Sigma^+$ | 3.18(10) [350] | 2.974 | 2.990 | 3.125 | 3.142 | 3.153 | 3.170 | 3.180 | 3.197 | 3.210 | 3.227 | 3.125 | 3.143 |
| YF | $X\ ^1\Sigma^+$ | 1.82(8) [377] | 1.711 | 1.749 | 1.800 | 1.840 | 1.848 | 1.890 | 1.848 | 1.890 | 1.853 | 1.895 | 1.830 | 1.871 |
| ZrO | $X\ ^1\Sigma^+$ | 2.551(11) [329] | 2.370 | 2.395 | 2.445 | 2.471 | 2.521 | 2.549 | 2.502 | 2.530 | 2.500 | 2.530 | 2.478 | 2.505 |

## Appendix to Chapter 7: Chemistry of AlF and CaF production in buffer gas sources

The efficiency of the AlF, CaF and MgF molecule formation in Al/Ca/Mg + $SF_6$/$NF_3$ collisions are presented in Fig.66, shown as the reaction probabilities for different products as a function of temperature. In particular, the productivity of MgF is simulated at 5000 K and 15000 K. The productivity of by-products is shown in Fig. 67.

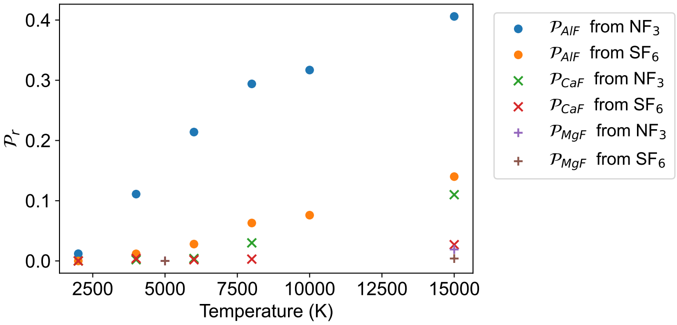


Figure 66: Reaction probability of $AlF_n$, $CaF_n$ and $MgF_n$ by-products for hot collisions of Al/Ca/Mg with $SF_6$ and $NF_3$ gases as a function of the temperature. Figure reproduced from ref. [228].

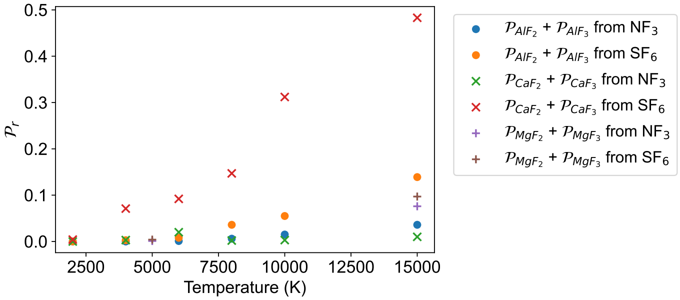


Figure 67: Reaction probability of (a) AlF, CaF and MgF and (b) $AlF_n$, $CaF_n$ and $MgF_n$ by-products for hot collisions of Al/Ca/Mg with $SF_6$ and $NF_3$ gases as a function of the temperature. Figure reproduced from ref. [228].

# BIBLIOGRAPHY

[1] Xuefei Xu, Wenjing Zhang, Mingsheng Tang, and Donald G. Truhlar. Do practical standard coupled cluster calculations agree better than Kohn–Sham calculations with currently available functionals when compared to the best available experimental data for dissociation energies of bonds to 3d transition metals? *Journal of chemical theory and computation*, 11(5):2036–2052, 2015.

[2] Wenjing Zhang, Donald G. Truhlar, and Mingsheng Tang. Tests of exchange-correlation functional approximations against reliable experimental data for average bond energies of 3d transition metal compounds. *Journal of chemical theory and computation*, 9(9):3965–3977, 2013.

[3] Wanyi Jiang, Nathan J. DeYonker, John J. Determan, and Angela K. Wilson. Toward accurate theoretical thermochemistry of first row transition metal complexes. *The Journal of Physical Chemistry A*, 116(2):870–885, 2012.

[4] Lan Cheng, Jurgen Gauss, Branko Ruscic, Peter B. Armentrout, and John F. Stanton. Bond dissociation energies for diatomic molecules containing 3d transition metals: Benchmark scalar-relativistic coupled-cluster calculations for 20 molecules. *Journal of chemical theory and computation*, 13(3):1044–1056, 2017.

[5] David E. Woon and Thom H. Dunning Jr. Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon. *The Journal of Chemical Physics*, 98(2):1358–1371, 1993.

[6] Thom H. Dunning Jr. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *The Journal of Chemical Physics*, 90(2):1007–1023, 1989.

[7] Angela K. Wilson, David E. Woon, Kirk A. Peterson, and Thom H. Dunning Jr. Gaussian basis sets for use in correlated molecular calculations. IX. The atoms gallium through krypton. *The Journal of Chemical Physics*, 110(16):7667–7676, 1999.

[8] Frank Jensen. The basis set convergence of the Hartree–Fock energy for $H_3^+$, $Li_2$ and $N_2$. *Theoretical Chemistry Accounts*, 104:484–490, 2000.

[9] Diptarka Hait, Yu Hsuan Liang, and Martin Head-Gordon. Too big, too small, or just right? A benchmark assessment of density functional theory for predicting the spatial extent of the electron density of small chemical systems. *The Journal of Chemical Physics*, 154(7):074109, 2021.

[10] Asger Halkier, Helena Larsen, Jeppe Olsen, Poul Jo/rgensen, and Jürgen Gauss. Full configuration interaction benchmark calculations of first-order one-electron properties of BH and HF. *The Journal of Chemical Physics*, 110(2):734–740, 1999.

[11] Alexander I. Johnson, Kushantha P. K. Withanage, Kamal Sharkas, Yoh Yamamoto, Tunna Baruah, Rajendra R. Zope, Juan E. Peralta, and Koblar A. Jackson. The effect of self-interaction error on electrostatic dipoles calculated using density functional theory. *The Journal of Chemical Physics*, 151(17):174106, 2019.

[12] Robin Grotjahn, Gregor J Lauter, Matthias Haasler, and Martin Kaupp. Evaluation of local hybrid functionals for electric properties: Dipole moments and static and dynamic polarizabilities. *The Journal of Physical Chemistry A*, 124(40):8346–8358, 2020.

[13] Ricardo A. Mata and Martin A. Suhm. Benchmarking quantum chemical methods: Are we heading in the right direction? *Angewandte Chemie International Edition*, 56(37):11011–11018, 2017.

[14] Zongtang Fang, Monica Vasiliu, Kirk A. Peterson, and David A. Dixon. Prediction of bond dissociation energies/heats of formation for diatomic transition metal compounds: CCSD(T) works. *Journal of chemical theory and computation*, 13(3):1057–1066, 2017.

[15] Yuri A. Aoto, Ana Paula de Lima Batista, Andreas Kohn, and Antonio GS de Oliveira-Filho. How to arrive at accurate benchmark values for transition metal compounds: Computation or experiment? *Journal of chemical theory and computation*, 13(11):5291–5316, 2017.

[16] Laura K. McKemmish. Molecular diatomic spectroscopy data. *WIREs Computational Molecular Science*, 11(5):e1520, 2021.

[17] K. P. Huber and G. Gerzberg. *Molecular Spectra and Molecular Structure*. Springer, New York, 1979.

[18] Luke W Bertels, Joonho Lee, and Martin Head-Gordon. Polishing the gold standard: The role of orbital choice in CCSD(T) vibrational frequency prediction. *Journal of chemical theory and computation*, 17(2):742–755, 2021.

[19] Juan Camilo Zapata and Laura K. McKemmish. Computation of dipole moments: A recommendation on the choice of the basis set and the level of theory. *The Journal of Physical Chemistry A*, 124(37):7538–7548, 2020.

[20] X. Liu, S. Truppe, G. Meijer, and J. Pérez-Ríos. The diatomic molecular spectroscopy database. *J Cheminform*, 12:31, 2020.

[21] Xiangyue Liu, Gerard Meijer, and Jesús Pérez-Ríos. A data-driven approach to determine dipole moments of diatomic molecules. *Phys. Chem. Chem. Phys.*, 22:24191–24200, 2020.

[22] Cheng Chin, V. V. Flambaum, and M. G. Kozlov. Ultracold molecules: new probes on the variation of fundamental constants. *New Journal of Physics*, 11(5):055048, 2009.

[23] M. R. Tarbutt, B. E. Sauer, J. J. Hudson, and E. A. Hinds. Design for a fountain of YbF molecules to measure the electron's electric dipole moment. *New Journal of Physics*, 15(5):053034, 2013.

[24] Ivan Kozyryev, Timothy C. Steimle, Phelan Yu, Duc-Trung Nguyen, and John M. Doyle. Determination of CaOH and $CaOCH_3$ vibrational branching ratios for direct laser cooling and trapping. *New Journal of Physics*, 21(5):052002, 2019.

[25] J. F. Barry, D. J. McCarron, E. B. Norrgard, M. H. Steinecker, and D. DeMille. Magneto-optical trapping of a diatomic molecule. *Nature*, 512(7514):286–289, 2014.

[26] H. J. Williams, S. Truppe, M. Hambach, L. Caldwell, N. J. Fitch, E. A. Hinds, B. E. Sauer, and M. R. Tarbutt. Characteristics of a magneto-optical trap of molecules. *New Journal of Physics*, 19(11):113035, nov 2017.

[27] Liang Xu, Yanning Yin, Bin Wei, Yong Xia, and Jianping Yin. Calculation of vibrational branching ratios and hyperfine structure of $^{24}Mg^{19}F$ and its suitability for laser cooling and magneto-optical trapping. *Phys. Rev. A*, 93:013408, Jan 2016.

[28] Matthew T. Hummon, Mark Yeo, Benjamin K. Stuhl, Alejandra L. Collopy, Yong Xia, and Jun Ye. 2D Magneto-Optical Trapping of Diatomic Molecules. *Phys. Rev. Lett.*, 110:143001, Apr 2013.

[29] Nathan Wells and Ian C. Lane. Electronic states and spin-forbidden cooling transitions of AlH and AlF. *Physical Chemistry Chemical Physics*, 13(42):19018–19025, 2011.

[30] Stefan Truppe, Silvio Marx, Sebastian Kray, Maximilian Doppelbauer, Simon Hofsäss, Hanns Christian Schewe, Nicole Walter, Jesús Pérez-Ríos, Boris G. Sartakov, and Gerard Meijer. Spectroscopic characterization of aluminum monofluoride with relevance to laser cooling and trapping. *Physical Review A*, 100(5):052513, 2019.

[31] Taylor N. Lewis, Chen Wang, John R. Daniel, Madhav Dhital, Christopher J. Bardeen, and Boerge Hemmerling. Optimizing pulsed-laser ablation production of AlCl molecules for laser cooling. *Phys. Chem. Chem. Phys.*, 23:22785–22793, 2021.

[32] D. DeMille. Quantum computation with trapped polar molecules. *Phys. Rev. Lett.*, 88:067901, Jan 2002.

[33] David B. Blasing, Jesús Pérez-Ríos, Yangqian Yan, Sourav Dutta, Chuan-Hsun Li, Qi Zhou, and Yong P. Chen. Observation of quantum interference and coherent control in a photochemical reaction. *Phys. Rev. Lett.*, 121:073202, Aug 2018.

[34] Lincoln D. Carr, David DeMille, Roman V. Krems, and Jun Ye. Cold and ultracold molecules: science, technology and applications. *New Journal of Physics*, 11(5):055049, may 2009.

[35] Rouven Essig, Jesús Pérez-Ríos, Harikrishnan Ramani, and Oren Slone. Direct detection of spin-(in)dependent nuclear scattering of sub-gev dark matter using molecular excitations, 2019.

[36] M. S. Safronova, D. Budker, D. DeMille, Derek F. Jackson Kimball, A. Derevianko, and Charles W. Clark. Search for new physics with atoms and molecules. *Rev. Mod. Phys.*, 90:025008, Jun 2018.

[37] HITRANonline, 2019.

[38] ExoMol, 2019.

[39] Nist chemistry webbook, 2019.

[40] The open spectral database, 2019.

[41] Christian P. Endres, Stephan Schlemmer, Peter Schilke, Jürgen Stutzki, and Holger S.P. Müller. The cologne database for molecular spectroscopy, CDMS, in the virtual atomic and molecular data centre, VAMDC. *Journal of Molecular Spectroscopy*, 327:95–104, 2016. New Visions of Spectroscopic Databases, Volume II.

[42] P.F Bernath and S McLeod. DiRef, a database of references associated with the spectra of diatomic molecules. *Journal of Molecular Spectroscopy*, 207(2):287, 2001.

[43] Frank J. Lovas and Eberhard Tiemann. Microwave Spectral Tables I. Diatomic Molecules. *Journal of Physical and Chemical Reference Data*, 3(3):609–770, 10 2009.

[44] F. J. Lovas, E. Tiemann, J. S. Coursey, S. A. Kotochigova, J. Chang, K. Olsen, and R. A. Dragoset. Molecular microwave spectral databases, 2005.

[45] The cologne databse for molecular spectroscopy, 2019.

[46] Sesam: SpEctroScopy of Atoms and Molecules, 2019.

[47] Peter F. Bernath. *Spectra of atoms and molecules*. Oxford university press, 2020.

[48] J. V. Lill, G. A. Parker, and J. C. Light. Discrete variable representations and sudden models in quantum scattering theory. *Chemical Physics Letters*, 89(6):483–489, 1982.

[49] J. C. Light, I. P. Hamilton, and J. V. Lill. Generalized discrete variable approximation in quantum mechanics. *The Journal of Chemical Physics*, 82(3):1400–1409, 1985.

[50] John C. Light and Tucker Carrington Jr. Discrete-variable representations and their utilization. *Advances in Chemical Physics*, 114:263–310, 2000.

[51] Mike Bostock. d3.js, 2019.

[52] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D$^3$ data-driven documents. *IEEE transactions on visualization and computer graphics*, 17(12):2301–2309, 2011.

[53] Database of spectroscopic constants of diatomic molecules, 2023.

[54] G. Herzberg. Molecular spectroscopy: a personal history. *Annu. Rev. Phys. Chem.*, 36:1, 1985.

[55] A. Kratzer. Die ultraroten Rotationsspektren der Halogenwasserstoffe. *Z. Phys.*, 3:289, 1920.

[56] R. Mecke. Zum Aufbau der Bandenspektra. *Z. Phys.*, 32:823, 1925.

[57] Philip M. Morse. Diatomic molecules according to the wave mechanics. II. Vibrational levels. *Physical review*, 34(1):57, 1929.

[58] C.H. Douglas Clark. XLIII. The relation between vibration frequency and nuclear separation for some simple non-hydride diatomic molecules. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 18:459–470, 1934.

[59] R. M. Badger. A relation between internuclear distances and bound force constants. *Journal of Chemical Physics*, 2:128, 1933.

[60] C.H. Douglas Clark. XLI. the application of a modified Morse formula to simple hydride diatomic molecules (Di-atoms). *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 19(126):476–485, 1935.

[61] C.H. Douglas Clark and John L. Stoves. A simple modification of Morse's rule. *Nature*, 133:873, 1934.

[62] Walter Gordy. A relation between bond force constants, bond orders, bond lengths, and the electronegativities of the bonded atoms. *The Journal of Chemical Physics*, 14:305–320, 1946.

[63] K. M. Guggenheimer. New regularities in vibrational spectra. *Proceedings of the Physical Society*, 58(4):456–468, 1946.

[64] C. H. Douglas Clark. Systematics of band-spectral constants. Part VII. The empirical form of relations involving group number. *Trans. Faraday Soc.*, 37:299–302, 1941.

[65] C. H. Douglas Clark and K. R. Webb. Systematics of band-spectral constants. Part VI. Interrelation of equilibrium bond constant and internuclear distance. *Trans. Faraday Soc.*, 37:293–298, 1941.

[66] J. W. Linnett. The relation between potential energy and interatomic distance in some diatomic molecules. *Trans. Faraday Soc.*, 36:1123–1134, 1940.

[67] R.A. Newing. XXX. On the interrelation of molecular constants for diatomic molecules.-II. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 29(194):298–301, 1940.

[68] J. W. Linnett. The relation between potential energy and interatomic distance in some di-atomic molecules II. *Trans. Faraday Soc.*, 38:1–9, 1942.

[69] Y. P. Varshni. Comparative study of potential energy functions for diatomic molecules. *Rev. Mod. Phys.*, 29:664, 1957.

[70] Y. P. Varshni. Correlation of molecular constants. II. Relation between force constant and equilibrium internuclear distance. *J. Chem. Phys.*, 28:1081, 1958.

[71] Ray Hefferlin. *Periodic systems and their relation to the systematic analysis of molecular data*. The Edwin Mellen Press, Queenston, Canada, 1989.

[72] Lionel Salem. Theoretical interpretation of force constants. *The Journal of Chemical Physics*, 38(5):1227–1236, 1963.

[73] Ho Jing Kim and Robert G. Parr. Integral Hellmann-Feynman theorem. *The Journal of Chemical Physics*, 41(9):2892–2897, 1964.

[74] Raymond F. Borkman and Robert G. Parr. Toward an understanding of potential-energy functions for diatomic molecules. *The Journal of Chemical Physics*, 48(3):1116–1126, 1968.

[75] W. T. King. Calculation of molecular force constants. *The Journal of Chemical Physics*, 49(6):2866–2867, 1968.

[76] Raymond F. Borkman, Gary Simons, and Robert G. Parr. Simple bond-charge model for potential-energy curves of heteronuclear diatomic molecules. *The Journal of Chemical Physics*, 50(1):58–65, 1969.

[77] Peter Politzer. Constant term in the energy function of a point-charge model of diatomic molecules. *The Journal of Chemical Physics*, 52(4):2157–2158, 1970.

[78] Alfred B. Anderson, Nicholas C. Handy, and Robert G. Parr. Relationships between vibrational force constants and quadrupole coupling constants for molecules and solids. *The Journal of Chemical Physics*, 50(8):3634–3635, 1969.

[79] Alfred B. Anderson and Robert G. Parr. Vibrational force constants from electron densities. *The Journal of Chemical Physics*, 53(8):3375–3376, 1970.

[80] Alfred B. Anderson and Robert G. Parr. Diatomic vibrational potential functions from integration of a Poisson equation. *The Journal of Chemical Physics*, 55(12):5490–5493, 1971.

[81] Gary Simons and Robert G. Parr. Development of the bond-charge model for vibrating diatomic molecules. *The Journal of Chemical Physics*, 55:4197–4202, 1971.

[82] Alfred B. Anderson. On effective molecular electronic charge densities and vibrational potential energy functions. *Journal of Molecular Spectroscopy*, 44(3):411–424, 1972.

[83] J. L. Gazquez and Robert G. Parr. Universal dissociation energy relationships for diatomic molecules. *Chemical Physics Letters*, 66(3):419–422, 1979.

[84] Krishnan Raghavachari, Gary W. Trucks, John A. Pople, and Martin Head-Gordon. A fifth-order perturbation comparison of electron correlation theories. *Chemical Physics Letters*, 157(6):479–483, 1989.

[85] Rodney J. Bartlett, J. D. Watts, S. A. Kucharski, and J. Noga. Noniterative fifth-order triple and quadruple excitation energy corrections in correlated methods. *Chemical Physics Letters*, 165(6):513–522, 1990.

[86] Narbe Mardirossian and Martin Head-Gordon. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Molecular Physics*, 115(19):2315–2372, 2017.

[87] John P. Perdew, Adrienn Ruzsinszky, Jianmin Tao, Viktor N. Staroverov, Gustavo E. Scuseria, and Gábor I. Csonka. Prescription for the design and selection of density functional approximations: More constraint satisfaction with fewer fits. *The Journal of Chemical Physics*, 123(6):062201, 2005.

[88] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical review letters*, 77(18):3865, 1996.

[89] Jianmin Tao, John P. Perdew, Viktor N. Staroverov, and Gustavo E. Scuseria. Climbing the density functional ladder: Nonempirical meta–generalized gradient approximation designed for molecules and solids. *Physical Review Letters*, 91(14):146401, 2003.

[90] Yan Zhao and Donald G. Truhlar. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four m06-class functionals and 12 other functionals. *Theoretical Chemistry Accounts*, 120(1-3):215–241, 2008.

[91] Michael J. Willatt, Félix Musil, and Michele Ceriotti. Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements. *Physical Chemistry Chemical Physics*, 20(47):29661–29668, 2018.

[92] Christopher K. I. Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[93] Xiangyue Liu, Gerard Meijer, and Jesús Pérez-Ríos. On the relationship between spectroscopic constants of diatomic molecules: a machine learning approach. *RSC Adv.*, 11:14552–14561, 2021.

[94] Jesús Pérez-Ríos. *An Introduction to Cold and Ultracold Chemistry*. Springer International Publishing, 2020.

[95] N. Balakrishnan. Perspective: Ultracold molecules and the dawn of cold controlled chemistry. *The Journal of Chemical Physics*, 145(15):150901, 2016.

[96] John C. Slater. Atomic radii in crystals. *The Journal of Chemical Physics*, 41(10):3199–3204, 1964.

[97] MATLAB. *9.7.0 (R2019b)*. The MathWorks Inc., Natick, Massachusetts, 2019.

[98] Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*, 2018.

[99] S. Glasstone. Recent advances in physical chemistry. by s. glasstone, ph.d., d.sc. 2nd ed. pp. viii + 498. london: J. & a. churchill, 1933. 15s. *Journal of the Society of Chemical Industry*, 53(18):380–380, 1934.

[100] Mansel Davies. Simple potential functions and the hydrogen halide molecules. *The Journal of Chemical Physics*, 17(4):374–379, 1949.

[101] D. F. Heath, J. W. Linnett, and P. J. Wheatley. Molecular force fields. Part XII.—Force constant relationships in the non-metallic hydrides. *Trans. Faraday Soc.*, 46:137–146, 1950.

[102] R. T. Birge. The quantum structure of the oh bands. *Phys. Rev.*, 25:240, 1925.

[103] Alfred B. Anderson and Robert G. Parr. Universal force constant relationships and a definition of atomic radius. *Chemical Physics Letters*, 10(3):293–296, 1971.

[104] S.K. McLamarrah, P.M. Sheridan, and Lucy M. Ziurys. The pure rotational spectrum of CoO ($x^4\delta_i$): Identifying the high-spin components. *Chemical Physics Letters*, 414(4-6):301–306, 2005.

[105] Dale J. Brugh, Michael D. Morse, Apostolos Kalemos, and Aristides Mavridis. Electronic spectroscopy and electronic structure of diatomic CrC. *The Journal of Chemical Physics*, 133(3):034303, 2010.

[106] S.K. Mishra, Raj K.S. Yadav, V.B. Singh, and S.B. Rai. Spectroscopic studies of diatomic indium halides. *Journal of physical and chemical reference data*, 33(2):453–470, 2004.

[107] Maria A. Garcia, Carolin Vietz, Fernando Ruipérez, Michael D. Morse, and Ivan Infante. Electronic spectroscopy and electronic structure of diatomic IrSi. *The Journal of Chemical Physics*, 138(15):154306, 2013.

[108] Timothy C. Steimle, Ruohan Zhang, and Hailing Wang. The electric dipole moment of magnesium deuteride, MgD. *The Journal of Chemical Physics*, 140(22):224308, 2014.

[109] Ryan S. DaBell, Raymond G. Meyer, and Michael D. Morse. Electronic structure of the 4d transition metal carbides: Dispersed fluorescence spectroscopy of MoC, RuC, and PdC. *The Journal of Chemical Physics*, 114(7):2938–2954, 2001.

[110] M.A. Burton and Lucy M. Ziurys. The pure rotational spectrum of the ZnBr radical ($x^2\sigma^+$): Trends in the zinc halide series. *The Journal of Chemical Physics*, 150(3):034303, 2019.

[111] Dale J. Brugh and Michael D. Morse. Resonant two-photon ionization spectroscopy of NiC. *The Journal of Chemical Physics*, 117(23):10703–10714, 2002.

[112] R.S. Ram and P.F. Bernath. Fourier transform infrared emission spectroscopy of a new $a^3\pi_i$-$x^3\sigma^-$ system of NiO. *Journal of Molecular Spectroscopy*, 155(2):315–325, 1992.

[113] R. S. Ram, S. Yu, I. Gordon, and P.F. Bernath. Fourier transform infrared emission spectroscopy of new systems of NiS. *Journal of Molecular Spectroscopy*, 258(1-2):20–25, 2009.

[114] Corey J. Evans, Lisa-Maria E. Needham, Nicholas R. Walker, Hansjochen Köckert, Daniel P. Zaleski, and Susanna L. Stephens. The pure rotational spectra of the open-shell diatomic molecules PbI and SnI. *The Journal of Chemical Physics*, 143(24):244309, 2015.

[115] Timothy C. Steimle, Wilton L. Virgo, and Tongmei Ma. The permanent electric dipole moment and hyperfine interaction in ruthenium monoflouride (RuF). *The Journal of Chemical Physics*, 124(2):024309, 2006.

[116] Ivan O. Antonov and Michael C. Heaven. Spectroscopic and theoretical investigations of UF and UF+. *The Journal of Physical Chemistry A*, 117(39):9684–9694, 2013.

[117] Leonid A. Kaledin, John E. McCord, and Michael C. Heaven. Laser spectroscopy of UO: Characterization and assignment of states in the 0-to 3-eV range, with a comparison to the electronic structure of ThO. *Journal of Molecular Spectroscopy*, 164(1):27–65, 1994.

[118] Shane M. Sickafoose, Adam W. Smith, and Michael D. Morse. Optical spectroscopy of tungsten carbide (WC). *The Journal of Chemical Physics*, 116(3):993–1002, 2002.

[119] Alonzo Martinez and Michael D. Morse. Spectroscopy of diatomic ZrF and ZrCl: 760–555 nm. *The Journal of Chemical Physics*, 135(2):024308, 2011.

[120] Irene Shim and Karl A. Gingerich. Electronic states and nature of bonding in the molecule MoC by all electron *ab initio* calculations. *The Journal of Chemical Physics*, 106(19):8093–8100, 1997.

[121] Irene Shim and Karl A. Gingerich. All-electron *abinitio* investigations of the three lowest-lying electronic states of the RuC molecule. *Chemical Physics Letters*, 317(3-5):338–345, 2000.

[122] C. A. Coulson. *The shape and structure of molecules*. Clarendon Press, Oxford, 1973.

[123] Maxim V. Ivanov, Felix H. Bangerter, and Anna I. Krylov. Towards a rational design of laser-coolable molecules: insights from equation-of-motion coupled-cluster calculations. *Phys. Chem. Chem. Phys.*, 21:19447–19457, 2019.

[124] M. D. Di Rosa. Laser-cooling molecules. *The European Physical Journal D - Atomic, Molecular, Optical and Plasma Physics*, 31(2):395–402, 2004.

[125] Benjamin L. Augenbraun, John M. Doyle, Tanya Zelevinsky, and Ivan Kozyryev. Molecular asymmetry and optical cycling: Laser cooling asymmetric top molecules. *Phys. Rev. X*, 10:031022, Jul 2020.

[126] S. Truppe, S. Marx, S. Kray, M. Doppelbauer, S. Hofsäss, H. C. Schewe, N. Walter, J. Pérez-Ríos, B. G. Sartakov, and G. Meijer. Spectroscopic characterization of aluminum monofluoride with relevance to laser cooling and trapping. *Phys. Rev. A*, 100:052513, Nov 2019.

[127] Mahmoud A. E. Ibrahim, X. Liu, and J. Pérez-Ríos. Spectroscopic constants from atomic properties: a machine learning approach. August 2023.

[128] Andrew McHutchon and Carl Rasmussen. Gaussian process training with input noise. *Advances in neural information processing systems*, 24, 2011.

[129] Shilin Hou and Peter F. Bernath. Relationship between dipole moments and harmonic vibrational frequencies in diatomic molecules. *The Journal of Physical Chemistry A*, 119(8):1435–1438, 2015.

[130] Shilin Hou and Peter F. Bernath. Relationships between dipole moments of diatomic molecules. *Phys. Chem. Chem. Phys.*, 17:4708–4713, 2015.

[131] Linus Pauling. *The nature of the chemical bond and the structure of molecules and crystals: An introduction to modern structural chemistry (3rd ed.)*. Cornell University Press, Ithaca, N.Y., 1986.

[132] Robert S. Mulliken. Electronic structures of molecules XI. Electroaffinity, molecular orbitals and dipole moments. *The Journal of Chemical Physics*, 3(9):573–585, 1935.

[133] C. A. Coulson. *Valence*. Clarendon Press, Oxford, Oxford, United Kingdom, 1952.

[134] Martin Klessinger. Polarity of covalent bonds. *Angewandte Chemie International Edition in English*, 9(7):500–512, 1970.

[135] William M. Haynes. *CRC handbook of chemistry and physics*. CRC press, 2014.

[136] T. Andersen, H. K. Haugen, and H. Hotop. Binding energies in atomic negative ions: III. *Journal of Physical and Chemical Reference Data*, 28(6):1511–1533, 1999.

[137] Steven G. Bratsch and J. J. Lagowski. Predicted stabilities of monatomic anions in water and liquid ammonia at 298.15 K. *Polyhedron*, 5(11):1763–1770, 1986.

[138] Atomic spectra database - ionization energies form, https://physics.nist.gov/physrefdata/asd/ionenergy.html.

[139] Johann V. Pototschnig, Andreas W. Hauser, and Wolfgang E. Ernst. Electric dipole moments and chemical bonding of diatomic alkali–alkaline earth molecules. *Physical Chemistry Chemical Physics*, 18(8):5964–5973, 2016.

[140] N. Bruce Hannay and Charles P. Smyth. The dipole moment of hydrogen fluoride and the ionic character of bonds. *Journal of the American Chemical Society*, 68(2):171–173, 1946.

[141] Philip J. Stephens, F. J. Devlin, C. F. N. Chabalowski, and Michael J. Frisch. *Ab initio* calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *The Journal of physical chemistry*, 98(45):11623–11627, 1994.

[142] Martin Kaupp, P. v. R. Schleyer, H. Stoll, and H. Preuss. Pseudopotential approaches to Ca, Sr, and Ba hydrides. Why are some alkaline earth $MX_2$ compounds bent? *The Journal of Chemical Physics*, 94(2):1360–1366, 1991.

[143] Thierry Leininger, Andreas Nicklass, Wolfgang Küchle, Hermann Stoll, Michael Dolg, and Andreas Bergner. The accuracy of the pseudopotential approximation: Non-frozen-core effects for spectroscopic constants of alkali fluorides XF (X= K, Rb, Cs). *Chemical Physics Letters*, 255(4-6):274–280, 1996.

[144] Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics*, 7(18):3297–3305, 2005.

[145] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji,

X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox. Gaussian 16 Revision C.01, 2016. Gaussian Inc. Wallingford CT.

[146] J. Hoeft, F.J. Lovas, E. Tiemann, and T. Törring. Dipole moments and hyperfine structure of the group IV/VI diatomic molecules. *The Journal of Chemical Physics*, 53(7):2736–2743, 1970.

[147] Christopher J. Barden, Jonathan C. Rienstra-Kiracofe, and Henry F. Schaefer III. Homonuclear 3d transition-metal diatomics: A systematic density functional theory study. *The Journal of Chemical Physics*, 113(2):690–700, 2000.

[148] Susumu Yanagisawa, Takao Tsuneda, and Kimihiko Hirao. An investigation of density functionals: The first-row transition metal dimer calculations. *The Journal of Chemical Physics*, 112(2):545–553, 2000.

[149] Gennady L. Gutsev and Charles W. Bauschlicher. Chemical bonding, electron affinity, and ionization energies of the homonuclear 3d metal dimers. *The Journal of Physical Chemistry A*, 107(23):4755–4767, 2003.

[150] Nathan E. Schultz, Yan Zhao, and Donald G. Truhlar. Databases for transition element bonding: Metal- metal bond energies and bond lengths and their use to test hybrid, hybrid meta, and meta density functionals and generalized gradient approximations. *The Journal of Physical Chemistry A*, 109(19):4388–4403, 2005.

[151] Filipp Furche and John P. Perdew. The performance of semilocal and hybrid density functionals in 3d transition-metal chemistry. *The Journal of Chemical Physics*, 124(4):044103, 2006.

[152] Michael Bühl and Hendrik Kabrede. Geometries of transition-metal complexes from density-functional theory. *Journal of chemical theory and computation*, 2(5):1282–1290, 2006.

[153] Kasper P Jensen, Björn O Roos, and Ulf Ryde. Performance of density functionals for first row transition metal systems. *The Journal of Chemical Physics*, 126(1):014103, 2007.

[154] Erich Goll, Hermann Stoll, Christian Thierfelder, and Peter Schwerdtfeger. Improved dipole moments by combining short-range gradient-corrected density-functional theory with long-range wave-function methods. *Physical Review A*, 76(3):032507, 2007.

[155] Christopher J. Cramer and Donald G. Truhlar. Density functional theory for transition metals and transition metal chemistry. *Physical Chemistry Chemical Physics*, 11(46):10757–10816, 2009.

[156] Anastassia Sorkin, Donald G. Truhlar, and Elizabeth A. Amin. Energies, geometries, and charge distributions of Zn molecules, clusters, and biocenters from coupled cluster, density functional, and neglect of diatomic differential overlap models. *Journal of chemical theory and computation*, 5(5):1254–1265, 2009.

[157] Diptarka Hait and Martin Head-Gordon. How accurate is density functional theory at predicting dipole moments? An assessment using a new database of 200 benchmark values. *Journal of chemical theory and computation*, 14(4):1969–1981, 2018.

[158] Diptarka Hait and Martin Head-Gordon. Communication: xDH double hybrid functionals can be qualitatively incorrect for non-equilibrium geometries: Dipole moment inversion and barriers to radical-radical association using XYG3 and XYGJ-OS. *The Journal of Chemical Physics*, 148(17):171102, 2018.

[159] Zhi Wei Seh, Jakob Kibsgaard, Colin F. Dickens, I. B. Chorkendorff, Jens K. Nørskov, and Thomas F. Jaramillo. Combining theory and experiment in electrocatalysis: Insights into materials design. *Science*, 355(6321):eaad4998, 2017.

[160] Federico Calle-Vallejo, Jakub Tymoczko, Viktor Colic, Quang Huy Vu, Marcus D Pohl, Karina Morgenstern, David Loffreda, Philippe Sautet, Wolfgang Schuhmann, and Aliaksandr S. Bandarenka. Finding optimal surface sites on heterogeneous catalysts by counting nearest neighbors. *Science*, 350(6257):185–189, 2015.

[161] Weiqi Wang, Xiangyue Liu, and Jesús Pérez-Ríos. Complex reaction network thermodynamic and kinetic autoconstruction based on *Ab Initio*

statistical mechanics: A case study of o2 activation on $ag_4$ clusters. *The Journal of Physical Chemistry A*, 125(25):5670–5680, 07 2021.

[162] Johannes T. Margraf, Ajith Perera, Jesse J. Lutz, and Rodney J. Bartlett. Single-reference coupled cluster theory for multi-reference problems. *The Journal of Chemical Physics*, 147(18):184101, 2017.

[163] Xiangyue Liu, Gerard Meijer, and Jesús Pérez-Ríos. A data-driven approach to determine dipole moments of diatomic molecules. *Phys. Chem. Chem. Phys.*, 22:24191–24200, 2020.

[164] Xiangyue Liu, Laura McKemmish, and Jesús Pérez-Ríos. The performance of CCSD(T) for the calculation of dipole moments in diatomics. *Physical Chemistry Chemical Physics*, 25(5):4093–4104, 2023.

[165] Henry F. Schaefer. *Methods of electronic structure theory*, volume 3. Springer Science & Business Media, 2013.

[166] Frank Jensen. The basis set convergence of the Hartree–Fock energy for $H_2$. *The Journal of Chemical Physics*, 110(14):6601–6605, 1999.

[167] Kim Aa Christensen and Frank Jensen. The basis set convergence of the density functional energy for $h_2$. *Chemical Physics Letters*, 317(3-5):400–403, 2000.

[168] Florian Weigend, Filipp Furche, and Reinhart Ahlrichs. Gaussian basis sets of quadruple zeta valence quality for atoms H–Kr. *The Journal of Chemical Physics*, 119(24):12753–12762, 2003.

[169] Dmitrij Rappoport and Filipp Furche. Property-optimized gaussian basis sets for molecular response calculations. *The Journal of Chemical Physics*, 133(13):134105, 2010.

[170] Frank Jensen. Polarization consistent basis sets: Principles. *The Journal of Chemical Physics*, 115(20):9113–9125, 2001.

[171] Harry Partridge. Near Hartree–Fock quality GTO basis sets for the first-and third-row atoms. *The Journal of Chemical Physics*, 90(2):1043–1047, 1989.

[172] Nikolai B. Balabanov and Kirk A. Peterson. Systematically convergent basis sets for transition metals. I. All-electron correlation consistent basis sets for the 3d elements Sc–Zn. *The Journal of Chemical Physics*, 123(6):064107, 2005.

[173] Rick A. Kendall, Thom H. Dunning Jr., and Robert J. Harrison. Electron affinities of the first-row atoms revisited. systematic basis sets and wave functions. *The Journal of Chemical Physics*, 96(9):6796–6806, 1992.

[174] Kirk A. Peterson and Thom H. Dunning Jr. Accurate correlation consistent basis sets for molecular core–valence correlation effects: The second row atoms Al–Ar, and the first row atoms B–Ne revisited. *The Journal of Chemical Physics*, 117(23):10548–10560, 2002.

[175] Detlev Figgen, Guntram Rauhut, Michael Dolg, and Hermann Stoll. Energy-consistent pseudopotentials for group 11 and 12 atoms: adjustment to multi-configuration Dirac–Hartree–Fock data. *Chemical physics*, 311(1-2):227–244, 2005.

[176] Detlev Figgen, Kirk A. Peterson, Michael Dolg, and Hermann Stoll. Energy-consistent pseudopotentials and correlation consistent basis sets for the 5d elements Hf–Pt. *The Journal of Chemical Physics*, 130(16):164108, 2009.

[177] Bernhard Metz, Hermann Stoll, and Michael Dolg. Small-core multiconfiguration-Dirac–Hartree–Fock-adjusted pseudopotentials for post-d main group elements: Application to PbH and PbO. *The Journal of Chemical Physics*, 113(7):2563–2569, 2000.

[178] Kirk A. Peterson. Systematically convergent basis sets with relativistic pseudopotentials. I. Correlation consistent basis sets for the post-d group 13–15 elements. *The Journal of Chemical Physics*, 119(21):11099–11112, 2003.

[179] Kirk A. Peterson, Detlev Figgen, Erich Goll, Hermann Stoll, and Michael Dolg. Systematically convergent basis sets with relativistic pseudopotentials. II. Small-core pseudopotentials and correlation consistent basis sets for the post-d group 16–18 elements. *The Journal of Chemical Physics*, 119(21):11113–11123, 2003.

[180] Kirk A. Peterson, Benjamin C. Shepler, Detlev Figgen, and Hermann Stoll. On the spectroscopic and thermochemical properties of ClO, BrO, IO, and their anions. *The Journal of Physical Chemistry A*, 110(51):13877–13883, 2006.

[181] Kirk A. Peterson, Detlev Figgen, Michael Dolg, and Hermann Stoll. Energy-consistent relativistic pseudopotentials and correlation consis-

tent basis sets for the 4d elements Y–Pd. *The Journal of Chemical Physics*, 126(12):124101, 2007.

[182] Kirk A. Peterson and Kazim E. Yousaf. Molecular core-valence correlation effects involving the post-d elements Ga–Rn: Benchmarks and new pseudopotential-based correlation consistent basis sets. *The Journal of Chemical Physics*, 133(17):174116, 2010.

[183] David Feller. The role of databases in support of computational chemistry calculations. *Journal of computational chemistry*, 17(13):1571–1586, 1996.

[184] Karen L. Schuchardt, Brett T. Didier, Todd Elsethagen, Lisong Sun, Vidhya Gurumoorthi, Jared Chase, Jun Li, and Theresa L. Windus. Basis set exchange: a community database for computational sciences. *Journal of chemical information and modeling*, 47(3):1045–1052, 2007.

[185] Benjamin P. Pritchard, Doaa Altarawy, Brett Didier, Tara D. Gibson, and Theresa L. Windus. New basis set exchange: An open, up-to-date resource for the molecular sciences community. *Journal of chemical information and modeling*, 59(11):4814–4820, 2019.

[186] J. F. Stanton, J. Gauss, L. Cheng, M. E. Harding, D. A. Matthews, and P. G. Szalay. CFOUR, Coupled-Cluster techniques for Computational Chemistry, a quantum-chemical program package. With contributions from A. Asthana, A.A. Auer, R.J. Bartlett, U. Benedikt, C. Berger, D.E. Bernholdt, S. Blaschke, Y. J. Bomble, S. Burger, O. Christiansen, D. Datta, F. Engel, R. Faber, J. Greiner, M. Heckert, O. Heun, M. Hilgenberg, C. Huber, T.-C. Jagau, D. Jonsson, J. Jusélius, T. Kirsch, M.-P. Kitsaras, K. Klein, G.M. Kopper, W.J. Lauderdale, F. Lipparini, J. Liu, T. Metzroth, L.A. Mück, D.P. O'Neill, T. Nottoli, J. Oswald, D.R. Price, E. Prochnow, C. Puzzarini, K. Ruud, F. Schiffmann, W. Schwalbach, C. Simmons, S. Stopkowicz, A. Tajti, J. Vázquez, F. Wang, J.D. Watts, C. Zhang, X. Zheng, and the integral packages MOLECULE (J. Almlöf and P.R. Taylor), PROPS (P.R. Taylor), ABACUS (T. Helgaker, H.J. Aa. Jensen, P. Jørgensen, and J. Olsen), and ECP routines by A. V. Mitin and C. van Wüllen. For the current version, see http://www.cfour.de.

[187] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, M. Schütz, et al. Molpro, version 2019.2, a package of*ab initio* programs, 2019. see https://www.molpro.net.

[188] Hans-Joachim W., P. J. Knowles, G. Knizia, F. R Manby, and Martin Schütz. Molpro: a general-purpose quantum chemistry program package. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2(2):242–253, 2012.

[189] David M. Bishop. Molecular vibrational and rotational motion in static and dynamic electric fields. *Reviews of Modern Physics*, 62(2):343, 1990.

[190] M. Brieger. Stark effect, polarizabilities and the electric dipole moment of heteronuclear diatomic molecules in 1$\sigma$ states. *Chemical physics*, 89(2):275–295, 1984.

[191] Andrew M. Teale, Ola B. Lutnæs, Trygve Helgaker, David J. Tozer, and Jürgen Gauss. Benchmarking density-functional theory calculations of NMR shielding constants and spin–rotation constants using accurate coupled-cluster calculations. *The Journal of Chemical Physics*, 138(2):024111, 2013.

[192] Victoria E. Ingamells, Manthos G. Papadopoulos, and Stavros G. Raptis. Vibrational effects on the polarizability and second hyperpolarizability of ethylene. *Chemical Physics Letters*, 307(5-6):484–492, 1999.

[193] Asger Halkier, Wim Klopper, Trygve Helgaker, and Poul Jo/rgensen. Basis-set convergence of the molecular electric dipole moment. *The Journal of Chemical Physics*, 111(10):4424–4430, 1999.

[194] Frank Neese and Edward F. Valeev. Revisiting the atomic natural orbital approach for basis sets: Robust systematic basis sets for explicitly correlated and conventional correlated *ab initio* methods? *Journal of chemical theory and computation*, 7(1):33–43, 2011.

[195] Florian Weigend, Andreas Köhn, and Christof Hättig. Efficient use of the correlation consistent basis sets in resolution of the identity mp2 calculations. *The Journal of Chemical Physics*, 116(8):3175–3183, 2002.

[196] Wibe A. De Jong, Robert J. Harrison, and David A. Dixon. Parallel douglas–kroll energy and gradients in NWChem: Estimating scalar relativistic effects using Douglas–Kroll contracted basis sets. *The Journal of Chemical Physics*, 114(1):48–53, 2001.

[197] David P. Tew, Wim Klopper, Miriam Heckert, and Jürgen Gauss. Basis set limit CCSD(T) harmonic vibrational frequencies. *The Journal of Physical Chemistry A*, 111(44):11242–11248, 2007.

[198] Asger Halkier, Henrik Koch, Ove Christiansen, Poul Jørgensen, and Trygve Helgaker. First-order one-electron properties in the integral-direct coupled cluster singles and doubles model. *The Journal of Chemical Physics*, 107(3):849–866, 1997.

[199] Evangelos Miliordos and Aristides Mavridis. Electronic structure and bonding of the early 3d-transition metal diatomic oxides and their ions: Sco0,±, tio0,±, cro0,±, and mno0,±. *The Journal of Physical Chemistry A*, 114(33):8536–8572, 2010.

[200] Fang Wang and Timothy C. Steimle. Hyperfine interaction and stark effect in the b $\pi$ 3-x$\sigma^{1+}$(0, 0) band of copper monofluoride, CuF. *The Journal of Chemical Physics*, 132(5):054301, 2010.

[201] Constantine Koukounas and Aristides Mavridis. *Ab initio* study of the diatomic fluorides FeF, CoF, NiF, and CuF. *The Journal of Physical Chemistry A*, 112(44):11235–11250, 2008.

[202] Aggelos Avramopoulos, Victoria E. Ingamells, Manthos G. Papadopoulos, and Andrzej J. Sadlej. Vibrational corrections to electric properties of relativistic molecules: The coinage metal hydrides. *The Journal of Chemical Physics*, 114(1):198–210, 2001.

[203] K. P. R. Nair and J. Hoeft. Hyperfine structure and stark effect in the rotational spectrum of diatomic AgI in its electronic ground state. *Physical Review A*, 29(4):1889, 1984.

[204] Erich Goll, Hermann Stoll, Christian Thierfelder, and Peter Schwerdtfeger. Improved dipole moments by combining short-range gradient-corrected density-functional theory with long-range wave-function methods. *Physical Review A*, 76(3):032507, 2007.

[205] J. Hoeft and K. P. R. Nair. Stark effect measurements in high temperature molecules: Hyperfine structure and Stark effect in the rotational spectrum of the silver iodide molecule. *Journal of Molecular Structure*, 97:347–350, 1983.

[206] Charles H. Townes and Arthur L. Schawlow. *Microwave spectroscopy*. Courier Corporation, 2013.

[207] Robert Alan Frosch and H. M. Foley. Magnetic hyperfine structure in diatomic molecules. *Physical Review*, 88(6):1337, 1952.

[208] John M. Brown and Alan Carrington. *Rotational spectroscopy of diatomic molecules*. Cambridge university press, 2003.

[209] N. Walter, M. Doppelbauer, S. Marx, J. Seifert, X. Liu, J. Pérez-Ríos, B. G. Sartakov, S. Truppe, and G. Meijer. Spectroscopic characterization of the a$^3\pi$ state of aluminum monofluoride. *The Journal of Chemical Physics*, 156(12):124306, 2022.

[210] Jens Peder Dahl and John Avery. *Local density approximations in quantum chemistry and solid state physics*. Springer Science & Business Media, 2013.

[211] Jesus Aldegunde and Jeremy M. Hutson. Hyperfine structure of $^2\Sigma$ molecules containing alkaline-earth-metal atoms. *Physical Review A*, 97(4):042505, 2018.

[212] Pablo J. Bruna and Friedrich Grein. The $X^2\Pi$ and $A^2\Sigma+$ states of $FH^+$, $ClH^+$ and $BrH^+$: Theoretical study of their g-factors and fine/hyperfine structures. *Molecular Physics*, 104(3):429–446, 2006.

[213] Jacek Bieroń, Pekka Pyykkö, Dage Sundholm, Vladimir Kellö, and Andrzej J. Sadlej. Nuclear quadrupole moments of bromine and iodine from combined atomic and molecular data. *Physical Review A*, 64(5):052507, 2001.

[214] Antoine Aerts and Alex Brown. A revised nuclear quadrupole moment for aluminum: Theoretical nuclear quadrupole coupling constants of aluminum compounds. *The Journal of Chemical Physics*, 150(22):224302, 2019.

[215] Pekka Pyykkö. Year-2017 nuclear quadrupole moments. *Molecular Physics*, 116(10):1328–1338, 2018.

[216] Takeshi Yanai, David P. Tew, and Nicholas C. Handy. A new hybrid exchange–correlation functional using the coulomb-attenuating method (CAM-B3LYP). *Chemical Physics Letters*, 393(1-3):51–57, 2004.

[217] Thom H. Dunning. Gaussian basis sets for use in correlated molecular calculations. I. the atoms boron through neon and hydrogen. *J. Chem. Phys.*, 90, 1989.

[218] Rick A. Kendall, Thom H. Dunning, and Robert J. Harrison. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.*, 96, 1992.

[219] David E. Woon and Thom H. Dunning. Gaussian basis sets for use in correlated molecular calculations. III. the atoms aluminum through argon. *Journal of Chemical Physics*, 98, 1993.

[220] E. S. Shuman, J. F. Barry, D. R. Glenn, and D. DeMille. Radiative force from optical cycling on a diatomic molecule. *Physical review letters*, 103(22):223001, 2009.

[221] Valentina Zhelyazkova, Anne Cournol, Thomas E. Wall, Aki Matsushima, Jonathan J. Hudson, E. A. Hinds, M. R. Tarbutt, and B. E. Sauer. Laser cooling and slowing of CaF molecules. *Physical Review A*, 89(5):053416, 2014.

[222] J. Lim, J. R. Almond, M. A. Trigatzis, J. A. Devlin, N. J. Fitch, B. E. Sauer, M. R. Tarbutt, and E. A. Hinds. Laser cooled YbF molecules for measuring the electron's electric dipole moment. *Physical review letters*, 120(12):123201, 2018.

[223] S. Truppe, M. Hambach, S. M. Skoff, N. E. Bulleid, J. S. Bumby, R. J. Hendricks, E. A. Hinds, B. E. Sauer, and M. R. Tarbutt. A buffer gas beam source for short, intense and slow molecular pulses. *Journal of Modern Optics*, 65(5-6):648–656, 2018.

[224] Sidney C. Wright, Maximilian Doppelbauer, Simon Hofsäss, H. Christian Schewe, Boris Sartakov, Gerard Meijer, and Stefan Truppe. Cryogenic buffer gas beams of AlF, CaF, MgF, YbF, Al, Ca, Yb and NO–a comparison. *Molecular Physics*, 121(17-18):e2146541, 2022.

[225] O. Schullian, J. Loreau, N. Vaeck, A. van der Avoird, B. R. Heazlewood, C. J. Rennick, and T. P. Softley. Simulating rotationally inelastic collisions using a direct simulation monte carlo method. *Molecular Physics*, 113(24):3972–3978, 2015.

[226] Thomas Gantner, Manuel Koller, Xing Wu, Gerhard Rempe, and Martin Zeppenfeld. Buffer-gas cooling of molecules in the low-density regime: Comparison between simulation and experiment. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 53(14):145302, jun 2020.

[227] Yuiki Takahashi, David Shlivko, Gabriel Woolls, and Nicholas R. Hutzler. Simulation of cryogenic buffer gas beams. *Phys. Rev. Research*, 3:023018, Apr 2021.

[228] X. Liu, W. Wang, S. C. Wright, M. Doppelbauer, G. Meijer, S. Truppe, and J. Pérez-Ríos. The chemistry of AlF and CaF production in buffer gas sources. *The Journal of Chemical Physics*, 157(7):074305, 08 2022.

[229] Axel D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical review A*, 38(6):3098, 1988.

[230] Axel D. Becke. A new mixing of Hartree–Fock and local density-functional theories. *The Journal of Chemical Physics*, 98(2):1372–1377, 1993.

[231] James K. Parker, Nancy L. Garland, and H. H. Nelson. Kinetics of the reaction Al+ $SF_6$ in the temperature range 499- 813 K. *The Journal of Physical Chemistry A*, 106(2):307–311, 2002.

[232] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of Chemical Physics*, 132(15):154104, 2010.

[233] S. Hofsäss, M. Doppelbauer, S. C. Wright, S. Kray, B. G. Sartakov, J. Pérez-Ríos, G. Meijer, and S. Truppe. Optical cycling of AlF molecules. *New Journal of Physics*, 23(7):075001, jun 2021.

[234] George T. Armstrong, Sidney Marantz, and Charles F. Coyle. Heat of formation of nitrogen trifluoride and the NF bond energy. *Journal of the American Chemical Society*, 81(14):3798–3798, 1959.

[235] T. Kiang and R. N. Zare. Stepwise bond dissociation energies in sulfur hexafluoride. *Journal of the American Chemical Society*, 102(12):4024–4029, 1980.

[236] Philip D. Gregory, Matthew D. Frye, Jacob A. Blackmore, Elizabeth M. Bridge, Rahul Sawant, Jeremy M. Hutson, and Simon L. Cornish. Sticky collisions of ultracold RbCs molecules. *Nature Communications*, 10(1):3104, 2019.

[237] Arthur Christianen, Tijs Karman, and Gerrit C. Groenenboom. Quasiclassical method for calculating the density of states of ultracold collision complexes. *Physical Review A*, 100(3):032708, 2019.

[238] Jia K. Yao, Cooper A. Johnson, Nirav P. Mehta, and Kaden R. A. Hazzard. Complex collisions of ultracold molecules: A toy model. *Phys. Rev. A*, 104:053311, Nov 2021.

[239] James F. E. Croft and John L. Bohn. Long-lived complexes and chaos in ultracold molecular collisions. *Physical Review A*, 89(1):012714, 2014.

[240] Michael Mayle, Brandon P. Ruzic, and John L. Bohn. Statistical aspects of ultracold resonant scattering. *Phys. Rev. A*, 85:062712, Jun 2012.

[241] Michael Mayle, Goulven Quéméner, Brandon P. Ruzic, and John L. Bohn. Scattering of ultracold molecules in the highly resonant regime. *Phys. Rev. A*, 87:012709, Jan 2013.

[242] Yu Liu and Kang-Kuen Ni. Bimolecular chemistry in the ultracold regime. *Annual Review of Physical Chemistry*, 73(1):73–96, 2022. PMID: 34890257.

[243] Dibyendu Sardar, Arthur Christianen, Hui Li, and John L. Bohn. Four-body singlet potential-energy surface for reactions of calcium monofluoride. *Physical Review A*, 107(3):032822, 2023.

[244] Xiangyue Liu, Weiqi Wang, and Jesús Pérez-Ríos. Molecular dynamics-driven global potential energy surfaces: Application to the AlF dimer. *The Journal of Chemical Physics*, 159(14):144103, 10 2023.

[245] John R. Kitchin. Machine learning in catalysis. *Nature Catalysis*, 1(4):230–232, 2018.

[246] Marwin H.S. Segler, Mike Preuss, and Mark P. Waller. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature*, 555(7698):604–610, 2018.

[247] L. M. Raff, M. Malshe, M. Hagan, D. I. Doughan, M. G. Rockley, and R. Komanduri. Ab initio potential-energy surfaces for complex, multichannel systems using modified novelty sampling and feedforward neural networks. *The Journal of Chemical Physics*, 122(8), 2005.

[248] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and E. Weinan. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Physical review letters*, 120(14):143001, 2018.

[249] Kristof T. Schütt, Huziel E. Sauceda, P.-J. Kindermans, Alexandre Tkatchenko, and K.-R. Müller. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24), 2018.

[250] Jie Cui and Roman V. Krems. Gaussian process model for collision dynamics of complex molecules. *Physical review letters*, 115(7):073202, 2015.

[251] Jie Cui and Roman V. Krems. Efficient non-parametric fitting of potential energy surfaces for polyatomic molecules with Gaussian processes. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 49(22):224001, 2016.

[252] Aditya Kamath, Rodrigo A. Vargas-Hernández, Roman V. Krems, Tucker Carrington, and Sergei Manzhos. Neural networks vs Gaussian process regression for representing potential energy surfaces: A comparative study of fit quality and vibrational spectrum accuracy. *The Journal of Chemical Physics*, 148(24):241702, 2018.

[253] Bastiaan J. Braams and Joel M. Bowman. Permutationally invariant potential energy surfaces in high dimensionality. *International Reviews in Physical Chemistry*, 28(4):577–606, 2009.

[254] Joel M. Bowman, Gabor Czako, and Bina Fu. High-dimensional*ab initio* potential energy surfaces for reaction dynamics calculations. *Physical Chemistry Chemical Physics*, 13(18):8094–8111, 2011.

[255] Chen Qu, Qi Yu, Brian L. Van Hoozen Jr, Joel M. Bowman, and Rodrigo A. Vargas-Hernández. Assessing Gaussian process regression and permutationally invariant polynomial approaches to represent high-dimensional potential energy surfaces. *Journal of Chemical Theory and Computation*, 14(7):3381–3396, 2018.

[256] Bin Jiang and Hua Guo. Permutation invariant polynomial neural network approach to fitting potential energy surfaces. *The Journal of Chemical Physics*, 139(5):054112, 2013.

[257] Jun Li, Bin Jiang, and Hua Guo. Permutation invariant polynomial neural network approach to fitting potential energy surfaces. II. Four-atom systems. *The Journal of Chemical Physics*, 139(20), 2013.

[258] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters*, 98(14):146401, 2007.

[259] Dilshana Shanavas Rasheeda, Alberto Martín Santa Daría, Benjamin Schröder, Edit Mátyus, and Jörg Behler. High-dimensional neural network potentials for accurate vibrational frequencies: the formic acid dimer benchmark. *Physical Chemistry Chemical Physics*, 24(48):29381–29392, 2022.

[260] Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B*, 87(18):184115, 2013.

[261] Felix Musil, Andrea Grisafi, Albert P. Bartók, Christoph Ortner, Gábor Csányi, and Michele Ceriotti. Physics-inspired structural representations for molecules and materials. *Chemical Reviews*, 121(16):9759–9815, 2021.

[262] Haoyan Huo and Matthias Rupp. Unified representation of molecules and crystals for machine learning. *Machine Learning: Science and Technology*, 3(4):045017, 2022.

[263] Tamás K. Stenczel, Zakariya El-Machachi, Guoda Liepuoniute, Joe D. Morrow, Albert P. Bartók, Matt I. J. Probert, Gábor Csányi, and Volker L. Deringer. Machine-learned acceleration for molecular dynamics in CASTEP. *The Journal of Chemical Physics*, 159(4), 2023.

[264] Ryosuke Jinnouchi, Jonathan Lahnsteiner, Ferenc Karsai, Georg Kresse, and Menno Bokdam. Phase transitions of hybrid perovskites simulated by machine-learning force fields trained on the fly with Bayesian inference. *Physical review letters*, 122(22):225701, 2019.

[265] Ryosuke Jinnouchi, Ferenc Karsai, and Georg Kresse. On-the-fly machine learning force field generation: Application to melting points. *Physical Review B*, 100(1):014105, 2019.

[266] Ryosuke Jinnouchi, Ferenc Karsai, Carla Verdi, Ryoji Asahi, and Georg Kresse. Descriptors representing two-and three-body atomic distributions and their effects on the accuracy of machine-learned interatomic potentials. *The Journal of Chemical Physics*, 152(23), 2020.

[267] Ryosuke Jinnouchi, Kazutoshi Miwa, Ferenc Karsai, Georg Kresse, and Ryoji Asahi. On-the-fly active learning of interatomic potentials for large-scale atomistic simulations. *The Journal of Physical Chemistry Letters*, 11(17):6946–6955, 2020.

[268] Xiaoxiao Lu, Chenyao Shang, Lulu Li, Rongjun Chen, Bina Fu, Xin Xu, and Dong H. Zhang. Unexpected steric hindrance failure in the gas phase $F^- + (CH_3)_3CI$ $S_N2$ reaction. *Nature Communications*, 13(1):4427, 2022.

[269] Johann Gasteiger and Jure Zupan. Neural networks in chemistry. *Angewandte Chemie International Edition in English*, 32(4):503–527, 1993.

[270] Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4):045002, 2019.

[271] Kristof T Schütt, Stefan Chmiela, O. Anatole Von Lilienfeld, Alexandre Tkatchenko, Koji Tsuda, and Klaus-Robert Müller. Machine learning meets quantum physics. *Lecture Notes in Physics*, 2020.

[272] April M. Miksch, Tobias Morawietz, Johannes Kästner, Alexander Urban, and Nongnuch Artrith. Strategies for the construction of machine-learning potentials for accurate and efficient atomic-scale simulations. *Machine Learning: Science and Technology*, 2(3):031001, 2021.

[273] Heather J. Kulik, Thomas Hammerschmidt, Jonathan Schmidt, Silvana Botti, Miguel AL Marques, Mario Boley, Matthias Scheffler, Milica Todorović, Patrick Rinke, Corey Oses, et al. Roadmap on machine learning in electronic structure. *Electronic Structure*, 4(2):023004, 2022.

[274] Marcel F. Langer, Alex Goeßmann, and Matthias Rupp. Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning. *npj Computational Materials*, 8(1):41, 2022.

[275] Yafu Guan, Shuo Yang, and Dong H. Zhang. Construction of reactive potential energy surfaces with Gaussian process regression: Active data selection. *Molecular Physics*, 116(7-8):823–834, 2018.

[276] Arthur Christianen, Tijs Karman, Rodrigo A. Vargas-Hernández, Gerrit C. Groenenboom, and Roman V. Krems. Six-dimensional potential energy surface for NaK–NaK collisions: Gaussian process representation with correct asymptotic form. *The Journal of Chemical Physics*, 150(6):064106, 2019.

[277] Elena Uteva, Richard S. Graham, Richard D. Wilkinson, and Richard J. Wheatley. Active learning in Gaussian process interpolation of potential energy surfaces. *The Journal of Chemical Physics*, 149(17):174114, 2018.

[278] Justin S. Smith, Ben Nebgen, Nicholas Lubbers, Olexandr Isayev, and Adrian E. Roitberg. Less is more: Sampling chemical space with active learning. *The Journal of Chemical Physics*, 148(24), 2018.

[279] Jonathan Vandermause, Yu Xie, Jin Soo Lim, Cameron J Owen, and Boris Kozinsky. Active learning of reactive Bayesian force fields applied to heterogeneous catalysis dynamics of H/Pt. *Nature Communications*, 13(1):5183, 2022.

[280] Jonathan Vandermause, Steven B. Torrisi, Simon Batzner, Yu Xie, Lixin Sun, Alexie M. Kolpak, and Boris Kozinsky. On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. *npj Computational Materials*, 6(1):20, 2020.

[281] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[282] Enzo Marinari and Giorgio Parisi. Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett.*, 19(6):451, 1992.

[283] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314(1-2):141–151, 1999.

[284] Erik Bitzek, Pekka Koskinen, Franz Gähler, Michael Moseler, and Peter Gumbsch. Structural relaxation made simple. *Physical review letters*, 97(17):170201, 2006.

[285] Boris M. Smirnov. *Reference Data on Atomic Physics and Atomic Processes*. Springer-Verlag, Berlin, Germany, 2008.

[286] J. Hoeft, F. J. Lovas, E. Tiemann, and T. Törring. The rotational spectra and dipole moments of AgF and CuF by microwave absorption. *Zeitschrift für Naturforschung A*, 25(1):35–39, 1970.

[287] Andrzej J. Sadlej and Miroslav Urban. Mutual dependence of relativistic and electron correlation contributions to molecular properties: the dipole moment of AgH. *Chemical Physics Letters*, 176(3-4):293–302, 1991.

[288] Timothy C. Steimle, Ruohan Zhang, Chengbing Qin, and Thomas D. Varberg. Molecular-beam optical Stark and Zeeman study of the [17.8] $0^{+}$–$X^{1}\Sigma^{+}(0, 0)$ band system of AuF. *The Journal of Physical Chemistry A*, 117(46):11737–11744, 2013.

[289] Ruohan Zhang, Yuanqin Yu, Timothy C. Steimle, and Lan Cheng. The electric dipole moments in the ground states of gold oxide, AuO, and gold sulfide, AuS. *The Journal of Chemical Physics*, 146(6):064307, 2017.

[290] W. E. Ernst, Jörn Kändler, and T. Törring. Hyperfine structure and electric dipole moment of BaF X $^{2}\Sigma^{+}$. *The Journal of Chemical Physics*, 84(9):4769–4773, 1986.

[291] L. Wharton, M. Kaufman, and W. Klemperer. Electric resonance spectrum and dipole moment of BaO. *The Journal of Chemical Physics*, 37(3):621–626, 1962.

[292] C. A. Melendres, A. J. Hebert, and K. Street Jr. Radio-frequency Stark spectra and dipole moment of BaS. *The Journal of Chemical Physics*, 51(2):855–856, 1969.

[293] Frank J. Lovas and Donald R. Johnson. Microwave spectrum of BF. *The Journal of Chemical Physics*, 55(1):41–44, 1971.

[294] Ritchie Thomson and F. W. Dalby. An experimental determination of the dipole moments of the X ($^1\Sigma$) and A ($^1\Pi$) states of the BH molecule. *Canadian Journal of Physics*, 47(11):1155–1158, 1969.

[295] Alexandre A. Radzig and Boris M. Smirnov. *Reference data on atoms, molecules, and ions*, volume 31. Springer Science & Business Media, 2012.

[296] T. Törring, W. E. Ernst, and S. Kindt. Dipole moments and potential energies of the alkaline earth monohalides from an ionic model. *The Journal of Chemical Physics*, 81(10):4614–4619, 1984.

[297] Jinhai Chen and Timothy C. Steimle. The permanent electric dipole moment of calcium monodeuteride. *The Journal of Chemical Physics*, 128(14):144312, 2008.

[298] W. J. Childs, L. S. Goodman, U. Nielsen, and V. Pfeufer. Electric-dipole moment of CaF (X $^2\Sigma^+$) by molecular beam, laser-rf, double-resonance study of Stark splittings. *The Journal of Chemical Physics*, 80(6):2283–2287, 1984.

[299] Peter Schwerdtfeger, John S. McFeaters, Michael J. Liddell, Jan Hrušák, and Helmut Schwarz. Spectroscopic properties for the ground states of AuF, AuF$^+$, AuF$_2$, and Au$_2$F$_2$: A pseudopotential scalar relativistic Mo/ller–Plesset and coupled-cluster study. *The Journal of Chemical Physics*, 103(1):245–252, 1995.

[300] Toshiaki Okabayashi, Fumi Koto, Kazuhiro Tsukamoto, Emi Yamazaki, and Mitsutoshi Tanimoto. Pure rotational spectrum of gold monoxide (AuO) in the X $^2\Pi_{3/2}$ state. *Chemical Physics Letters*, 403(1-3):223–227, 2005.

[301] Austin J. Parsons, Samuel P. Gleason, and Thomas D. Varberg. High-resolution spectroscopy of the a$^4\Sigma_{3/2}$- X$_1$ $^2\Pi_{3/2}$ system of gold monosulphide in the near infrared. *Molecular Physics*, 116(23-24):3547–3553, 2018.

[302] P.F. Bernath, R.W. Field, B. Pinchemel, Y. Lefebvre, and J. Schamps. Laser spectroscopy of CaBr: $A^2\Pi$-$X^2\Sigma^+$ and $B^2\Sigma^+$-$X^2\Sigma^+$ systems. *Journal of Molecular Spectroscopy*, 88(1):175–193, 1981.

[303] W. E. Ernst, Jörn Kändler, J. Lüdtke, and T. Törring. Precise measurement of the ground state dipole moment of CaI. *The Journal of Chemical Physics*, 83(6):2744–2747, 1985.

[304] C. R. Byfleet, A. Carrington, and D. K. Russell. Electric dipole moments of open-shell diatomic molecules. *Molecular Physics*, 20(2):271–277, 1971.

[305] Edward William Kaiser. Dipole moment and hyperfine parameters of $H^{35}Cl$ and $D^{35}Cl$. *The Journal of Chemical Physics*, 53(5):1686–1703, 1970.

[306] Barbara Fabricant and J. S. Muenter. Molecular beam Zeeman effect and dipole moment sign of ClF. *The Journal of Chemical Physics*, 66(12):5274–5277, 1977.

[307] Takayoshi Amano, Shuji Saito, Eizi Hirota, Yonezo Morino, D. R. Johnson, and F. X. Powell. Microwave spectrum of the ClO radical. *Journal of Molecular Spectroscopy*, 30(1-3):275–289, 1969.

[308] Ritchie Thomson and F. W. Dalby. Experimental determination of the dipole moments of the X ($^2\Sigma^+$) and B ($^2\Sigma^+$) states of the CN molecule. *Canadian Journal of Physics*, 46(24):2815–2819, 1968.

[309] Co A. Burrus. Stark effect from 1.1 to 2.6 millimeters wavelength: $PH_3$, $PD_3$, DI, and CO. *The Journal of Chemical Physics*, 28(3):427–429, 1958.

[310] Hailing Wang, Xiujuan Zhuang, and Timothy C. Steimle. The permanent electric dipole moments of cobalt monofluoride, CoF, and monohydride, CoH. *The Journal of Chemical Physics*, 131(11):114315, 2009.

[311] Xiujuan Zhuang and Timothy C. Steimle. The electric dipole moment of cobalt monoxide, CoO. *The Journal of Chemical Physics*, 140(12):124301, 2014.

[312] Jinhai Chen, Timothy C. Steimle, and Anthony J. Merer. The permanent electric dipole moment of chromium monodeuteride, CrD. *The Journal of Chemical Physics*, 127(20):204307, 2007.

[313] Timothy C. Steimle, J. Scott Robinson, and Damian Goodridge. The permanent electric dipole moments of chromium and vanadium mononitride: CrN and VN. *The Journal of Chemical Physics*, 110(2):881–889, 1999.

[314] Timothy C. Steimle, David F. Nachman, Jeffrey E. Shirley, Charles W. Bauschlicher Jr., and Stephen R. Langhoff. The permanent electric dipole moment of chromium monoxide. *The Journal of Chemical Physics*, 91(4):2049–2053, 1989.

[315] Gisbert Winnewisser and Robert L. Cook. The dipole moment of carbon monosulfide. *Journal of Molecular Spectroscopy*, 28(2):266–268, 1968.

[316] T. L. Story Jr. and A. J. Hebert. Dipole moments of KI, RbBr, RbI, CsBr, and CsI by the electric deflection method. *The Journal of Chemical Physics*, 64(2):855–858, 1976.

[317] A. J. Hebert, F. J. Lovas, C. A. Melendres, C. D. Hollowell, T. L. Story Jr., and K. Street Jr. Dipole moments of some alkali halide molecules by the molecular beam electric resonance method. *The Journal of Chemical Physics*, 48(6):2824–2825, 1968.

[318] J. McGurk, H. L. Tigelaar, S. L. Rock, C. L. Norris, and W. H. Flygare. Detection, assignment of the microwave spectrum and the molecular Stark and Zeeman effects in CSe, and the Zeeman effect and sign of the dipole moment in CS. *The Journal of Chemical Physics*, 58(4):1420–1424, 1973.

[319] Fang Wang and Timothy C. Steimle. Hyperfine interaction and Stark effect in the b $^3\Pi$ -X $^1\Sigma^+$(0, 0) band of copper monofluoride, CuF. *The Journal of Chemical Physics*, 132(5):054301, 2010.

[320] Xiujuan Zhuang, Sarah E. Frey, and Timothy C. Steimle. Permanent electric dipole moment of copper monoxide, CuO. *The Journal of Chemical Physics*, 132(23):234312, 2010.

[321] Timothy C. Steimle, Wen-Lie Chang, David F. Nachman, and John M. Brown. Electronic properties of CuS: Experimental determination of the magnetic hyperfine interactions and permanent electric dipole moment. *The Journal of Chemical Physics*, 89(12):7172–7179, 1988.

[322] Timothy C. Steimle, Wilton L. Virgo, and David A. Hostutler. The permanent electric dipole moments of iron monocarbide, FeC. *The Journal of Chemical Physics*, 117(4):1511–1516, 2002.

[323] Timothy C. Steimle, Jinhai Chen, Jeremy J. Harrison, and John M. Brown. A molecular-beam optical Stark study of lines in the (1, 0) band of the $F^4\Delta_{7/2}$–$X^4\Delta_{7/2}$ transition of iron monohydride, FeH. *The Journal of Chemical Physics*, 124(18):184307, 2006.

[324] Timothy C. Steimle, David F. Nachman, Jeffrey E. Shirley, and Anthony J. Merer. The permanent electric dipole moment of iron monoxide. *The Journal of Chemical Physics*, 90(10):5360–5363, 1989.

[325] John W. Raymonda, John S. Muenter, and William A. Klemperer. Electric dipole moment of Sio and Geo. *The Journal of Chemical Physics*, 52(7):3458–3461, 1970.

[326] F.J. Hoeft, J.and Lovas, E. Tiemann, and T Törring. Elektrisches Dipolmoment von GeSe, GeTe, PbSe und PbTe. *Zeitschrift für Naturforschung A*, 25:539, 1970.

[327] J.S. Muenter and William Klemperer. Hyperfine structure constants of HF and DF. *The Journal of Chemical Physics*, 52(12):6033–6037, 1970.

[328] Anh Le, Timothy C. Steimle, Leonid Skripnikov, and Anatoly V. Titov. The molecular frame electric dipole moment and hyperfine interactions in hafnium fluoride, HfF. *The Journal of Chemical Physics*, 138(12):124313, 2013.

[329] R.D. Suenram, F.J. Lovas, G.T. Fraser, and K. Matsumura. Pulsed-nozzle fourier-transform microwave spectroscopy of laser-vaporized metal oxides: Rotational spectra and electric dipole moments of YO, LaO, ZrO, and HfO. *The Journal of Chemical Physics*, 92(8):4724–4733, 1990.

[330] P.M. Sheridan, M.A. Brewster, and Lucy M. Ziurys. Rotational rest frequencies for CrO ($X\,^5\Pi_r$) and CrN ($X\,^4\Sigma^-$). *The Astrophysical Journal*, 576(2):1108, 2002.

[331] James F. Harrison. Electronic structure of the transition metal nitrides TiN, VN, and CrN. *The Journal of Physical Chemistry*, 100(9):3513–3519, 1996.

[332] A. Durand, J.C. Loison, and J. Vigué. Spectroscopy of pendular states: Determination of the electric dipole moment of ICl in the $X^1\Sigma^+(v`' = 0)$ and $A^3\Pi_1$ ($v' = 6 - 29$) levels. *The Journal of Chemical Physics*, 106(2):477–484, 1997.

[333] F.J. Hoeft, J.and Lovas, E. Tiemann, and T Törring. Microwave absorption spectra of AlF, GaF, InF, and TiF. *Zeitschrift für Naturforschung A*, 25(7):1029–1035, 1970.

[334] Andrew J. Marr, ME. Flores, and T.C. Steimle. The optical and optical/Stark spectrum of iridium monocarbide and mononitride. *The Journal of Chemical Physics*, 104(21):8183–8196, 1996.

[335] Xiujuan Zhuang, Timothy C. Steimle, and Colan Linton. The electric dipole moment of iridium monofluoride, IrF. *The Journal of Chemical Physics*, 133(16):164310, 2010.

[336] R. Van Wachem, F.H. De Leeuw, and A. Dymanus. Dipole moments of KF and KBr measured by the molecular-beam electric-resonance method. *The Journal of Chemical Physics*, 47(7):2256–2258, 1967.

[337] A.j. Hebert, F.W. Breivogel Jr., and K. Street Jr. Radio-frequency and microwave spectra of LiBr by the molecular-beam electric-resonance method. *The Journal of Chemical Physics*, 41(8):2368–2376, 1964.

[338] Lennard Wharton, L. Peter Gold, and William Klemperer. Dipole moment of lithium hydride. *The Journal of Chemical Physics*, 33(4):1255–1255, 1960.

[339] F.W. Breivogel Jr., A.J. Hebert, and K. Street Jr. Radio-frequency and microwave spectra of $^{6}$Li$^{127}$I by the molecular-beam electric-resonance method. *The Journal of Chemical Physics*, 42(5):1555–1558, 1965.

[340] P.J. Dagdigian, J. Graff, and L. Wharton. Stark effect of NaLi $X^1\Sigma^+$. *The Journal of Chemical Physics*, 55(10):4980–4982, 1971.

[341] Timothy C. Steimle, Ruohan Zhang, and Hailing Wang. The electric dipole moment of magnesium deuteride, MgD. *The Journal of Chemical Physics*, 140(22):224308, 2014.

[342] Hailing Wang, Wilton L. Virgo, Jinhai Chen, and Timothy C. Steimle. Permanent electric dipole moment of molybdenum carbide. *The Journal of Chemical Physics*, 127(12):124302, 2007.

[343] D.A. Fletcher, K.Y. Jung, and T.C. Steimle. Molecular beam optical Stark spectroscopy of MoN. *The Journal of Chemical Physics*, 99(2):901–905, 1993.

[344] C.D. Hollowell, A.J. Hebert, and K. Street Jr. Radio-frequency and microwave spectra of NaF by the molecular-beam electric-resonance method. *The Journal of Chemical Physics*, 41(11):3540–3545, 1964.

[345] Paul J. Dagdigian. Ground state dipole moment of NaH. *The Journal of Chemical Physics*, 71(5):2328–2329, 1979.

[346] D.A. Fletcher, D. Dai, T.C. Steimle, and K. Balasubramanian. The permanent electric dipole moment of NbN. *The Journal of Chemical Physics*, 99(11):9324–9325, 1993.

[347] Jeffrey A. Gray, Steven F. Rice, and R.W. Field. The electric dipole moment of NiH $X^2\Delta_{5/2}$ and $B^2\Delta_{5/2}$. *The Journal of Chemical Physics*, 82(10):4717–4718, 1985.

[348] A.R. Hoy, J.W.C. Johns, and A.R.W. McKellar. Stark spectroscopy with the CO laser: dipole moments, hyperfine structure, and level crossing effects in the fundamental band of NO. *Canadian Journal of Physics*, 53(19):2029–2039, 1975.

[349] David D. Nelson Jr., Aram Schiffman, and David J. Nesbitt. The dipole moment function and vibrational transition intensities of OH. *The Journal of Chemical Physics*, 90(10):5455–5465, 1989.

[350] J. Hoeft, F. Lovas, E. Tiemann, R. Tischer, and T. Törring. Elektrisches Dipolmoment und Mikrowellenrotationsspektrum von SnO, SnS, PbO und PbS. *Zeitschrift für Naturforschung A*, 24(8):1222–1226, 1969.

[351] F.C. Wyse, E.L. Manson, and W. Gordy. Millimeter wave rotational spectrum and molecular constants of $^{31}P^{14}N$. *The Journal of Chemical Physics*, 57(3):1106–1108, 1972.

[352] Hiroyuki Kanata, Satoshi Yamamoto, and Shuji Saito. The dipole moment of the PO radical determined by microwave spectroscopy. *Journal of Molecular Spectroscopy*, 131(1):89–95, 1988.

[353] T.C. Steimle, K.Y. Jung, and B.-Z. Li. The permanent electric dipole moment of PtO, PtS, PtN, and PtC. *The Journal of Chemical Physics*, 103(5):1767–1772, 1995.

[354] Chengbing Qin, Ruohan Zhang, Fang Wang, and Timothy C. Steimle. The permanent electric dipole moment and hyperfine interactions in platinum monofluoride, PtF. *The Journal of Chemical Physics*, 137(5):054309, 2012.

[355] K.Y. Jung, T.C. Steimle, D. Dai, and K. Balasubramanian. Experimental determination of dipole moments, hyperfine interactions, and *ab initio* predictions for PtN. *The Journal of Chemical Physics*, 102(2):643–652, 1995.

[356] Timothy C. Steimle and Wilton L. Virgo. The permanent electric dipole moments of WN and ReN and nuclear quadrupole interaction in ReN. *The Journal of Chemical Physics*, 121(24):12411–12420, 2004.

[357] Tongmei Ma, Jamie Gengler, Zhong Wang, Hailing Wang, and Timothy C. Steimle. Molecular beam optical Stark study of rhodium mononitride. *The Journal of Chemical Physics*, 126(24):244312, 2007.

[358] Jamie Gengler, Tongmei Ma, Allan G. Adam, and Timothy C. Steimle. A molecular beam optical Stark study of the [15.8] and [16.0] $^2\Pi_{1/2}$ –X $^4\Sigma^-$ (0, 0) band systems of rhodium monoxide, RhO. *The Journal of Chemical Physics*, 126(13):134304, 2007.

[359] Timothy C. Steimle, Wilton L. Virgo, and Tongmei Ma. The permanent electric dipole moment and hyperfine interaction in ruthenium monoflouride (RuF). *The Journal of Chemical Physics*, 124(2):024309, 2006.

[360] Jeffrey Shirley, Chris Scurlock, and Timothy Steimle. Molecular-beam optical Stark spectroscopy of ScO. *The Journal of Chemical Physics*, 93(3):1568–1575, 1990.

[361] T.C. Steimle, A.j. Marr, and DM Goodridge. Dipole moments and hyperfine interactions in scandium monosulfide, ScS. *The Journal of Chemical Physics*, 107(24):10406–10414, 1997.

[362] W.L. Meerts and A. Dymanus. A molecular beam electric resonance study of the hyperfine $\Lambda$ doubling spectrum of OH, OD, SH, and SD. *Canadian Journal of Physics*, 53(19):2123–2141, 1975.

[363] W.L. Meerts and A. Dymanus. The hyperfine lambda-doubling spectrum of sulfur hydride in the $^2\Pi_{3/2}$ state. *The Astrophysical Journal*, 187:L45, 1974.

[364] J. Lovas, FJ. Hoeft, E. Tiemann, and T Törring. Elektrisches Dipolmoment und Mikrowellen-rotationsspektrum von SiS. *Zeitschrift für Naturforschung A.*, 24:1422, 1969.

[365] F.X. Powell and David R. Lide Jr. Microwave spectrum of the SO radical. *The Journal of Chemical Physics*, 41(5):1413–1419, 1964.

[366] W. E. Ernst, Jörn Kändler, S. Kindt, and T. Törring. Electric dipole moment of SrF $X^2\Sigma^+$ from high-precision Stark effect measurements. *Chemical Physics Letters*, 113(4):351–354, 1985.

[367] Fang Wang, Anh Le, Timothy C. Steimle, and Michael C. Heaven. Communication: The permanent electric dipole moment of thorium monoxide, ThO, 2011.

[368] Anh Le, Michael C. Heaven, and Timothy C. Steimle. The permanent electric dipole moment of thorium sulfide, ThS. *The Journal of Chemical Physics*, 140(2):024307, 2014.

[369] T.C. Steimle, J.E. Shirley, B. Simard, M. Vasseur, and P. Hackett. A laser spectroscopic study of gas-phase TiH. *The Journal of Chemical Physics*, 95(10):7179–7182, 1991.

[370] B. Simard, H. Niki, and P.A. Hackett. The permanent dipole moment of TiN and the nuclear magnetic hyperfine structure in its $X^2\Sigma^+$ and $a^2\Pi$ electronic states. *The Journal of Chemical Physics*, 1990.

[371] Timothy C. Steimle and Wilton Virgo. The permanent electric dipole moments of the $X^3\Delta$, $E^3\Pi$, $A^3\Phi$ and $B^3\Pi$ states of titanium monoxide, TiO. *Chemical Physics Letters*, 381(1-2):30–36, 2003.

[372] R. v. Boeckh, G. Gräff, and R. Ley. Die Abhängigkeit innerer und äußerer Wechselwirkungen des TlF-Moleküls von der Schwingung, Rotation und Isotopie. *Zeitschrift für Physik*, 179(3):285–313, 1964.

[373] R. D. Suenram, Gerald T. Fraser, Francis J. Lovas, and C. W. Gillies. Microwave spectra and electric dipole moments of $X^4\Sigma^{-1}_{12}$ VO and NbO. *Journal of Molecular Spectroscopy*, 148(1):114–122, 1991.

[374] Xiujuan Zhuang and Timothy C. Steimle. The permanent electric dipole moment of vanadium monosulfide. *The Journal of Chemical Physics*, 132(23):234304, 2010.

[375] Fang Wang and Timothy C. Steimle. Communication: Electric dipole moment and hyperfine interaction of tungsten monocarbide, WC, 2011.

[376] B.E. Sauer, Jun Wang, and E.A. Hinds. Laser-rf double resonance spectroscopy of $^{174}$YbF in the $X^2\Sigma^+$ state: Spin-rotation, hyperfine interactions, and the electric dipole moment. *The Journal of Chemical Physics*, 105(17):7412–7420, 1996.

[377] Jeffrey Shirley, Chris Scurlock, Timothy Steimle, Benoit Simard, Michael Vasseur, and P.A. Hackett. Molecular beam optical Stark spectroscopy of YF. *The Journal of Chemical Physics*, 93(12):8580–8585, 1990.

[378] Joshua H. Bartlett, Ivan O. Antonov, and Michael C. Heaven. Spectroscopic and theoretical investigations of ThS and $ThS^+$. *The Journal of Physical Chemistry A*, 117(46):12042–12048, 2013.

[379] T.M. Dunn, Louise K. Hanson, and Kenneth A. Rubinson. Rotational analysis of the red electronic emission system of titanium nitride. *Canadian Journal of Physics*, 48(14):1657–1663, 1970.

[380] A.E. Douglas and P.M. Veillette. The electronic spectrum of TiN. *The Journal of Chemical Physics*, 72(10):5378–5380, 1980.

[381] Walter J. Balfour, Anthony J. Merer, Hideaki Niki, Benoit Simard, and Peter A. Hackett. Rotational, fine, and hyperfine analyses of the (0, 0) band of the D $^3\Pi$–$X^3\Delta$ system of vanadium mononitride. *The Journal of Chemical Physics*, 99(5):3288–3303, 1993.

[382] B. Simard, C. Masoni, and P.A. Hackett. Spectroscopy and photophysics of refractory molecules at low temperature: The $d^1\Sigma^+$-$X^3\Delta_1$ intercombination system of vanadium nitride. *Journal of Molecular Spectroscopy*, 136(1):44–55, 1989.

[383] Anthony J. Merer. Spectroscopy of the diatomic 3d transition metal oxides. *Annual Review of Physical Chemistry*, 40(1):407–438, 1989.

[384] R. S. Ram and P. F. Bernath. High-resolution fourier-transform emission spectroscopy of the $A^4\Pi$–$X^4\Sigma$- system of WN. *Journal of the Optical Society of America B*, 11(1):225–230, 1994.

[385] K. P. R. Nair and J. Hoeft. Electric dipole moment of the diatomic AgBr molecule. *Chemical Physics Letters*, 102(5):438–439, 1983.

[386] Yuyan Liu, Yuanqing Guo, Jieli Lin, Guangming Huang, Chuanxi Duan, and Fengyan Li. Measurement of the electric dipole moment of NO (x $^2\pi$ $v$= 0, 1) by mid-infrared laser magnetic resonance spectroscopy. *Molecular Physics*, 99(17):1457–1461, 2001.

[387] Jeffrey Shirley, Chris Scurlock, and Timothy Steimle. Molecular-beam optical stark spectroscopy of ScO. *The Journal of Chemical Physics*, 93(3):1568–1575, 1990.

[388] J. Hoeft, F. J. Lovas, and T. Törring. Elektrisches dipolmoment und mikrowellen-rotationsspektrum von SiS. *Zeitschrift für Naturforschung A*, 24(9):1422–1423, 1969.

# LIST OF PUBLICATIONS

[1] Xiangyue Liu, Stefan Truppe, Gerard Meijer, and Jesús Pérez-Ríos. The diatomic molecular spectroscopy database. *Journal of Cheminformatics*, 12(1):1–8, 2020.

[2] Xiangyue Liu, Gerard Meijer, and Jesús Pérez-Ríos. On the relationship between spectroscopic constants of diatomic molecules: A machine learning approach. *RSC advances*, 11(24):14552–14561, 2021.

[3] Weiqi Wang, Xiangyue Liu, and Jesús Pérez-Ríos. Complex reaction network thermodynamic and kinetic autoconstruction based on *ab initio* statistical mechanics: A Case study of $O_2$ activation on $Ag_4$ clusters. *The Journal of Physical Chemistry A*, 125(25):5670–5680, 2021.

[4] Xiangyue Liu, Gerard Meijer, and Jesús Pérez-Ríos. A data-driven approach to determine dipole moments of diatomic molecules. *Phys. Chem. Chem. Phys.*, 22:24191–24200, 2020.

[5] Nicole Walter, Maximilian Doppelbauer, Silvio Marx, Johannes Seifert, Xiangyue Liu, Jesús Pérez-Ríos, Boris G Sartakov, Stefan Truppe, and Gerard Meijer. Spectroscopic characterization of the $a^3\Pi$ state of aluminum monofluoride. *The Journal of Chemical Physics*, 156(12), 2022.

[6] Xiangyue Liu, Weiqi Wang, Sidney C. Wright, Maximilian Doppelbauer, Gerard Meijer, Stefan Truppe, and Jesús Pérez-Ríos. The chemistry of AlF and CaF production in buffer gas sources. *The Journal of Chemical Physics*, 157(7), 2022.

[7] Xiangyue Liu, Laura McKemmish, and Jesús Pérez-Ríos. The performance of CCSD(T) for the calculation of dipole moments in diatomics. *Physical Chemistry Chemical Physics*, 25(5):4093–4104, 2023.

[8] Xiangyue Liu, Weiqi Wang, and Jesús Pérez-Ríos. Molecular dynamics-driven global potential energy surfaces: Application to the AlF dimer. *The Journal of Chemical Physics*, 159(14):144103, 2023.

[9] Mahmoud A. E. Ibrahim, X. Liu, and J. Pérez-Ríos. Spectroscopic constants from atomic properties: a machine learning approach. *Digital Discovery*, 3:34–50, 2024.

[10] Weiqi Wang, Xiangyue Liu, and Jesús Pérez-Ríos. AlF-AlF Reaction Dynamics between 200 K and 1000 K: Reaction Mechanisms and Intermediate Complex Characterization. *Molecules*, 29(1), 2024.