Free Energy Methods Involving Quantum Physics, Path Integrals, and Virtual Screenings

Development, Implementation and Application in Drug Discovery

Dissertation

zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)

eingereicht beim
Fachbereich Mathematik und Informatik
der Freien Universität Berlin

vorgelegt von
Christoph Gorgulla

Berlin 2018

This thesis was created within the graduate programs of the Berlin Mathematical School (BMS) and the International Max Planck Research School for Computational Biology and Scientific Computing (IMPRS-CBSC) of the Max Planck Institute for Molecular Genetics (MPIMG), as well as within the Department of Mathematics and Computer Science and the Department of Physics of the Freie Universität Berlin.

Erstgutachter (Betreuer)

Prof. Dr. Christof Schütte
Freie Universität Berlin
Fachbereich Mathematik und Informatik
Arnimallee 6
14195 Berlin, Germany

Zweitgutachter

Priv. Doz. Dr. Konstantin Fackeldey
Technische Universität Berlin
Institut für Mathematik
Straße des 17. Juni 136
10623 Berlin, Germany

Tag der Disputation: 3. Mai 2018

To my grandmother,
and all the kind, wise, and caring
people in the world.

Abstract

Computational science has the potential to solve most of the problems which pharmaceutical research is facing these days. In this field the most pivotal property is arguably the free energy of binding. Yet methods to predict this quantity with sufficient accuracy, reliability and efficiency remain elusive, and are thus not yet able to replace experimental determinations, which remains one of the unattained holy grails of computer-aided drug design (CADD). The situation is similar for methods which are used to identify promising new drug candidates with high binding affinity, which resembles a closely related endeavor in this field.

In this thesis the development of a new free energy method (QSTAR) was in the focus. It is able to explicitly take into account the quantum nature of atomic nuclei which so far was not done in binding free energy simulations of biomolecular systems. However, it can be expected to play a substantial role in such systems in particular due to the abundance of hydrogen atoms which posses one of the strongest nuclear delocalizations of all atoms. To take these nuclear quantum effects into account Feynman's path integral formulation is used and combined in a synergistic way with a novel alchemical transformation scheme. QSTAR makes also available the first readily available single topology approach for electronic structure methods (ESMs). Moreover, an extended alchemical scheme for relative binding free energies was developed to address van der Waals endpoint problems. QSTAR and the alchemical schemes were implemented in HyperQ, a new free energy simulation suite which is highly automated and scalable.

Most ESMs methods become soon prohibitively expensive with the size of the system, a restriction which can be circumvented by quantum mechanics/molecular mechanics (QM/MM) methods. In order to be able to apply QSTAR together with ESMs on biomolecular systems an enhanced QM/MM scheme was developed. It is a method for diffusive systems based on restraining potentials, and allows to define QM regions of customizable shape while being computationally fast. It was implemented in a novel client for i-PI, and together with HyperQ allows to carry out free energy simulations of biomolecular systems with potentials of very high accuracy.

One of the most promising ways to identify new hit compounds in CADD is provided by structure-based virtual screenings (SBVSs) which make use of free energy methods. In this thesis it is argued that the larger the scale of virtual screenings the higher their success. And a novel workflow system was developed called Virtual Flow, allowing to carry out SBVS-related tasks on computer clusters with virtually perfect scaling behavior and no practically relevant bounds regarding the number of nodes/CPUs. Two versions were implemented, VFLP and VFVS, dedicated to the preparation of large ligand databases and for carrying out the SBVS procedure itself.

As a primary application of the new methods and software a dedicated drug design project was started involving three regions on the novel target EBP1, expected to be located on protein-protein interfaces which are extremely challenging to inhibit. Three multistage SBVSs were carried out each involving more than 100 million compounds. Subsequent experimental binding assays indicated a remarkably high true hit rate of above 30 %. Subsequent fluorescence microscopy of one selected compound exhibited favorable biological activities in cancer cells. Other applied projects included computational hit and lead discovery for several other types of anti-cancer drugs, anti-Herpes medications, as well as antibacterials.

Acknowledgements

Each one of us can make a difference. Together we make change.

Barbara Mikulski

There are many people who have supported my doctoral research in one way or another, and I am more than grateful to all of them.

At first I would like to sincerely thank the person who made this PhD project possible, my supervisor Christof Schütte. On the one hand he provided what was most important to me: Freedom. Freedom to choose and design the contents of my PhD, and freedom to find and follow my own paths. This freedom transformed my PhD into a wonderful adventure within the realm of science. On the other hand, during this journey Christof Schütte provided me with all the support required whenever there was a need.

I am also deeply grateful to my two mentors/thesis advisors, Petra and Max, for their constant support, their exceptional kindness, and the trust they seemed to have in me from the early days on. I feel more than fortunate to having had them as my mentors, and it was a true pleasure to work with them. To Noemi, my mentor of the Berlin Mathematical School (BMS), I am also grateful for her advices on several matters.

A substantial part of this work would not have been possible without all the collaborating groups and people involved, and I would like to sincerely thank all of them for their efforts. Among them are Haribabu Arthanari and Gerhard Wagner of the Dana Farber Cancer Institute and the Harvard Medical School for their support in bringing my applied project to the experimental level, and for having invited me to visit their research groups. I am thankful to Andras Boeszoermenyi, my primary coworker in Boston regarding the experimental work, for all his time and the interesting discussion during my visit. And to our collaborators, Nancy Kedersha and Pavel Ivanov in the Brigham and Women's Hospital in Boston, who have done excellent work in regard to the fluorescence microscopy experiments. Magdalena Czuban from the Charité in Berlin I would like to thank for having worked with me on our new collaborative project, Michelle Ceriotti as well as Riccardo Petraglia from the EPFL for their support in relation to i-PI/i-QI, and Jonathan LaRochelle, as well as a few members of the Naar Lab and the Coen Lab with whom I had the joy to work with in several collaborative projects. I would also like to acknowledge Enamine, in particular Olga Tarhanova, for having supported my applied drug discovery project by freely synthesizing compounds for us from their vast virtual compound libraries.

Also, it was a true pleasure to having had such wonderful colleagues and fellow students in the Computational Biophysics Group, the Biocomputing Group, the Arthanari Lab, the Wagner Lab, the Berlin Mathematical School (BMS) and the International Max Planck Research School for Computational Biology and Scientific Computing (IMPRS-CBSC). Thank you all for having provided such a pleasant and supportive working environment.

To my friends Moritz and Brigitte I would like to express my gratitude for reading my thesis and their valuable comments, Moritz also for his support regarding Python, and Brigitte for being so meticulous in her reading and encouraging. Also Gerhard König and Han Cheng Lie I would like to thank for various discussions.

Among the administrative staff of the research groups and graduate schools I would like to thank in particular Fabian Feutlinske, Kirsten Kelleher, Chris Seyffer, Benita Ross, Dominique Schneider, Tanja Fagel, Annette Schumann-Welde, and Dorothé Auth for their constant and kind support throughout my PhD.

Considerable computational resources were required during this PhD, and I would like to thank all the sources which have granted me computing time and additional technical support. Among the high performance computers used are the Sheldon, Yoshi, and Soroban Linux clusters of the Freie Universität Berlin, the HLRN III supercomputer in Berlin/Hanover, the ERISOne cluster of Partners Healthcare/Dana Farber Cancer Institute, the Orchestra and O2 clusters of the Harvard Medical School, the Coral cluster of the Wagner Lab in the Harvard Medical School, and the Odyssey cluster of the Faculty of Harvard University. Among the technical and scientific staff related to these resources I would like to thank in particular Jens Dreger, Christian Tuma, Guido Laubender, Francesco Pontiggia, James Cuff and Scott Yockel for their support in various ways.

I would also like to acknowledge the free and open source software communities which provide the world altruistically with such magnificent and versatile software, but which we often take for granted. Their contributions played a crucial role in various aspects of this work, and using their packages has often been a delight.

This work was funded by several institutions/groups (mainly via scholarships or traveling allowances), and I would like to thank all of them for their support. These are the Einstein Center of Mathematics Berlin/BMS, the Computational Biophysics Group/Freie Universität Berlin, the IMPRS-CBSC/Max Planck Institute for Molecular Genetics, the Arthanari Lab/Dana Farber Cancer Institute, and the Schneider Group/Berlin Institute of Technology.

Content Overview

I Theory and Mathematical Modeling 9

II Model and Method Implementations 103

III Applications in Drug Discovery 161

Contents

(in Detail)

III Applications in Drug Discovery 161

List of Figures

Listings

Acronyms

ABFE absolute binding free energy.

ADMET absorption, distribution, metabolism, excretion and toxicity.

AMCS atomic maximum common substructure.

BAR Bennett acceptance ratio.

BOMD Born-Oppenheimer molecular dynamics.

CADD computer-aided drug design.

CDM canonical density matrix.

CGenFF CHARMM General Force Field.

CPF canonical partition function.

CSS common substructure.

DRMS distributed resource management system.

EBP1 ErbB3-binding protein 1.

EMA European Medicines Agency.

FDA U.S. Food and Drug Administration.

FEM free energy method.

FEP Free Energy Pertubation.

FES free energy simulation.

FIRES Flexible Inner Region Ensemble Separator.

HCV Hepatitis C virus.

HPC high performance computer.

ITEO imaginary time evolution operator.

L ligand.

LOMAP Lead Optimization Mapper.

LS ligand and solvent.

MBAR Multistate Bennett Acceptance Ratio.

MC Monte Carlo.

MD molecular dynamics.

MED15 mediator of RNA polymerase II transcription subunit 15.

MM molecular mechanics.

MPI Message Passing Interface.

MSP molecular system pair.

NBE new biological entity.

NME new molecular entity.

NMR nuclear magnetic resonance.

NQE nuclear quantum effect.

PDB Protein Data Bank.

PDBX Protein Data Bank Extended.

PEARL Parallel Endpoint Atom Removal and Location.

PEARL-B PEARL Branch-consideration.

PEARL-N PEARL Neighbor-separation.

PEARL-P PEARL Primitive.

PEARL-XH PEARL-X Hydrogen-single-step.

PES potential energy surface.

PIMC path integral Monte Carlo.

PIMD path integral molecular dynamics.

PPI protein-protein interface.

QM quantum mechanics.

QM/MM quantum mechanics/molecular mechanics.

QSTAR Quantum Sphere Transformation Alchemical Route.

QUASAR Quantum Adaptive Sphere Assembly Restraints.

R receptor.

RBFE relative binding free energy.

RDF radial distribution function.

RLS receptor, ligand and solvent.

RMSD root-mean-square deviation.

RTEO real time evolution operator.

S solvent.

SHP2 Src-homology 2 domain-containing phosphatase 2.

SMILES Simplified Molecular Input Line Entry Specification.

SREBP sterol regulatory element binding protein.

TDS thermodynamic state.

TDSE time-dependent Schrödinger equation.

TI Thermodynamic Integration.

TMCS topological maximum common substructure.

UFD Uniform Force Distribution.

VFLP Virtual Flow for Ligand Preparation.

VFVS Virtual Flow for Virtual Screening.

WHAM Weighted Histogram Analysis Method.

Nomenclature

$\bar{A}$	The closure of the operator A	15
β	Thermodynamic beta, i.e. $1/k_BT$	27
$\hbar = h/2\pi$	The reduced Planck constant	18
$\langle\cdot,\cdot\rangle$	The standard inner product of the Hilbert on which it is defined	15
$\mathcal{H}$	A separable Hilbert space. If it is a quantum mechanical Hilbert space it is over $\mathbb{C}$	15
$\mathcal{L}_{\mathrm{d}}(\mathcal{H})$	The space of all linear operators which are densely defined on the Hilbert space $\mathcal{H}$	15
$\mathcal{L}_{\mathrm{esa}}(\mathcal{H})$	The space of all essentially self-adjoin operators on the Hilbert space $\mathcal{H}$	16
$\mathcal{L}_{\mathrm{pd}}(\mathcal{H})$	The space of all positive definite operators on the Hilbert space $\mathcal{H}$	16
$\mathcal{L}_{\mathrm{p}}(\mathcal{H})$	The space of all positive operators on the Hilbert space $\mathcal{H}$	16
$\mathcal{L}_{\mathrm{sa}}(\mathcal{H})$	The space of all self-adjoint operators on the Hilbert space $\mathcal{H}$	16
$\mathcal{L}_{\mathrm{sb}}(\mathcal{H})$	The space of all semibounded operators on the Hilbert space $\mathcal{H}$	16
$\mathcal{L}_{\mathrm{s}}(\mathcal{H})$	The space of all symmetric operators on the Hilbert space $\mathcal{H}$	15
$\mathcal{U}(\mathcal{H})$	The space of all linear operators on the Hilbert space $\mathcal{H}$ which are (possibly) unbounded	15
$\Theta(A)$	The numerical range of an operator A on a Hilbert space.	15
$\Theta_l(A)$	The lower bound of the numerical range $\Theta(A)$	15
$\Theta_u(A)$	The upper bound of the numerical range $\Theta(A)$	15
$\times$	Scalar multiplication	29
$A > 0$	The operator A is positive definite	16
$A \geq 0$	The operator A is positive (semidefinite)	16
$D(A)$	The domain of an operator A	15
$R_{\mathrm{THR}}(b, K)$	True hit rate of the K top scoring screening compounds with binding strength threshold b	97
$W(\boldsymbol{r}, \kappa)$	The total QSTAR potential	72

Chapter 1

Introduction

> *The miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift, which we neither understand nor deserve.*
>
> Eugene Wigner
> *The Unreasonable Effectiveness of Mathematics in the Natural Sciences* [329]

In pharmaceutical research and development of new medications a variety of serious challenges and critical problems exist these days. Among them are the following:

- **Many Incurable Diseases.** In our world there are innumerably many human diseases, many of which can reduce the quality of life significantly or be even lethal. Alone the number of monogenic diseases (i.e. diseases caused by a mutation of a single gene) such as familial Alzheimer's disease or monogenic Parkinson's disease is estimated to be higher than 10 000 [336]. To date only a minor fraction of all known diseases is curable, which is also indicated by the number of approved drugs (see below).

- **Few New Medications.** Despite the circumstance that mankind has made tremendous efforts to find cures for most of the diseases which it had to face throughout its history, to date only about 1 600 to 1 800 small molecule drugs (NMEs: new molecular entities) and approximately 160 biological macromolecules (NBEs: new biological entities) have been approved by authorities such as the U.S. Food and Drug Administration (FDA) [331, 330, 298] or the European Medicines Agency (EMA) [83]. And despite the advances in technology, the rate at which new drugs are appearing has not increased, but rather remains constant at merely around 30 new therapeutic entities (NMEs and NBEs) on average (see also Figure 1.1 on the following page) [208]. These numbers reduce even further if only curative drugs are considered (while neglecting agents for symptomatic treatments, for instance analgesics). It follows that there are far more pharmaceutical companies than there are new drugs per year, and if the largest of them manage to bring a handful of new medications on the market in an entire year it is considered to be a substantial success.

- **High Development Costs.** The research and development costs of a single successful drug lie on average between \$ 2.5 and \$ 3 billion (these numbers include the costs of drug development projects which have failed) [70]. Moreover, the development costs per approved drug are rising dramatically, possibly even at an exponential rate as was suggested in [209]. Such rates, even if subexponential, are not sustainable in the medium and long term.

- **Long Development/Approval Times.** The average development time of approved drugs, from the early beginning to the final approval, is on average 10 to 12 years [302].
- **Animal Experiments.** Approximately 50-100 million higher animals (vertebrates) are used each year for testing new drug candidates in preclinical trials in the U.S. alone. These animals often suffer during their time in the lab, and in most cases have to leave their lives even when surviving the experiments unharmed. In some countries, many of the living beings which are used for such purposes are not covered by laws which provide a more strict protection in research (and other situations). This is for instance the case in the U.S, where the Animal Welfare Act does not cover certain classes of higher vertebrates such as birds, mice, rats and fish [207, 13].[1]

A question which arises in this context is now, what are the causes of the above mentioned problems? One cause is related to the circumstance that most of the drugs which are easy to discovered were already identified, and that the degree of difficulty of finding new drugs has risen substantially. Therefore more and more effort and/or better methods have to be used in order to find new medications. While both approaches are taken, the vast majority of pharmaceutical research and development is happening on the experimental level, which is extremely expensive and also related to the other problems mentioned above.

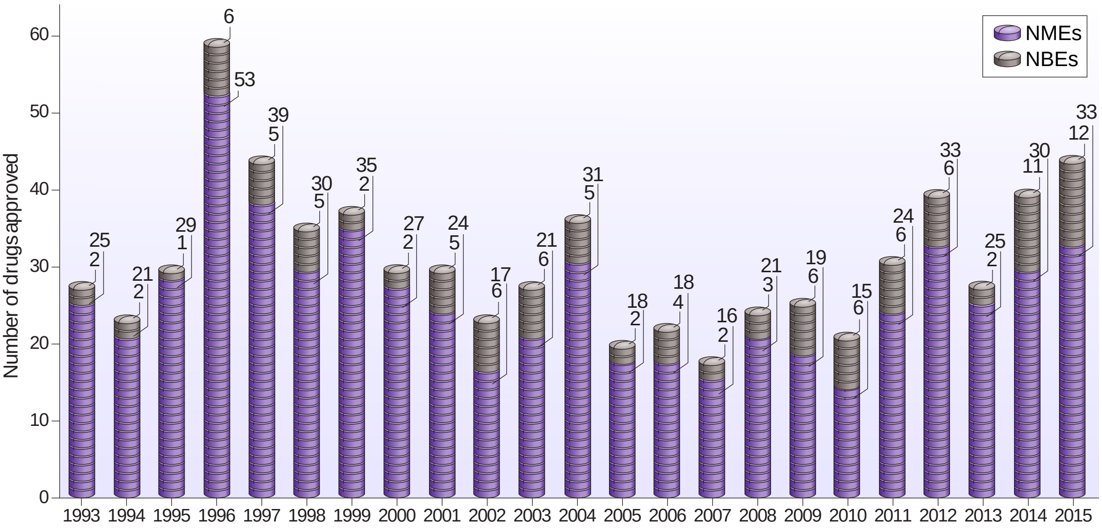


Figure 1.1: Number of newly approved drugs by the FDA in the recent decades. The average number of all new therapeutic drugs per year remains roughly the same. Adapted with permission from [208].

[1] In the European Union the EU Directive 2010/63/EU was introduced which applies to all vertebrates, and thus provides a certain level of protection to a relatively large class of animals.

Computational Drug Development

However, there is an alternative approach to experimental research which can be applied to drug development, namely computational drug development. And it has the potential to solve all of the above mentioned problems. Some of them can already be tackled by computational approaches such as the finding of new drug candidates for incurable diseases, because these approaches allow to investigate the causes of the diseases in ways which are often not possible or affordable yet on the experimental level. And this in turn can lead to increases in the number of newly approved drugs. Computational approaches also have the potential to dramatically reduce the drug development costs since the major resources which are needed are high performance computers (HPCs). While HPCs are very expensive, their cost is relatively small when compared to the expenditures related to experimental research. For instance, the biggest pharmaceutical companies spend typically between \$ 5 and \$ 10 billion each year for research and development. In comparison, the fastest supercomputers in the world cost roughly between \$ 0.05 and \$ 0.5 billion, and since they are normally used for many years (typically 5 to 10) their cost further decreases. It increases again moderately when adding the running costs (which is typically only a fraction per year of the initial costs). Yet the computational facilities of pharmaceutical companies do not seem to come even near the lower bound of the mentioned range regarding their computational facilities, even though pharmaceutical research is one of the most promising applications of HPCs.[2] Computational approaches can also speed up the drug development time since computations can be run faster than carrying out experiments if enough computing power is available. Another reason why it can reduce the costs is because it can help to decrease the attrition rate of drug development projects, and may in the far future also be able to replace preclinical and clinical trials. Such endevears might become feasible with the aid of quantum computers, which are able to simulate molecular systems on the quantum mechanical level natively at speeds many orders of magnitudes higher than classical computers (as suggested by Richard Feynman as early as 1982 [87]). If clinical trials can someday be replaced by computer simulations, even if only to a certain extent, not only will money be saved and development time reduced, it will also spare the life of animals and reduce the number of required human probands.

While computational science has the potential to solve all of the above mentioned problems, at the time being it can only do so in a very limited way. There are at least two reasons for that, namely that the methods which are used are either

(1) not reliable enough regarding precision and accuracy,

(2) not efficient enough for the computational resources which are available,

or possibly both at the same time. It follows that either faster hardware or improved methods are required, ideally both.

The drug development process consists of several consecutive stages:

(1) Drug discovery/design.

(2) Drug evaluation (preclinical and clinical trials).

[2] This impression was obtained by the author during discussions with several employees of different pharmaceutical companies which worked in their associated computational divisions.

(3) Authority review/approval procedures.

(4) Post-approval research/monitoring.

Sometimes also basic research to understand the disease and identify possible targets is in addition included as the very first step of the procedure, and while it is indeed required for carrying out target-based drug development, usually such studies are considered as falling under the field of molecular biology/pathophysiology. Computational approaches are at the present particularly useful for the first step, drug discovery (as well as basic research), while the replacement of preclinical and clinical trials remains a long term goal. Drug design can further be divided into the computer-aided drug design (CADD) in which new ligands (NMEs) are in the focus, and into computer aided biologics design where biological macromolecules (NBEs) are being developed. While the former is already a well-established field, the latter is still in its infancy.

Binding Free Energies

Computational methods in CADD are normally used to compute certain physico-chemical or pharmacological properties of small organic molecules and larger biomolecular systems. The most important property in this field is arguably the free energy of binding between small organic compounds and the target structure. The binding affinity is the quantity which is normally used as the key criterion when identifying new binders, but also in the successive steps of hit and lead optimization it is meticulously monitored to either further improve it or to make sure it remains sufficiently high when other properties are optimized. The higher the binding affinity, the lower the required dose of the drug, and therefore the lower the side effects which are almost always present to some degree (even if not sensible). Furthermore, the higher the binding affinity of the compounds the more room is there for improving other important properties (in particular those related to the absorption, distribution, metabolism, excretion and toxicity (ADMET) characteristics).

The reliable prediction of binding affinities is one of the greatest challenges of CADD, despite substantial efforts which have been made over the past decades. It is seen as one of the holy grails of the field of CADD [7, 198], and closely related also the identification of new ligands with high affinities to their targets [236].

When trying to improve the computational methods which are used to identify new drug candidates based on the estimation of the binding affinities, there exist two approaches. The free energy methods (FEMs) themselves can be improved. Or the methods which employ the FEMs for the identification of high-affinity binders, such as the virtual screening procedures, can be advanced. In this thesis both approaches are taken, the development of a novel FEM method, as well as efforts to advance structure-based virtual screening procedures. And as we will see in section 5.4 on page 100, both approaches can be elegantly combined.

Structure-based CADD can be divided into four major parts.

(1) **Structure Preparation.** The preparation of the target structure such that it is ready to be used within the hit identification procedure. This step can include homology modeling, structure refinement, conformation sampling, binding-site identification, clustering, and other related tasks.

(2) **Hit Identification/Design.** The finding of small molecules which are predicted to bind to the target sufficiently strong.

(3) **Hit to Lead Optimization.** Hit confirmation (by more reliable methods) and hit expansion (to obtain improved lead-like compounds).

(4) **Lead Optimization.** Optimization of the lead compounds regarding their binding affinities, physico-chemical and pharmacological properties such as ADMET).

While virtual screenings can be used in the second and in the third step (in the third one to screen for analogs), FEMs are of importance in all parts but the first stage. Free energy simulation (FES) methods, a special type of FEMs, depend upon force fields/potentials, but the potentials can also be used for other purposes such as vanilla molecular dynamics (MD) simulations. One type of potential is provided by quantum mechanics/molecular mechanics (QM/MM) methods, one of which is developed in this thesis which is particularly useful also for free energy simulations. Also free energy simulations can be used within virtual screenings, in particular they are attractive to be used in high-accuracy rescoring procedures. The conceivable use-cases of these three types of methods (virtual screenings, free energy methods, and QM/MM potentials) are graphically illustrated in Figure 1.2.

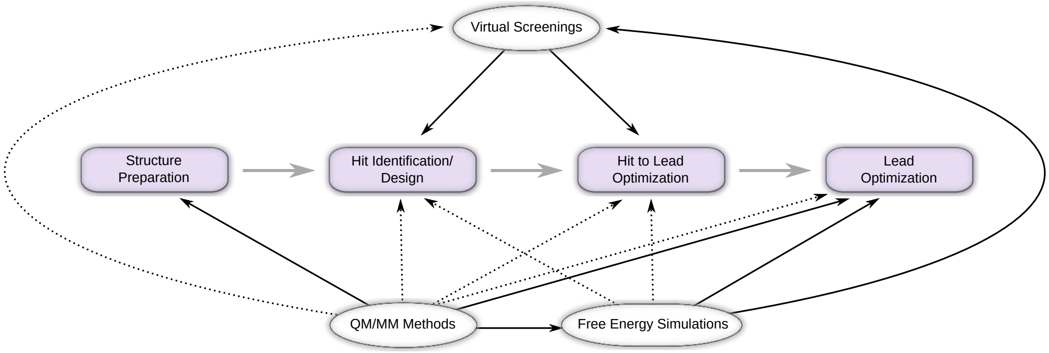


Figure 1.2: Possible application of virtual screenings, free energy simulation methods, as well as QM/MM methods in the four primary stages of computer aided drug design (pale violet). Black arrows indicate direct applicability (e.g. via MD simulations in the case of QM/MM methods), and dotted lines indicate indirect applicability via one of the other methods.

Besides their application in CADD, FEMs have a wide range of applications in physics, chemistry, biology, pharmacy and material science/nanotechnology.

Theory, Implementation and Application

While a major focus of this thesis are theoretical aspects and developments of new methods, they were also implemented and applied in this PhD. If a method is not implemented it cannot be used, and it has happened all too often in the history of science that methods were not made available and became forgotten. The new implementations presented in this thesis are also applied to real systems and projects for different reasons. On the one hand application in real projects provides valuable feedback on how well the methods and implementations work in real world problems, and thus also which aspects

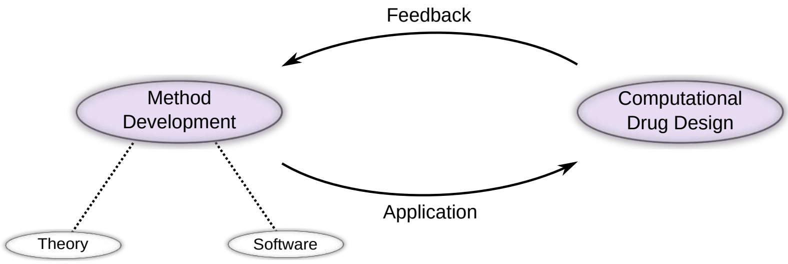


Figure 1.3: Bidirectional flow of information in a method development/application refinement loop.

need to be improved see also Figure 1.3). Also, the new methods can directly contribute to these projects and help to advance them, which can be more useful than only testing the methods on toy problems.

Real-world applications are one of the primary reasons why science is carried out and supported. Yet they are often of great challenge due to several reasons. One of them lies in the circumstance that real-world problems are often highly disciplinary and require knowledge not only of one scientific field, but of multiple disciplines such as in the case of research related to the development of new medications. Another major challenge is that real world problems are often intrinsically highly complicated and nonlinear, making for instance analytical solutions in most cases impossible.

Relevant Scientific Fields

While the work in this thesis touches multiple scientific disciplines, partially depending on the precise subproject, it can mostly be classified as applied mathematics (which itself is based on pure mathematics), and thus mathematics as a whole. The core of this thesis involving mathematical modeling, and quantum physics falls also into the field of theoretical/mathematical physics (which is one of the major branches of applied mathematics). The implementation and application of the new free energy method can moreover be associated to computational/quantum chemistry, computational biology, and CADD, all of which fall into the field of computational science. Also the work done related to virtual screenings falls into CADD, a branch of computational drug development, and thus computational science. While being a highly interdisciplinary field, computational science is a also subdiscipline of applied mathematics. In addition to the above mentioned areas, an essential part of computational science consists of computer science since software needs to be implemented to carry out the desired computations and simulations. All major disciplines which are relevant for this thesis are illustrated in Figure 1.4 on the facing page.

General Contextualization in Science

Scientific work can be characterized in various ways. One of them is by subject (on different scales, from entire sciences to specific topics). Another way of classification is by the stage of the method development process. Scientific work is normally either about the development of new methods or the application of methods (while in the former case

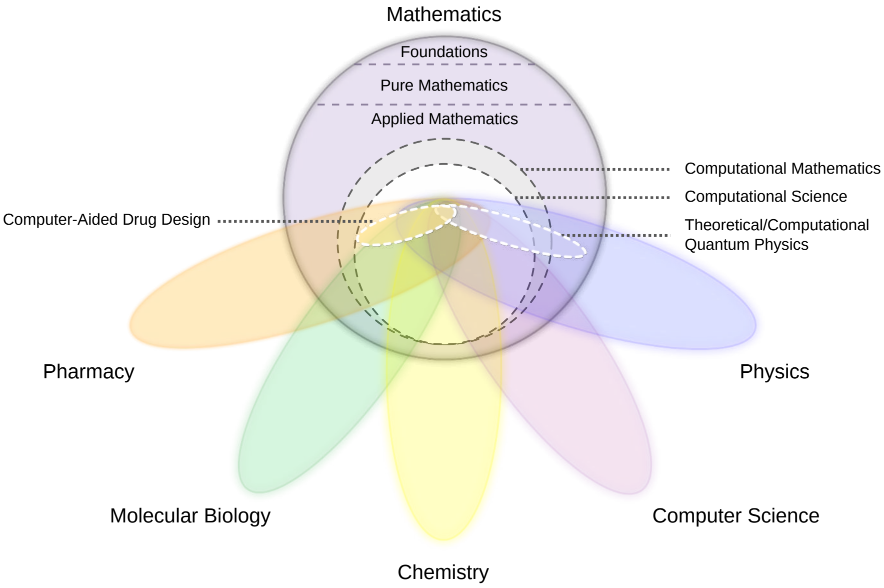


Figure 1.4: Venn diagram (non-rigorous) showing the scientific disciplines which are relevant in this thesis. The work has its root in mathematical modeling and quantum physics (both subbranches of applied mathematics), and has overlaps with computational/quantum chemistry as well as computational molecular biology, involves computer science (due the implementation of the new methods), and extends into the field of computer-aided drug design and pharmacy (due to the work on virtual screenings and the applications).

already existing methods are often applied as well), leading to the following classes:

(1) Method discovery.

(2) Method implementation.

(3) Method application.

Furthermore, scientific activities can often be classified by the mode in which knowledge is generated:

(1) Theory.

(2) Simulation.

(3) Experiment.

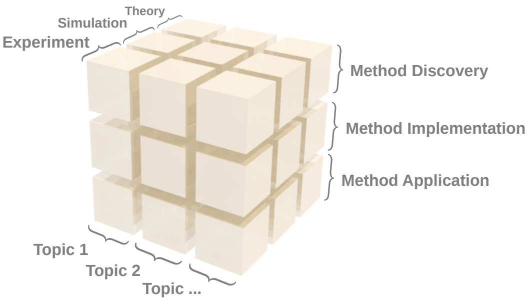


Figure 1.5: Classification of scientific tasks/activities by subject, knowledge generation type, as well as the method development-based stage.

All three classification schemes are shown in a unified manner in figure 1.5 on the previous page.

The work in this thesis was already characterized by the method development-based stages as well as the involved subjects/sciences. When considering topics in this context, they might be chosen as FEMs, the QM/MM methodology, and virtual screenings. Regarding the knowledge generation mode, while the core of this thesis lies in the theory, significant knowledge was also generated by simulation (since they are the purpose of the new methods), and experimental knowledge was produced for verification purposes.

Chapter Overview

The first part (I) in this thesis is dedicated to theory and theoretical method development, while in part II the new methods and software for CADD is implemented. In part III the new software is applied to real world applications. The relationship between the various chapters is illustrated in Figure 1.6.

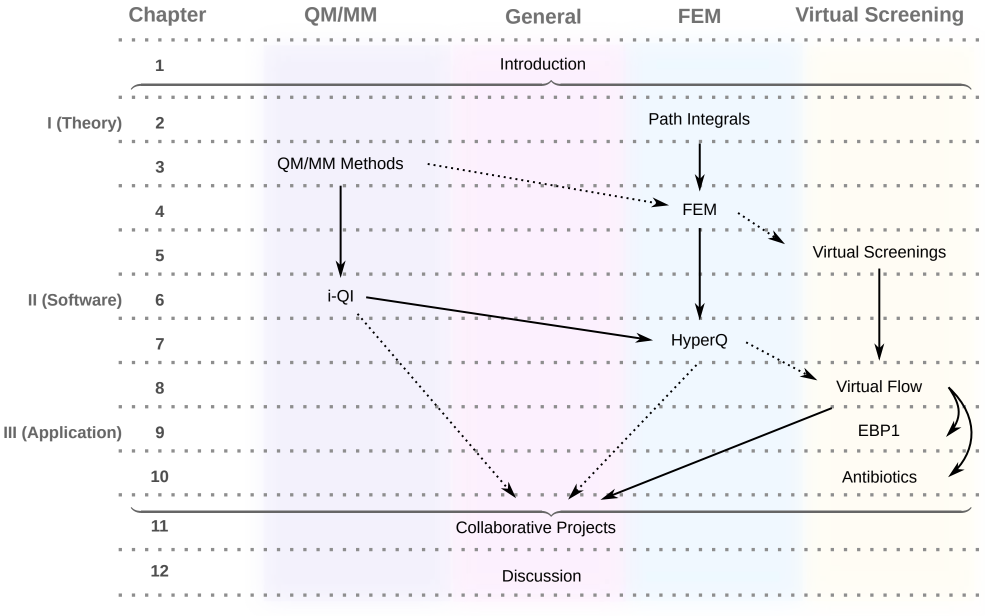


Figure 1.6: Overview of chapters in this thesis. Solid arrows imply a relation of a more direct nature, while dotted arrows indicate relations of a more indirect or optional nature.

Part I

Theory and Mathematical Modeling

Chapter 2

Path Integral Approaches

> *One might think this means that imaginary numbers are just a mathematical game having nothing to do with the real world. ... [However,] It turns out that a mathematical model involving imaginary time predicts not only effects we have already observed but also effects we have not been able to measure yet nevertheless believe in for other reasons. So what is real and what is imaginary? Is the distinction just in our minds?*
>
> Stephen Hawking
> *The Universe in a Nutshell*

Contents

2.1 Introduction

The dynamical behavior of quantum mechanical systems can be described in a number of different but equivalent ways.[1] Five of the more common formulations of quantum mechanics are the following:[2]

1. The *Schrödinger picture*, also referred to as *wave mechanics.*

2. The *Heisenberg representation*, of which *matrix mechanics* is a special case.[3]

3. The *Dirac representation*, also called the *interaction picture.*

4. The *phase space formulation*, also known as the *Wigner representation* or *Wigner-Weyl quantization.*

5. The *Feynman path integral formulation*, also referred to as the *sum-over-histories representation.*

[1] References valid for larger parts of this and other chapters are listed in appendix A on page 241.

[2] In addition there are several other alternative formulations, surveys of them can be found for instance in [76, 77, 272].

[3] Some authors use the terms *Heisenberg representation/picture* and *matrix mechanics/ representation* interchangeably.

The most well known formalism, wave mechanism, is based on the famous time-dependent Schrödinger equation (TDSE). And while the Schrödinger picture might be the most widely used and intuitive formulation, each of the other representations has their unique advantages as well. This is in particular true also for the path integral formalism, which is of remarkable elegance and usefulness in certain theories and applications, and which establishes additional connections between classical mechanics and quantum mechanics as well as between quantum processes and stochastic processes. It is a central ingredient in the newly developed free energy method laid out in Chapter 4 via the path integral molecular dynamics (PIMD) methodology (described section 2.4 on page 34), and therefore the fundamentals of this approach are summarized in this chapter.

The path integral formalism has also many other applications. It is of fundamental importance in relativistic quantum field theories, including quantum electrodynamics and quantum chromodynamics [172, 41]. Moreover it has applications in (equilibrium and non-equilibrium) quantum statistical mechanics, mathematical finance/econophysics, quantum chemistry, quantum gravity, quantum cosmology, quantum information science and other areas [297, 115, 96, 148, 220, 113, 105].

Historical Notes

The matrix mechanics formulation, which is the special case of the Heisenberg picture in the position representation, is widely seen as the first consistent formulation of quantum mechanics and with that as the birth of modern quantum mechanics. It was conceived by Werner Heisenberg in 1925, and fully formulated by him, Pascual Jordan and Max Born later in the same year [117, 37, 36]. For this achievement the Nobel Prize in Physics of the year 1932 was awarded to Werner Heisenberg *'for the creation of quantum mechanics'* [285]. Shortly thereafter wave mechanics was developed by Erwin Schrödinger in late 1925 [248], followed by the Dirac picture in 1926 [74]. The first ideas related to phase space representation of quantum mechanics can be found in articles by Hermann Weyl and Eugene Wigner of the years 1927 and 1932, respectively, and the formalism was at first fully developed by Groenewold in 1946 [323, 328, 104].

The path integral formalism of quantum mechanics has at least two independent historical roots. Only one of them is well known, namely the one which has its origin in an article published by Paul Dirac in 1933. Dirac conceived a connection between classical mechanics and quantum mechanics via the action integral which is central in Lagrangian mechanics [75]. The idea was picked up by Richard Feynman who developed the complete path integral formalism during his doctoral and postdoctoral research during the 1940s, culminating in his seminal publication of 1948 [85, 89]. The new approach to quantum mechanics was soon widely adopted and applied, and it was named after him. Also therefore it is remarkable that Paul Dirac and Richard Feynman have in fact not been the first ones to discovered and use the path integral formalism (even though Richard Feynman was the first who completed its full formulation). It was Gregor Wentzel who used the central ideas of this approach already in 1924 [321, 322]. However, his findings did not gain widespread knowledge causing it to fall into oblivion, and it was rediscovered only in the 1960s by Thomas S. Kuhn [11]. What is even more remarkable is that Wentzel's discovery precedes the matrix mechanics formulation of Werner Heisenberg by approximately one year. It therefore seems that the presumed latest major approach to quantum dynamics was in fact the overlooked first one, even though it was fully developed only later by Feynman [10].

Types of Path Integrals

The Feynman path integral is not the only type of path integral. The other major types are line integrals (i.e. integrals of functions evaluated along a path), and certain types of functional integrals.

Functional integrals are integrals in which the integrand is a functional defined on a space of functions, and the domain of integration is a continuous region of that space. The first functional integral was developed by Daniell, published in 1919 [68]. It is not necessarily a path integral, but can be one in specific instances. A more widely known functional integral is the Wiener integral, which is an abstract Lebesgue integral based on the Wiener measure [326, 327]. It was developed by Norbert Wiener and published in 1921 in order to provide a rigorous mathematical framework for the proof of existence of atoms by Albert Einstein, which was one of his achievements during his annus mirabilis in 1905 [81].

The Feynman path integral is sometimes considered to be a special type of functional integration, but sometimes also as a third type of path integral. The reason for this ambiguity is that while the path integrals in Feynman's formulation appear to be functional integrals, not all of them are all well-defined, and mathematical research is still ongoing to provide a rigorous theoretical framework for them. Interestingly, the well-defined functional integrals in the path integral formulation of quantum mechanics are Wiener integrals. In this light one can see this part of the Feynman path integral formulation as an application or special case of the Wiener integral. A visual overview of the various path integral types can be seen in Figure 2.1.

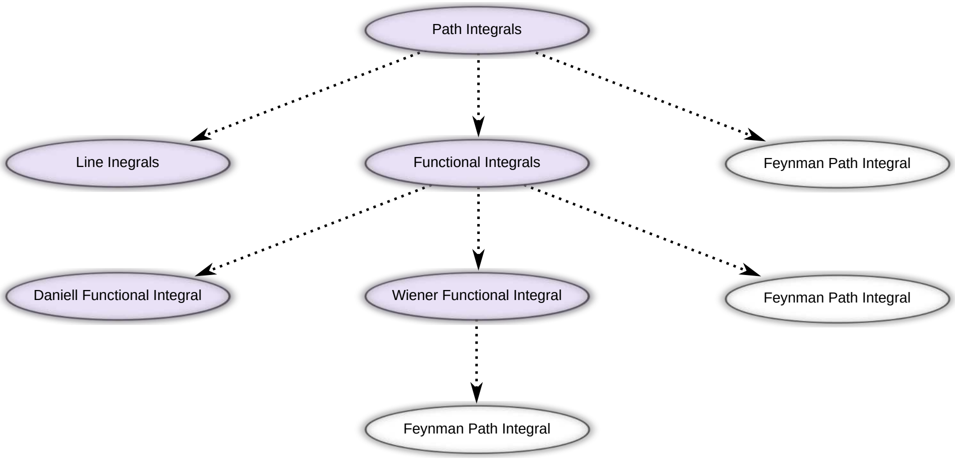


Figure 2.1: Overview of the various integrals which can be called *path integrals*. The Feynman path integral can be classified differently, and there seems to be no consensus in the literature. It can be seen as a separate type of path integral, as a special type of functional integration, and as a specific application of the Wiener functional integral. The ambiguity arises mainly due to the fact that the Feynman path integrals contains elements which are mathematically not well founded/defined.

Chapter Overview

In the next section the path integral formalism will be derived, assuming that the framework of quantum mechanics is already established via its axioms. After that we will in Section 2.3 turn to the relevant application of the path integral formalism in quantum statistical mechanics. These results will subsequentially be used in Section 2.4 which is dedicated to the PIMD technique.

2.2 Path Integral Formulation of Quantum Mechanics

The path integral formulation of quantum mechanics can serve as the starting point to arrive at quantum mechanics motivated by its close ties to classical mechanics, which was the approach by Feynman and Hibbs in [88, 250]. Alternatively it can be deduced by any of the other alternative formulations of quantum mechanics, which we will describe in this section.

There are different versions of the Feynman path integral formula, depending on whether

- the time is real or imaginary,
- the integral is over the phase or configuration space,
- the integral is over continuous or discretized paths.

The parts of quantum mechanics in which the time is imaginary is often also called *Euclidean quantum mechanics* (the reason is that in quantum field theory the Minkowski space is transformed into Euclidean space if the imaginary time is considered instead of the real time) [271]. The real time formulation leads to serious mathematical problems which are yet to be solved. For imaginary times these problems were remedied already in the middle of the last century by Mark Kac in 1949 [137]. While real time quantum mechanics is of central importance in physics, Euclidean quantum mechanics also has several critical applications. Moreover, it is possible to obtain results about real-time quantum mechanics from the corresponding imaginary representation via the path integral formalism and vice versa. It is also the imaginary time path integral which is relevant for this thesis due to its applications in quantum statistical mechanics. In this section we will at first turn to the Lie-Trotter-Kato product formulas, which serve as starting points for deriving both the real and the imaginary time versions of path integral formulation, to which we will come afterwards.

2.2.1 Fundamentals

In this subsection several concepts of mathematics and theoretical physics which are relevant in this thesis will be reviewed and the associated notation defined.

Unbounded Linear Operators

The theory of unbounded (as well as bounded) linear operators is central in the mathematical treatment of quantum mechanics, of which the most relevant classes will be defined here. Throughout this thesis the symbol $\mathcal{H}$ always refers to a Hilbert space which is separable (because only this type is relevant for us), but several concepts can principally also be defined for wider classes of spaces (such as Banach or certain topological spaces). The inner product of the Hilbert space will be denoted by $\langle \cdot, \cdot \rangle$. The norm $\| \cdot \|$ which is used in this chapter is always the one induced by the inner product of the corresponding Hilbert space unless stated otherwise.

Definition 2.2.1 (Unbounded Linear Operators)
Let $\mathcal{H}$ be a Hilbert space. Then a map $A : D(A) \to \mathcal{H}$ is called an unbounded linear operator on $\mathcal{H}$ if its domain $D(A)$ is a linear subspace of $\mathcal{H}$ and A is linear. The space of all unbounded linear operators on $\mathcal{H}$ is denoted by $\mathcal{U}(\mathcal{H})$.

It should be noted that here the concept of unbounded linear operators is a generalization of the bounded linear operators and thus include them as a special case (therefore an unbounded operator is possibly unbounded rather than necessarily).

One particularly important class of unbounded are the densely defined operators.

Definition 2.2.2 (Densely Defined Operators)
Let $\mathcal{H}$ be a Hilbert space, and let $A \in \mathcal{U}(\mathcal{H})$. Then A is said to be densely defined on $\mathcal{H}$ if its domain $D(A)$ is dense in $\mathcal{H}$ [28]. The space of all densely defined operators on $\mathcal{H}$ is denoted by $\mathcal{L}_{\mathrm{d}}(\mathcal{H})$.

An often useful tool when working with operators on a Hilbert space is the so-called *numerical range.*

Definition 2.2.3 (Numerical Range)
Let $\mathcal{H}$ be a Hilbert space, and let $A \in \mathcal{L}(\mathcal{H})$. Then the numerical range of A is defined by $\Theta(A) := \{\langle Ax, x\rangle | x \in D(A) \wedge \|x\| = 1\}$ [28].

A special type of densely defined operators are the *symmetric operators* [28, 281].

Definition 2.2.4 (Symmetric Operators)
Let $\mathcal{H}$ be a Hilbert space, and let $A \in \mathcal{L}_{\mathrm{d}}(\mathcal{H})$. Then A is said to be symmetric if

$$\langle Ax, x\rangle = \langle x, Ax\rangle \quad \forall x \in D(A). \tag{2.1}$$

In this case

(i) the lower bound of $\Theta(A)$ is defined by $\Theta_l(A) := \inf \Theta(A)$,

(ii) the upper bound of $\Theta(A)$ is defined by $\Theta_u(A) := \sup \Theta(A)$.

The space of all symmetric operators on $\mathcal{H}$ is denoted by $\mathcal{L}_{\mathrm{s}}(\mathcal{H})$.

If the Hilbert space is over $\mathbb{C}$, then A is symmetric if and only if $\Theta(A) \subseteq \mathbb{R}$ [281]. Symmetric operators are closable and again symmetric, i.e. the closure of the graph of these operators is always again the graph of linear operators in $\mathcal{L}_{\mathrm{s}}$.

The notion of symmetry is closely related to self-adjointness. For bounded operators they are equivalent, but for unbounded operators they are generally not the same.

Definition 2.2.5 (Self-adjoint Operators)
Let $\mathcal{H}$ be a Hilbert space, and let $A \in \mathcal{L}_{\mathrm{d}}(\mathcal{H})$. Then A is said to be self-adjoint if A is symmetric and $D(A) = D(A^*)$, where A^* is the adjoint of A [28]. The space of all self-adjoint operators on $\mathcal{H}$ is denoted by $\mathcal{L}_{\mathrm{sa}}(\mathcal{H})$.

A generalization of self-adjoint operator are essentially self-adjoint operators.

Definition 2.2.6 (Essentially Self-adjoint Operators)
Let $\mathcal{H}$ be a Hilbert space, and let $A \in \mathcal{L}_{\mathrm{d}}(\mathcal{H})$. Then A is said to be essentially self-adjoint if A is symmetric and if $\bar{A} \in \mathcal{L}_{\mathrm{sa}}$ [281]. The space of all essentially self-adjoint operators on $\mathcal{H}$ is denoted by $\mathcal{L}_{\mathrm{esa}}(\mathcal{H})$.

Another important special case of symmetric operators are positive operators.

Definition 2.2.7 (Positive Operators)
Let $\mathcal{H}$ be a Hilbert space, and let $A \in \mathcal{L}_{\mathrm{s}}(\mathcal{H})$. Then A is said to be

(i) positive (semidefinite) if $\Theta_l(A) \geq 0$, which is also denoted by $A \geq 0$,
(ii) positive definite if $\Theta_l(A) > 0$, which is also denoted by $A > 0$ [28, 237].

The space of all positive operators on $\mathcal{H}$ is denoted by $\mathcal{L}_{\mathrm{p}}(\mathcal{H})$, and the space of all positive definite operators by $\mathcal{L}_{\mathrm{pd}}(\mathcal{H})$.

Some authors use the term positive only for strictly positive operators.

An additional type of symmetric operators which play a role in quantum mechanics are semibounded operators [108].

Definition 2.2.8 (Semibounded Operators)
Let $\mathcal{H}$ be a Hilbert space, and let $A \in \mathcal{L}_{\mathrm{s}}(\mathcal{H})$. Then A is said to be semibounded if $\inf \Theta_l(A) > -\infty$. The space of all semibounded operators on $\mathcal{H}$ is denoted by $\mathcal{L}_{\mathrm{sb}}(\mathcal{H})$.

Semibounded operators are alternatively characterized by the condition that $\exists\, b \in \mathbb{R}$ such that

$$\langle Ax, x\rangle \geq b\|x\|^2 \quad \forall x \in D(A), \tag{2.2}$$

which is again equivalent to the condition that $\exists\, b \in \mathbb{R}$ such that $A + b\,\mathrm{I} \geq 0$. Semibounded operators are also called *bounded from below* [108], a property which also exists for bounded operators, but for them it is defined differently: $\exists\, b > 0$ such that

$$\langle Ax, x\rangle \geq b\|x\|^2 \quad \forall x \in D(A) = \mathcal{H}. \tag{2.3}$$

Therefore special care has to be taken, and we use the term *semibounded* for unbounded operators to distinguish it from the notion for bounded operators. For bounded operators boundedness from below also implies positivity, which is not generally the case for unbounded operators.

Quantum Mechanical Notation

In this thesis for abstract quantum mechanical object the Dirac notation, also called bra-ket notation, is used which was introduced by Paul Dirac in year 1939 [73].[4] In the case of two algebraic states belonging to some complex Hilbert space $\mathcal{H}$ normally in the form of $|\psi\rangle$ or $|\phi\rangle$, where each vector (in this notation) is called a *ket*. The linear functional which is defined by the ket $|\psi\rangle$ via

$$\langle\psi|(|\phi\rangle) := \langle\psi,\phi\rangle, \quad |\phi\rangle \in \mathcal{H}, \tag{2.4}$$

is on the other hand denoted by $\langle\psi|$ and called a *bra*. The inner product $\langle\psi,\phi\rangle$ is also written as $\langle\psi|\phi\rangle$ and is sometimes called a *bra-ket* or *bracket*. If an operator $\hat{A}$ is applied on the ket $|\phi\rangle$ and subsequentially the inner product taken with the bra $\langle\psi|$ this can, based on the above bra-ket notation, be denoted by $\langle\psi|(\hat{A}|\phi\rangle) = \langle\psi|\hat{A}|\phi\rangle$. Sometimes $\hat{A}|\phi\rangle$ is also denoted by $|\hat{A}\phi\rangle$. On the other hand also $\langle\psi|\hat{A}$ can be interpreted as linear functional due to the circumstance that each pair of one bra $\langle\psi|$ and one linear operator $\hat{A}$ on $D(\mathcal{H})$ give rise to a new linear functional on $D(\mathcal{H})$, a ket, which in the Dirac notation can be denoted by $\langle\psi|\hat{A}$, defined by

$$(\langle\psi|\hat{A})|\phi\rangle := \langle\psi|\hat{A}|\phi\rangle = \langle\psi|\hat{A}\phi\rangle, \quad \phi \in D(\hat{A}). \tag{2.5}$$

Another important concept in the Dirac notation are *ket-bras* of the form $|\psi\rangle\langle\phi|$ which are operators defined by

$$(|\psi\rangle\langle\phi|)|\chi\rangle := |\psi\rangle(\langle\phi|\chi\rangle), \quad |\chi\rangle \in \mathcal{H}, \tag{2.6}$$

where $|\psi\rangle, |\phi\rangle \in \mathcal{H}$.

If the states are represented in some coordinate representation as functions the Dirac notation is dropped and they are written as ordinary mathematical functions, e.g. $\psi = \psi(\mathbf{q})$ instead of $|\psi\rangle$, where $\mathbf{q}$ are some generalized coordinates. The situation is similar for quantum mechanical operators. As abstract quantities a hat is located above them, e.g. $\hat{H}$ for a quantum mechanical Hamiltonian, while in the coordinate representation these operators are written again as ordinary mathematical functions.

Lie-Trotter Product Formulas

The Lie-Trotter product formula plays a central role in the derivation of the path integral formulation both in real time and in imaginary time. It also has other applications, such as in the development of symplectic integrators for ordinary differential equations [120]. The first version of the formula goes back to Sophus Lie who has discovered it in the 19th century for finite-dimensional matrices [178]. It was subsequentially generalized to include certain classes of unbounded linear operators, most notably by Hale Trotter in 1959 and Tosio Kato in 1970s, but also several others [295, 141, 142]. Therefore the various versions are known under different, not clearly defined names, which include usually one or more of the contributors mentioned before.

We will introduce in the following mainly the versions which are of particular relevance to us.

[4] The Dirac presentation is for instance summarized in [108].

Theorem 2.2.9 (Lie-Trotter Product Formulas)
Let $\mathcal{H}$ be a Hilbert space over $\mathbb{C}$, let $t \geq 0$, and let A, $B \in \mathcal{L}_{\text{sa}}(\mathcal{H})$ such that $A + B \in \mathcal{L}_{\text{esa}}(D(A) \cap D(B))$. Then we have:

(i) The sequence of operators $(e^{\frac{it}{n}A}e^{\frac{it}{n}B})^n$ converges to $e^{it(A+B)}$ in the strong operator topology, i.e.

$$\lim_{n\to\infty} \left\| e^{it(A+B)}x - \left(e^{\frac{it}{n}A}e^{\frac{it}{n}B}\right)^n x \right\| = 0 \qquad (\forall x \in \mathcal{H}) \tag{2.7}$$

(ii) If A and B are also semibounded, then the sequence of operators $(e^{-\frac{t}{n}A}e^{-\frac{t}{n}B})^n$ converges to $e^{-t(A+B)}$ in the strong operator topology, i.e.

$$\lim_{n\to\infty} \left\| e^{-t(A+B)}x - \left(e^{-\frac{t}{n}A}e^{-\frac{t}{n}B}\right)^n x \right\| = 0 \qquad (\forall x \in \mathcal{H}) \tag{2.8}$$

The norms occuring in the above theorem is norm of the Hilbert space which is induced by its inner product. This and similar versions of the above theorem, as well as proofs, can for instance be found in [108, 101]. Another version which differs in the form of the splitting is the *symmetric Trotter formula*, also called *Strang splitting* [270], which states that under certain conditions we have

$$\left(e^{\frac{t}{2n}A}e^{\frac{t}{n}B}e^{\frac{t}{2n}A}\right)^n \xrightarrow{n} e^{t(A+B)} \tag{2.9}$$

for $t \geq 0$.

2.2.2 Path Integral Formulation in Real Time

In this subsection the path integral formulation in real time will be derived. For this purpose we will at first look more closely at the time evolution operator.

Time Evolution Operator

According to the axioms of quantum mechanics, every quantum system is completely defined by its Hamiltonian operator $\hat{H}$. This is also true for its evolution in time. By the spectral theory for bounded and unbounded operators, it can be shown that the solution of the following general initial value problem associated with the Schrödinger equation

$$|\psi(0)\rangle = |\psi_0\rangle, \tag{2.10}$$

$$i\hbar\frac{\partial}{\partial t}|\psi(t)\rangle = \hat{H}|\psi(t)\rangle, \quad t \geq 0, \tag{2.11}$$

is given by

$$|\psi(t)\rangle = e^{-\frac{i}{\hbar}\hat{H}t}|\psi_0\rangle, \quad t \geq 0. \tag{2.12}$$

Here $\hbar = h/2\pi$ is the reduced Planck constant, and t is the time.

The form of the solution motivates the definition of the real time evolution operator (RTEO).

Definition 2.2.10 (Real Time Evolution Operator)
Given a Hamiltonian operator $\hat{H}$ defined on the separable Hilbert space $\mathcal{H}$. Then we define the real time evolution operator associated with $\hat{H}$ by

$$\hat{U}(t) := e^{-\frac{i}{\hbar}\hat{H}t}, \quad t \geq 0. \tag{2.13}$$

It is explicitly mentioned that the time is real in this case to distinguish it from the imaginary time evolution operator (ITEO), which we will encounter later in this exposition.

We now assume that the Hamiltonian is of the form

$$\hat{H} = \hat{K} + \hat{V} \tag{2.14}$$

such that it is an element in $\mathcal{L}_{\text{esa}}(D(\hat{K}) \cap D(\hat{V}))$, where $\hat{K}$ is the canonical quantization of the classical kinetic energy given by

$$\hat{K} = K(\hat{\boldsymbol{p}}) = \frac{\hat{\boldsymbol{p}}^2}{2m}, \tag{2.15}$$

and $\hat{V} = V(\hat{\boldsymbol{x}})$ is a general potential only dependent on $\hat{\boldsymbol{x}}$. Here $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{p}}$ denote the d-dimensional position and momentum operators, respectively. Transforming the initial value problem given by (2.10)-(2.11) into the position representation leads in this case to

$$\psi(\boldsymbol{x}, 0) = \psi_0(\boldsymbol{x}) \tag{2.16}$$

$$i\hbar\frac{\partial}{\partial t}\psi(\boldsymbol{x}, t) = -\frac{\hbar^2}{2m}\nabla^2\psi(\boldsymbol{x}, t) + V(\boldsymbol{x})\psi(\boldsymbol{x}, t), \quad t \geq 0. \tag{2.17}$$

It can be shown that the above Cauchy problem has a weak solution (in the distributional sense) under relatively permissive conditions on the potential operator, such as $V \in \mathcal{L}^{\infty}_{loc}(\mathbb{R}^n)$ is bounded from below, where $\mathcal{L}^{\infty}_{loc}(\mathbb{R}^n)$ is the space of locally ∞-integrable functions defined on $\mathbb{R}^n$). This solution is also called a fundamental solution of the Cauchy problem, which we will denote by $U(\boldsymbol{x}, \boldsymbol{x'}, t)$. It satisfies the following conditions [281]:

$$U(\boldsymbol{x}, \boldsymbol{x'}, 0) = \delta(\boldsymbol{x} - \boldsymbol{x'}) \tag{2.18}$$

$$\psi(\boldsymbol{x'}, \boldsymbol{t}) = \int_{\mathbb{R}^n} U(\boldsymbol{x}, \boldsymbol{x'}, t)\psi_0(\boldsymbol{x})d\boldsymbol{x}, \quad t \geq 0. \tag{2.19}$$

It is no coincidence that we have chosen the symbol U for the fundamental solution, because it is the distributional kernel of the RTEO $\hat{U}(t)$. In other words, it is the RTEO in its position representation:

$$U(\boldsymbol{x}, \boldsymbol{x'}, t) = \langle\boldsymbol{x'}|\hat{U}(t)|\boldsymbol{x}\rangle, \quad t \geq 0. \tag{2.20}$$

It is also called the time propagator or the (complex) probability amplitude of the quantum system [281], and sometimes also the matrix representation of the RTEO (and the values of it its matrix elements) [297].

One of the big challenges in quantum mechanics is to determine the precise form of the propagator for specific quantum systems. And one of the major advantages of the Feynman path integral is its ability to provide such a form in a generic way [281].

Derivation

We will now derive the path integral formulation in real time in the case of a particle in one dimension, where the potential is of the form given in equation (2.14) above, i.e. $\hat{H} = \hat{K} + \hat{V}$ belonging to $\mathcal{L}_{\text{esa}}(D(\hat{K}) \cap D(\hat{V}))$, the potential $V(x)$ sufficiently well behaved, and the kinetic energy in its position representation given by

$$K = -\frac{\hbar^2}{2m}\nabla^2. \tag{2.21}$$

For the RTEO we have:

$$\begin{aligned}
\hat{U}(t) &= e^{-\frac{i}{\hbar}\hat{H}t} && (2.22)\\
&= e^{-\frac{i}{\hbar}(\hat{K}+\hat{V})t} && (2.23)\\
&= \lim_{P\to\infty}\left(e^{-\frac{i}{\hbar P}\hat{K}t}e^{-\frac{i}{\hbar P}\hat{V}t}\right)^P && (2.24)\\
&= \lim_{P\to\infty}\hat{\Omega}(t)^P, && (2.25)
\end{aligned}$$

where $\hat{\Omega}$ is defined by

$$\hat{\Omega}(t) := e^{-\frac{i}{\hbar P}\hat{K}t}e^{-\frac{i}{\hbar P}\hat{V}t}, \quad t \geq 0, \tag{2.26}$$

and equation (2.24) follows from the Lie-Trotter product formula (2.7). The quantity P occurring in the exponentials can be interpreted as partitioning the time into P smaller intervals of equal size, which motivates the definition

$$\Delta t := \frac{t}{P}. \tag{2.27}$$

We will try to find the position representation of a single $\hat{\Omega}$ operator. The diagonal elements of the distributional kernel of the operator

$$e^{-\frac{i}{\hbar}\hat{V}\Delta t} \tag{2.28}$$

are naturally given by

$$\langle x|e^{-\frac{i}{\hbar}\hat{V}\Delta t}|x\rangle = e^{-\frac{i}{\hbar}V(x)\Delta t} \tag{2.29}$$

since the position eigenstates $|x\rangle$ are the eigenstates of the this operator as well. For the other part of $\hat{\Omega}$ involving the momentum operator $\hat{K}$, i.e.

$$e^{-\frac{i}{\hbar}\hat{K}\Delta t}, \tag{2.30}$$

its position representation is given by

$$\begin{aligned}
\langle x'|e^{-\frac{i}{\hbar}\hat{K}\Delta t}|x\rangle &= \frac{1}{2\pi\hbar}\int_{\mathbb{R}} e^{\frac{i}{\hbar}(p(x'-x)-\frac{p^2}{2m}\Delta t)}dp && (2.31)\\
&= \sqrt{\frac{m}{2\pi i\hbar\Delta t}}e^{\frac{im}{2\hbar}\frac{(x'-x)^2}{\Delta t}}. && (2.32)
\end{aligned}$$

The first equation can be obtained by solving the TDSE for the free particle by the Fourier method (see for instance [281]), and the second part follows by evaluating the integral by using the Fresnel integral. Using the above expressions we obtain for the position representation of $\hat{\Omega}$:

$$\begin{aligned}
\Omega(x,x',t) &= \langle x'|\hat{\Omega}|x\rangle && (2.33)\\
&= \langle x'|e^{-\frac{i}{\hbar P}\hat{K}t}e^{-\frac{i}{\hbar}\hat{V}\Delta t}|x\rangle && (2.34)\\
&= e^{-\frac{i}{\hbar}V(x)\Delta t}\langle x'|e^{-\frac{i}{\hbar}\hat{K}\Delta t}|x\rangle && (2.35)\\
&= \sqrt{\frac{m}{2\pi i\hbar\Delta t}}e^{-\frac{i}{\hbar}V(x)\Delta t}e^{\frac{im}{2\hbar}\frac{(x'-x)^2}{\Delta t}} && (2.36)\\
&= \mathscr{N}_P e^{\frac{i}{\hbar}\left(\frac{m}{2}\frac{(x'-x)^2}{\Delta t}-V(x)\Delta t\right)}, && (2.37)
\end{aligned}$$

where the constant factor $\mathscr{N}_P$ is defined by

$$\mathscr{N}_p := \sqrt{\frac{m}{2\pi i\hbar\Delta t}} = \sqrt{\frac{mP}{2\pi i\hbar t}}, \quad P \in \mathbb{N}. \tag{2.38}$$

Now we will come back to the product of such operators, for which we define

$$\hat{\Omega}_P := \hat{\Omega}^P, \quad P \in \mathbb{N}. \tag{2.39}$$

For the position representation of $\hat{\Omega}_P$ we find that

$$\begin{aligned}
\Omega_P(x,x',t) &= \langle x'|\hat{\Omega}_P|x\rangle && (2.40)\\
&= \langle x'|\hat{\Omega}\hat{I}\hat{\Omega}...\hat{\Omega}\hat{I}\hat{\Omega}|x'\rangle && (2.41)\\
&= \langle x'|\hat{\Omega}\left(\int_{\mathbb{R}}|x\rangle\langle x|dx\right)\hat{\Omega}...\hat{\Omega}\left(\int_{\mathbb{R}}|x\rangle\langle x|dx\right)\hat{\Omega}|x'\rangle && (2.42)\\
&= \int ... \int_{\mathbb{R}^{P-1}} \langle x'|\hat{\Omega}|x_{P-1}\rangle ... \langle x_1|\hat{\Omega}|x\rangle dx_1...dx_{P-1} && (2.43)\\
&= \int ... \int_{\mathbb{R}^{P-1}} \prod_{k=0}^{P-1}\Omega(x_k,x_{k+1},t)\prod_{k=1}^{P-1}dx_k. && (2.44)
\end{aligned}$$

In equation (2.42) we have used the identity operator $\hat{I}$ in the form of

$$\hat{I} = \int_{\mathbb{R}}|x\rangle\langle x|dx \tag{2.45}$$

overall $P-1$ times. In the subsequent equations the integrals were rearranged and the coordinates relabeled such that $x_0 = x$ and $x_P = x'$. Expanding the distributional kernels of equation (2.44) we obtain

$$\begin{aligned}
\Omega_P(x,x',t) &= \langle x'|\hat{\Omega}_P|x\rangle && (2.46)\\
&= \mathscr{N}_P^P \int ... \int\limits_{\mathbb{R}^{P-1}} \prod_{k=0}^{P-1} e^{\frac{i}{\hbar}\left(\frac{m}{2}\frac{(x_{k+1}-x_k)^2}{\Delta t}-V(x_k)\Delta t\right)}\prod_{k=1}^{P-1}dx_k && (2.47)\\
&= \mathscr{N}_P^P \int ... \int\limits_{\mathbb{R}^{P-1}} e^{\frac{i}{\hbar}\sum\limits_{k=1}^{P-1}\left(\frac{m}{2}\frac{(x_{k+1}-x_k)^2}{\Delta t}-V(x_k)\Delta t\right)}\prod_{k=1}^{P-1}dx_k. && (2.48)
\end{aligned}$$

Finally coming back to the time evolution operator, we obtain for its distributional kernel:

$$U(x,x',t) = \lim_{P\to\infty} \Omega_P(x,x',t) \tag{2.49}$$

$$= \lim_{P\to\infty} \mathscr{N}_P^P \underset{\mathbb{R}^{P-1}}{\int\cdots\int} e^{\frac{i}{\hbar}\sum\limits_{k=0}^{P-1}\left(\frac{m}{2}\frac{(x_{k+1}-x_k)^2}{\Delta t}-V(x_k)\Delta t\right)} \prod_{k=1}^{P-1} dx_k. \tag{2.50}$$

Equation (2.50) is called the *discretized path integral formulation* of the real time propagator, which is mathematically well defined [281].

We will now try to take the limit in (2.50). The values of the x_k can be interpreted as the values of a path $x : [0,t] \to \mathbb{R}$ evaluated at the time points $s_k := k\Delta t$, implying that the difference between two consecutive time points is Δt. Therefore in the limit the term

$$\frac{m}{2}\frac{(x_{k+1}-x_k)^2}{\Delta t} = \frac{m\Delta t}{2}\left(\frac{x_{k+1}-x_k}{\Delta t}\right)^2 \tag{2.51}$$

which is present in the integrand turns into a derivative of the path $x = x(s)$. And the sum in the exponent, which can be seen as Riemann sums, turns into a Riemann integral, yielding

$$U(x,x',t) = \lim_{P\to\infty} \mathscr{N}_P^P \underset{\mathbb{R}^{P-1}}{\int\cdots\int} e^{\frac{i}{\hbar}\sum\limits_{k=1}^{P-1}\left(\frac{m\Delta t}{2}\left(\frac{x_{k+1}-x_k}{\Delta t}\right)^2-V(x_k)\Delta t\right)} \prod_{k=1}^{P-1} dx_k \tag{2.52}$$

$$= \mathscr{N}_\infty \int\limits_{\mathcal{P}(\mathbb{R})_{x,0}^{x',t}} e^{\frac{i}{\hbar}\int_0^t\left(\frac{m}{2}\left(\frac{dx}{ds}(s)\right)^2-V(x(s))\right)ds}\mathcal{D}x, \tag{2.53}$$

where

$$\mathscr{N}_\infty = \lim_{P\to\infty} \mathscr{N}_P^P, \quad \mathcal{D}x = \lim_{P\to\infty} \prod_{k=1}^{P-1} dx_k, \tag{2.54}$$

and the integral is taken over the path space

$$\mathcal{P}(\mathbb{R})_{x,0}^{x',t} = \left\{x : [0,t] \to \mathbb{R} \;\middle|\; x(0) = x, x(t) = x'\right\}. \tag{2.55}$$

Equation (2.53) is the well-known (continuous) *Feynman path integral formula* in the configuration space for the real time propagator (sometimes the normalization constant $\mathscr{N}_\infty$ is left out for reasons mentioned in the discussion below).

It is also possible to obtain a path integral formula in the phase space by not evaluating the momentum integrals in equation (2.31) but instead rearranging them in a similar fashion as with the configuration integrals [281]. This leads to the discretized path integral formula

$$U(x,x',t) = \tag{2.56}$$

$$\lim_{P\to\infty}\left(\frac{1}{2\pi\hbar}\right)^P \underset{\mathbb{R}^{2P-1}}{\int\cdots\int} e^{\frac{i}{\hbar}\sum\limits_{k=0}^{P-1}\left(p_k(x_{k+1}-x_k)^2-\left(\frac{p_k^2}{2m}+V(x_k)\right)\Delta t\right)} dp_0 \prod_{k=1}^{P-1} dp_k dx_k. \tag{2.57}$$

If the limit is taken a continuous version of the phase space formula can be obtained in a similar fashion as before for the configuration space version.

Interpretation

We now try to interpret the path integral formula in the configuration space by looking again at the derivation above. The splitting of the operator $e^{-\frac{i}{\hbar}(\hat{K}+\hat{V})t}$ into a product of P operators $\hat{\Omega}$ (via the Trotter formula) in equation (2.24) can be seen as the slicing of the time into P equal pieces of size Δt [297].Furthermore, the introduction of $P-1$ times the identity operator $\hat{I}$ expanded in the position representation in equation (2.42), together with the subsequent reordering of the integrals in equation (2.43), can be interpreted as measuring at each time slice the probability amplitude of the particle at each point in space for having been at each point in space of the last time slice (via the terms $\langle x_{k+1}|\hat{\Omega}|x_k\rangle$, which is the probability amplitude of a particle at point x_{k+1} after time Δt before having been localized at position x), and then integrating of the configuration space of each time slice to obtain the total probability amplitude [297]. This can also be seen as introducing a lattice at each intermediary timepoint with infinitely small cells through which the particle has to pass. This allows one to interpret the $P-1$ dimensional integral as an integral over all piecewise linear paths the particle can take between x and x', where each path is weighted by a complex value. Also, the sum in the exponent of the discretized path integral formula is a sum over each edge of the piecewise linear path. The time slicing procedure is illustrated visually in Figure 2.2. In the limit of $P \to \infty$ the number of time slices becomes continuous and the paths become smooth, as illustrated in Figure 2.3 on the next page [297].

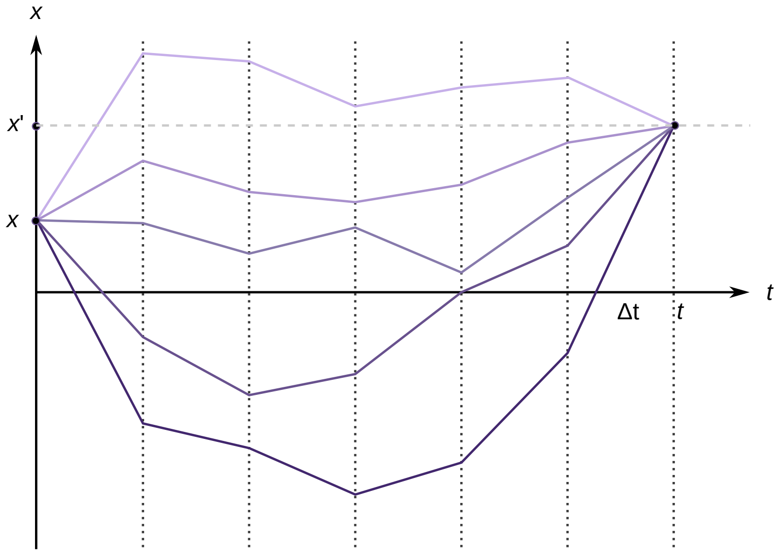


Figure 2.2: Time slicing procedure in the path integral formalism. The vertical lines (grey) represent the time slices, and the grating/openings within each time are infinitely small and have the cardinality of the continuum. Shown are several paths connecting the starting point x and the final point x'.

When looking at the continuous path integral formula given by equation (2.53), it seems clear why this representation of the propagator is called the path integral formulation: Because there is an integral over a set of paths. However, it is interesting to note that in the exponent there is a another integral occurring which can be interpreted as a line integral, which is a another type of path integral. But also the discretized path integral formulation deserves its name since there is an integral over the space of piecewise linear paths.

Connection to Classical Mechanics

In quantum mechanics particles generally spread out in space and are delocalized, and the probability amplitude alters in the entire space during the time evolution. It is therefore remarkable that one can express the delocalized time evolution as an integral over all possible paths the particles can take from x to x', and this circumstance already establishes a connection to classical mechanics since there particles also move along paths. However, the path integral formulation also shows a key difference between the two theories. While in classical mechanics particles can be interpreted as moving along a single well-defined trajectory, in quantum mechanics they move along all possible paths at the same time.

Moreover, the above connection is of a more deep nature as pointed out in [281]. In classical mechanics, the Lagrangian L of a mechanical system with kinetic energy K_{class} and potential energy V_{class} is given by

$$L(\boldsymbol{x}, \dot{\boldsymbol{x}}, t) = K_{\text{class}}(\dot{\boldsymbol{x}}) - V_{\text{class}}(\boldsymbol{x}, t) \tag{2.58}$$

$$= \frac{m\dot{\boldsymbol{x}}}{2} - V_{\text{class}}(\boldsymbol{x}, t). \tag{2.59}$$

And the classical action functional is defined by

$$S = S[\boldsymbol{x}(s)] = \int_{t_0}^{t_1} L(\boldsymbol{x}, \dot{\boldsymbol{x}}, s)ds, \tag{2.60}$$

where $\boldsymbol{x} = \boldsymbol{x}(s)$ is a classical trajectory parametrized by $s \in [t_0, t_1]$.

When now looking at the continuous Feynman path integral formula (2.53), one can see that in the exponent the integral is in fact over the classical Lagrangian, and that

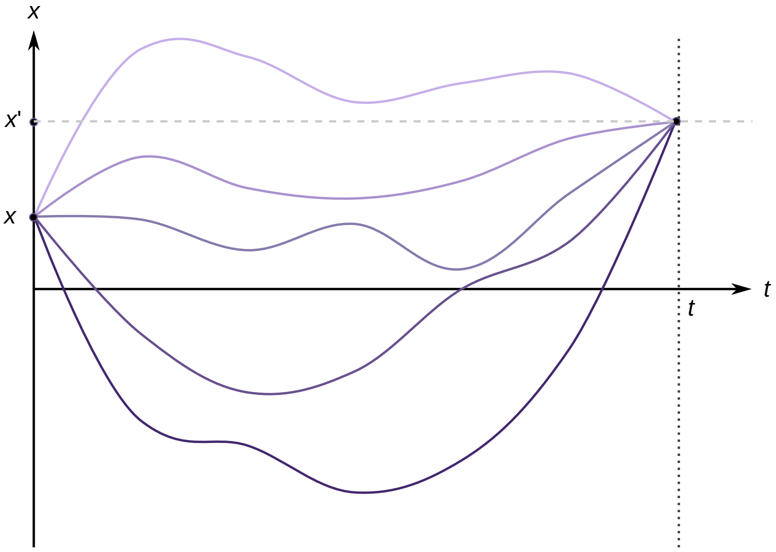


Figure 2.3: Continuous trajectories in the path integral representation connecting the starting point x and the final point x'.

the integral itself represents the action integral. Thus equation (2.53) can be written as

$$U(x,x',t) = \mathscr{N}_\infty \int\limits_{\mathcal{P}(\mathbb{R})_{x,0}^{x',t}} e^{\frac{i}{\hbar}S[x]}\mathcal{D}x. \tag{2.61}$$

This is an astonishing relation, which can be interpreted as each possible path between x and x' which contributes to the total probability amplitude is weighted by a complex number determined by term $e^{\frac{i}{\hbar}S[x]}$ which involves its classical action.

In classical mechanics the particles move along paths for which the classical action is stationary, i.e.

$$\frac{\delta S}{\delta x(t)} = 0, \tag{2.62}$$

where the derivative is the variational derivative. But in quantum mechanics, also paths are generally relevant at which the action is not stationary. However, if we consider the classical limit where $\hbar \to 0$, then with decreasing $\hbar$ the exponential weight starts to oscillate faster and faster when the action changes its value, and therefore these contributions starts to cancel each other out. In the limit only the weights of those paths are not canceling out for which the action is stationary, i.e. the classical path [281]. This consideration allows one to understand intuitively how clearly defined classical movements can arise from the principles of quantum mechanics.

Mathematical Problems

The Feynman path integral formula, while appearing intuitively logical, poses severe mathematical problems.

1. The sequence of normalization constants $\mathscr{N}_P^P$ is not converging, i.e. $\mathscr{N}_\infty$ is not well defined.

2. The sequence of product Lebesgue measures $\prod_{k=1}^{P-1} dx_k$ is not converging, i.e. $\mathcal{D}x$ is not well defined.

3. The action integral requires paths which have square-integrable derivatives, but the path space $\mathcal{P}(\mathbb{R})_{x,0}^{x',t}$ over which is integrated contains a much wider class of trajectories.

Therefore the continuous versions of the Feynman path integral formula (in real time) are not well-defined, and only the discretized versions have a mathematically rigorous meaning.

Multiparticle Systems in Higher Dimensions

When considering multiple particles the situation becomes more complex due to the possibility of entanglement of quantum particles. According to the standard model of particle physics two principal types of quantum particles exist, bosons and fermions. When multiple particles of the same type are being entangled they have different eigenstates. Bosons (e.g. photons) exhibit symmetric eigenstates, and follow Bose-Einstein statistics. Fermions on the other hand (e.g. electrons and protons) have antisymmetric eigenfunctions, and can be described by the Fermi-Dirac statistics. Antisymmetry also leads to the Pauli exclusion principle which prevents two fermions of being in the same state simultaneously. If quantum entanglement is not present or is neglected, then the particles are said to be non-interacting, and in this case the Boltzmann statistics is valid for them. For all three cases a path integral formulation can be derived.

We will summarize the for us relevant case of non-interacting particles. Given a system of N non-interacting particles in d-dimensional space, the discretized path integral formula in real-time is given by

$$U(x,x',t) = \tag{2.63}$$
$$\lim_{P\to\infty}\left(\prod_{i=1}^{N}\mathscr{N}_{P,i}^{dP}\right)\int_{\mathbb{R}^{Nd(P-1)}}\cdots\int e^{\frac{i}{\hbar}\sum\limits_{k=0}^{P-1}\left(\sum\limits_{i=1}^{N}\frac{m_i}{2}\frac{\left(\boldsymbol{r}_{k+1}^{(i)}-\boldsymbol{r}_{k}^{(i)}\right)^2}{\Delta t}-V\left(\boldsymbol{r}_k^{(1)},\dots,\boldsymbol{r}_k^{(N)}\right)\Delta t\right)}\prod_{i=1}^{N}\prod_{k=1}^{P-1}d\boldsymbol{r}_k^{(i)},$$

where $\boldsymbol{r}_k^{(i)}$ is the d-dimensional coordinate of particle i at intermediary time $s_k = k\Delta t$, and

$$\mathscr{N}_{P,i} = \sqrt{\frac{m_i}{2\pi i\hbar\Delta t}},\quad P\in\mathbb{N},\;\; i\in\{1,\dots,N\}. \tag{2.64}$$

The continuous version, again representing the limit of the corresponding discretized version (2.64) in a non-rigorous fashion, takes the shape of

$$U(x,x',t) = \mathscr{N}_\infty^{'}\int\limits_{\mathcal{P}(\mathbb{R}^d)_{\boldsymbol{r}^{(1)},\dots,\boldsymbol{r}^{(N)},0}^{\boldsymbol{r}'^{(1)},\dots,\boldsymbol{r}'^{(N)},t}} e^{\frac{i}{\hbar}\sum\limits_{i=1}^{N}\int_0^t\left(\frac{m_i}{2}\left(\dot{\boldsymbol{r}}^{(i)}\right)^2-V\left(\boldsymbol{r}^{(1)},\dots,\boldsymbol{r}^{(N)}\right)\right)ds}\prod_{i=1}^{N}\mathcal{D}\boldsymbol{r}^{(i)}, \tag{2.65}$$

where the normalization constant is formally given by

$$\mathscr{N}_\infty^{'} = \lim_{P\to\infty}\left(\prod_{i=1}^{N}\mathscr{N}_{P,i}^{dP}\right) \tag{2.66}$$

which again as in the one-dimensional case diverges and $\boldsymbol{r}^i,\; i\in\{1,\dots,N\}$, are the paths parametrized by $s\in[0,t]$.

2.2.3 Path Integral Formulation in Imaginary Time

Closely related to the RTEO is the ITEO (introduced on page 19), whose distributional kernel can be represented via the path integral formalism in a similar manner as the RTEO. One of the key advantages of the imaginary time variant (also called Euclidean path integrals) is that it leads to mathematically well defined continuous path integral formulas.

Imaginary Time Evolution Operator

The ITEO is defined by

$$\hat{\rho}(\beta) := e^{-\beta\hat{H}},\quad \beta\in[0,\infty). \tag{2.67}$$

It is also called the canonical density matrix (CDM) of the quantum system [297], because it is a generalization of the CDM of quantum statistical mechanics in the sense that here the parameter β can be freely chosen. In its field of origin the value is also called the *thermodynamic beta* or *inverse thermodynamic temperature*, and its value is given by

$$\beta = \frac{1}{k_B T}, \tag{2.68}$$

where T is the temperature of the system and k_B the Boltzmann constant. Here in the current context of pure quantum mechanics β is often simply referred to as the *imaginary time* [271] or the *inverse temperature* [297].

There exists a very close and direct relationship between the ITEO and the RTEO in the sense that both can be transformed into each other by the so-called (backward) *Wick rotation*.[5] The former is obtained by evaluating the latter at

$$t = -i\hbar\beta, \quad \beta \in [0, \infty), \tag{2.69}$$

which is seemingly an *imaginary time*, and hence the name ITEO of this operator. Similarly the RTEO can be obtained by the inverse of the above transformation, referred by us to as the *forward Wick rotation*, i.e. evaluating the CDM at

$$\beta = \frac{i}{\hbar} t, \quad t \in [0, \infty). \tag{2.70}$$

Both transformations are illustrated in Figure 2.4.

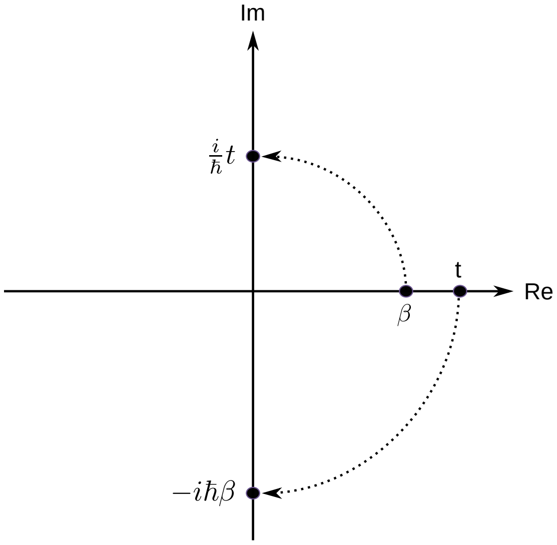


Figure 2.4: Transformation of real numbers into an imaginary numbers by the forward and backward Wick rotations, which are used to transform the imaginary time to the real time and vice versa within the time evolution operators.

[5] Some authors use the forward and the backward Wick transformations in the reverse direction.

Derivation

The path integral representation of the distributional kernel of the ITEO can be obtained in different ways. One of them is simply applying the Wick rotation on the corresponding path integral formula of the RTEO. Alternatively, one can derive it from the beginning analogously to the case of the RTEO which was shown above. A few modifications are required, such as using part (ii) of theorem 2.2.9 instead of part (i), or by using the version of the Strang splitting shown in equation (2.9).

Using the Strang splitting for a single particle in one dimension leads to the following representation of $\hat{\rho}$ in the position representation [297]:

$$\rho(x,x',\beta) = \lim_{P\to\infty} \mathscr{R}_P^P \underset{\mathbb{R}^{P-1}}{\int \cdots \int} e^{-\frac{1}{\hbar}\sum_{k=0}^{P-1}\left(\frac{m}{2\hbar}\frac{(x_{k+1}-x_k)^2}{\Delta\beta}+\frac{\hbar}{2}(V(x_{k+1})+V(x_k))\Delta\beta\right)} \prod_{k=1}^{P-1} dx_k, \tag{2.71}$$

where $\Delta\beta := \beta/P$, $x_0 = x$, $x_P = x'$, and

$$\mathscr{R}_p := \sqrt{\frac{m}{2\pi\hbar^2\Delta\beta}}. \tag{2.72}$$

Equation (2.71) is the *discretized path integral formula* of the ITEO in configuration space.

If $x = x'$, i.e. the starting point equals the end point, then equation (2.71) simplifies to

$$\rho(x,x,\beta) = \lim_{P\to\infty} \mathscr{R}_P^P \underset{\mathbb{R}^{P-1}}{\int \cdots \int} e^{-\frac{1}{\hbar}\sum_{k=0}^{P-1}\left(\frac{m}{2\hbar}\frac{(x_{k+1}-x_k)^2}{\Delta\beta}+\hbar V(x_k)\Delta\beta\right)} \prod_{k=1}^{P-1} dx_k, \tag{2.73}$$

These are the diagonal elements of $\hat{\rho}$, i.e. the amplitudes of the paths of a particle traveling in the imaginary time period $\beta = it/\hbar$ starting and ending in point x. These trajectories are also called *cyclic path*.

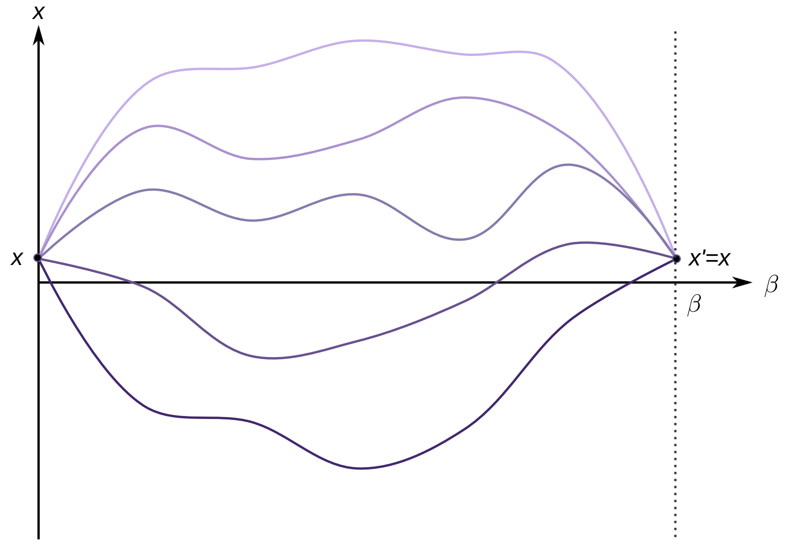


Figure 2.5: Cyclic paths in the path integral representation (in imaginary time), where the starting point x is the same as the end point x'.

The continuous path integral formulation we obtain again by taking the limit in the discretized version

$$\rho(x,x',\beta) = \lim_{P\to\infty} \mathscr{R}_\infty \int\limits_{\mathcal{P}(\mathbb{R})_{x,0}^{x',\beta}} e^{-\frac{1}{\hbar}\int\limits_0^{\hbar\beta}\left(\frac{m}{2}\left(\frac{dx}{ds}\right)^2+V(x)\right)ds}\mathcal{D}x, \tag{2.74}$$

where the space $\mathcal{P}(\mathbb{R})_{x,0}^{x',\beta}$ is defined according to equation (2.32), and

$$\mathscr{R}_\infty = \lim_{P\to\infty} \mathscr{R}_P. \tag{2.75}$$

In equation (2.74) the reduced Planck constant $\hbar$ is also present in the upper bound of the integral in the exponent, which follows by integration by substitution. In contrast to the corresponding version in real time equation (2.74) can be given a rigorous mathematical meaning. This was accomplished by Mark Kac in 1949 via the so-called *Feynman-Kac formula*. This theorem allows to identify a part of the above formula as the Wiener measure, which is a functional integral and a path integral as well. It follows that the Feynman path integral can be seen as

(1) containing path integrals of its own type (i.e. the not-well defined functional integrals of the real-time formulation),

(2) containing path integrals which are Wiener integrals,

(3) containing line integrals in the exponent, which represent a completely different class of path integrals.

Therefore, even though the term *Feynman path integral* usually refers to any of the functional integrals found in the path integral formulation of quantum mechanics, some of them can be identified with different type of functional path integral, one which is well defined.

Multiparticle System in Higher Dimensions

Regarding systems of N non-interacting particles in d dimensions, the situation is similar in the case of the RTEO described above on page 25. For such systems the *discretized path integral representation* of the distributional kernel of the ITEO is given by

$$\begin{aligned}\rho(\boldsymbol{r}^{(1)},...,\boldsymbol{r}^{(N)},\boldsymbol{r'}^{(1)},...,\boldsymbol{r'}^{(N)},\beta) &= \lim_{P\to\infty}\left(\prod_{i=1}^{N}\mathscr{R}_{P,i}^{dP}\right)\int\limits_{\mathbb{R}^{Nd(P-1)}}\cdots\int \\ &\times e^{-\sum\limits_{k=0}^{P-1}\left(\sum\limits_{i=1}^{N}\frac{m_i}{2\hbar^2}\frac{\left(\boldsymbol{r}_{k+1}^{(i)}-\boldsymbol{r}_k^{(i)}\right)^2}{\Delta\beta}+\frac{1}{2}\left(V\left(\boldsymbol{r}_{k+1}^{(1)},...,\boldsymbol{r}_{k+1}^{(N)}\right)+V\left(\boldsymbol{r}_k^{(1)},...,\boldsymbol{r}_k^{(N)}\right)\right)\Delta\beta\right)}\prod_{i=1}^{N}\prod_{k=1}^{P-1}d\boldsymbol{r}_k^{(i)},\end{aligned} \tag{2.76}$$

where × simply denotes scalar multiplication,

$$\mathscr{R}_{P,i} := \sqrt{\frac{m_i P}{2\pi\hbar^2\beta}}, \quad P\in\mathbb{N}, \ \ i\in\{1,...,N\}, \tag{2.77}$$

and $\boldsymbol{r}_0^{(i)} = \boldsymbol{r}^{(i)}$ as well as $\boldsymbol{r}_P^{(i)} = \boldsymbol{r'}^{(i)}$. For cyclic paths with in which $\boldsymbol{r}^{(i)} = \boldsymbol{r'}^{(i)}$ the formula simplifies again to

$$\rho(\boldsymbol{r}^{(1)}, ..., \boldsymbol{r}^{(N)}, \boldsymbol{r}^{(1)}, ..., \boldsymbol{r}^{(N)}, \beta) = \tag{2.78}$$

$$\lim_{P\to\infty} \left(\prod_{i=1}^{N} \mathscr{R}_{P,i}^{dP} \right) \int\limits_{\mathbb{R}^{Nd(P-1)}} e^{-\sum\limits_{k=0}^{P-1} \left(\sum\limits_{i=1}^{N} \frac{m_i}{2\hbar^2} \frac{\left(\boldsymbol{r}_{k+1}^{(i)} - \boldsymbol{r}_k^{(i)}\right)^2}{\Delta\beta} + V\left(\boldsymbol{r}_k^{(1)}, ..., \boldsymbol{r}_k^{(N)}\right) \Delta\beta \right)} \prod_{i=1}^{N} \prod_{k=1}^{P-1} d\boldsymbol{r}_k^{(i)}.$$

The corresponding continuous version is given by

$$\rho(x, x', \beta) = \mathscr{R}'_{\infty} \int\limits_{\mathcal{P}(\mathbb{R}^d)_{\boldsymbol{r}^{(1)}, ..., \boldsymbol{r}^{(N)}, 0}^{\boldsymbol{r'}^{(1)}, ..., \boldsymbol{r'}^{(N)}, \beta}} e^{-\frac{1}{\hbar} \sum\limits_{i=1}^{N} \int_0^{\hbar\beta} \left(\frac{m_i}{2} \left(\dot{\boldsymbol{r}}^{(i)}\right)^2 + V\left(\boldsymbol{r}^{(1)}, ..., \boldsymbol{r}^{(N)}\right)\right) ds} \prod_{i=1}^{N} \mathcal{D}\boldsymbol{r}^{(i)}, \tag{2.79}$$

where

$$\mathscr{R}'_{\infty} = \lim_{P\to\infty} \left(\prod_{i=1}^{N} \mathscr{R}_{P,i}^{dP} \right). \tag{2.80}$$

All of the path integral representations shown in this subsection are expressed as configuration space integrals, but as for the real time formulation one can also obtain phase space formulations for the imaginary time propagator.

2.3 Quantum Statistical Mechanics

In statistical mechanics, one of the most important ensembles is the canonical ensembles. And in the theory about this ensemble the CDM plays a central role, which is defined by

$$\hat{\rho}(N, T, V) = e^{-\beta \hat{H}}, \tag{2.81}$$

where β is again the *thermodynamic inverse temperature* given by $\beta = 1/k_B T$ (not to be confused with the more general β of the previous subsection), N is the number of particles, V is the volume of the system, and T the is the absolute temperature. The CDM occurs for instance in the canonical partition function (CPF) as well as in the normalized CDM (the quantum analog to the canonical phase space distribution function of classical statistical mechanics), and thus also in all ensemble averages. We will look at all of these objects, and will find that the path integral formulation of the CDM can in many cases be used to represent them in an alternative way. This alternative representation will subsequentially be useful in the PIMD methodology discussed in section 2.4 on page 34.

2.3.1 Canonical Partition Functions

The partition function plays a key role in statistical mechanics, and for a canonical ensemble with Hamiltonian $\hat{H}$ the CPF Q is given by

$$Q(N, V, T) = \mathrm{Tr}(\hat{\rho}(N, V, T)), \quad T \in [0, \infty). \tag{2.82}$$

For a single particle in a one-dimensional (i.e. $N = 1$, effectively reducing the number of variables of Q and ρ to only two) connected volume $\mathcal{V}$ which is bounded (e.g. an interval $[a, b] \subseteq \mathbb{R}$), the configuration space representation of the CPF becomes

$$Q(V,T) = \int\limits_{\mathcal{V}} \langle x|\hat{\rho}(V,T)|x\rangle dx. \tag{2.83}$$

$$= \int\limits_{\mathcal{V}} \rho(x,x,\beta)dx \tag{2.84}$$

$$= \int\limits_{\mathcal{V}} \lim_{P\to\infty} \mathscr{R}_P^P \underbrace{\int \cdots \int}_{\mathcal{V}^{P-1}} e^{-\frac{1}{\hbar}\sum\limits_{k=0}^{P-1}\left(\frac{m}{2\hbar}\frac{(x_{k+1}-x_k)^2}{\Delta\beta}+\hbar V(x_k)\Delta\beta\right)} \prod_{k=1}^{P-1} dx_k dx \tag{2.85}$$

$$= \lim_{P\to\infty} \mathscr{R}_P^P \underbrace{\int \cdots \int}_{\mathcal{V}^{P}} e^{-\frac{1}{\hbar}\sum\limits_{k=0}^{P-1}\left(\frac{m}{2\hbar}\frac{(x_{k+1}-x_k)^2}{\Delta\beta}+\hbar V(x_k)\Delta\beta\right)} \prod_{k=0}^{P-1} dx_k, \tag{2.86}$$

where in the last equation the coordinate x was relabeled to x_0 (which moreover equals x_P in this case since the paths are cyclic). Thus the CPF can be seen as the integral over all the diagonal elements of the CPF.

For systems consisting of N noninteracting particles in a d-dimensional volume $\mathcal{V}$ we obtain, analogously to the one-dimensional case, for the CPF in the position representation (by employing equation (2.78)):

$$\begin{aligned} Q(N,V,T) &= \int\limits_{\mathcal{V}} \rho(\boldsymbol{r}^{(1)},...,\boldsymbol{r}^{(N)},\boldsymbol{r}^{(1)},...,\boldsymbol{r}^{(N)},\beta) \prod_{i=1}^{N} d\boldsymbol{r}^{(i)} \\ &= \int\limits_{\mathcal{V}} \lim_{P\to\infty} \left(\prod_{i=1}^{N} \mathscr{R}_{P,i}^{dP}\right) \underbrace{\int \cdots \int}_{\mathcal{V}^{N(P-1)}} \\ &\quad \times e^{-\sum\limits_{k=0}^{P-1}\left(\sum\limits_{i=1}^{N}\frac{m_i}{2\hbar^2}\frac{\left(\boldsymbol{r}_{k+1}^{(i)}-\boldsymbol{r}_k^{(i)}\right)^2}{\Delta\beta}+V\left(\boldsymbol{r}_k^{(1)},...,\boldsymbol{r}_k^{(N)}\right)\Delta\beta\right)} \prod_{i=1}^{N}\prod_{k=1}^{P-1} d\boldsymbol{r}_k^{(i)} \prod_{i=1}^{N} d\boldsymbol{r}^{(i)} \end{aligned} \tag{2.87}$$

$$\begin{aligned} &= \lim_{P\to\infty} \left(\prod_{i=1}^{N} \mathscr{R}_{P,i}^{dP}\right) \underbrace{\int \cdots \int}_{\mathcal{V}^{NP}} \\ &\quad \times e^{-\sum\limits_{k=0}^{P-1}\left(\sum\limits_{i=1}^{N}\frac{m_i}{2\hbar^2}\frac{\left(\boldsymbol{r}_{k+1}^{(i)}-\boldsymbol{r}_k^{(i)}\right)^2}{\Delta\beta}+V\left(\boldsymbol{r}_k^{(1)},...,\boldsymbol{r}_k^{(N)}\right)\Delta\beta\right)} \prod_{i=1}^{N}\prod_{k=0}^{P-1} d\boldsymbol{r}_k^{(i)}, \end{aligned} \tag{2.88}$$

where $\boldsymbol{r}_0^{(i)} = \boldsymbol{r}_P^{(i)} = \boldsymbol{r}^{(i)}$ for $i \in \{1,...,N\}$. For practical purposes we now introduce the concept of the *finite CPF*, which is defined by

$$\begin{aligned} Q_P(N,V,T) &:= \left(\prod_{i=1}^{N} \mathscr{R}_{P,i}^{dP}\right) \underbrace{\int \cdots \int}_{\mathcal{V}^{NP}} \\ &\quad \times e^{-\sum\limits_{k=0}^{P-1}\left(\sum\limits_{i=1}^{N}\frac{m_i}{2\hbar^2}\frac{\left(\boldsymbol{r}_{k+1}^{(i)}-\boldsymbol{r}_k^{(i)}\right)^2}{\Delta\beta}+V\left(\boldsymbol{r}_k^{(1)},...,\boldsymbol{r}_k^{(N)}\right)\Delta\beta\right)} \prod_{i=1}^{N}\prod_{k=0}^{P-1} d\boldsymbol{r}_k^{(i)}, \end{aligned} \tag{2.89}$$

where $P \in \mathbb{N}$. It follows therefore by equation (2.86) that the CPF can also be written as

$$Q(N,V,T) = \lim_{P\to\infty} Q_P(N,V,T). \tag{2.90}$$

2.3.2 Expectation Values

The canonical density matrix $\hat{\rho}$ of quantum statistical mechanics corresponds, after normalization, to the *canonical phase space distribution function* of classical statistical mechanics. The *normalized CDM*, which we will denote by $\tilde{\rho}$, needs to have a trace of 1. Therefore it is defined by

$$\begin{align}\tilde{\rho}(N,V,T) &:= \frac{1}{\mathrm{Tr}(e^{-\beta\hat{H}})} e^{-\beta\hat{H}} \tag{2.91}\\ &= \frac{1}{Q(N,V,T)}\hat{\rho}(N,V,T) \tag{2.92}\end{align}$$

In other words, the normalization constant is the CPF encountered earlier, which is the motivation of its definition in the first place.

The normalized density matrix plays a central role when expressing expectation values of thermodynamic quantities. Let $\hat{A}$ be a quantum mechanical observable, then its expectation value is given by

$$\begin{align}\langle\hat{A}\rangle &= \mathrm{Tr}(\hat{A}\tilde{\rho}) \tag{2.93}\\ &= \frac{1}{Q(N,V,T)}\mathrm{Tr}(\hat{A}\hat{\rho}) \tag{2.94}\end{align}$$

If we consider now a single particle in one dimension, and if the operator observable can be expressed as a function of the position operator $\hat{x}$, i.e. $\hat{A} = A(\hat{x})$, then when expanding the expectation value in the position representation we obtain by using equation (2.73)

$$\begin{align}\langle\hat{A}\rangle &= \frac{1}{Q(V,T)}\mathrm{Tr}(\hat{A}\hat{\rho}) \tag{2.95}\\ &= \frac{1}{Q(V,T)}\int_{\mathcal{V}}\langle x|\hat{A}\hat{\rho}|x\rangle dx \tag{2.96}\\ &= \frac{1}{Q(V,T)}\int_{\mathcal{V}}\langle x|A(\hat{x})\hat{\rho}|x\rangle dx \tag{2.97}\\ &= \frac{1}{Q(V,T)}\int_{\mathcal{V}} A(x)\langle x|\hat{\rho}|x\rangle dx \tag{2.98}\\ &= \frac{1}{Q(V,T)}\lim_{P\to\infty}\mathscr{R}_P^P \underset{\mathcal{V}^P}{\int\cdots\int} A(x_0) e^{-\frac{1}{\hbar}\sum_{k=0}^{P-1}\left(\frac{m}{2\hbar}\frac{(x_{k+1}-x_k)^2}{\Delta\beta}+\hbar V(x_k)\Delta\beta\right)}\prod_{k=0}^{P-1} dx_k \tag{2.99}\end{align}$$

To obtain the last equation formula (2.73) was used, which is possible since $\langle x|\hat{\rho}|x\rangle = \rho(x,x,\beta)$ are the diagonal elements of the distributional kernel of the CDM $\hat{\rho}(\beta)$, and subsequentially the integrals were merged.

In equation (2.99) the function $A = A(x)$ is evaluated merely at the coordinate x_0, i.e. at one imaginary time point, and the other ones are neglected. However, if all the coordinates could be considered at the same time it would generally increase the convergence rate of the series (which is important in simulations). Fortunately, all the coordinates in this equation play an equivalent role and can therefore be exchanged with each other. If this is done in a cyclic fashion in the entire equation (2.99), and subsequentially only in the exponent (which is possible since the sum is over paths which are cyclic), equation (2.99) can also be rewritten more generally as

$$\langle \hat{A} \rangle = \frac{1}{Q(V,T)} \lim_{P\to\infty} \mathscr{R}_P^P \underset{\mathcal{V}^P}{\int \cdots \int} A(x_j) e^{-\frac{1}{\hbar}\sum\limits_{k=0}^{P-1}\left(\frac{m}{2\hbar}\frac{(x_{k+1}-x_k)^2}{\Delta\beta}+\hbar V(x_k)\Delta\beta\right)} \prod_{k=0}^{P-1} dx_k, \tag{2.100}$$

where $j \in \{1, ..., P\}$. When summing the right hand side of the above equation over all j and dividing it by P it follows that

$$\begin{aligned}\langle \hat{A} \rangle &= \frac{1}{P}\sum_{j=1}^{P} \frac{1}{Q(V,T)} \lim_{P\to\infty} \mathscr{R}_P^P \underset{\mathcal{V}^P}{\int \cdots \int} \\ &\quad \times A(x_j) e^{-\frac{1}{\hbar}\sum\limits_{k=0}^{P-1}\left(\frac{m}{2\hbar}\frac{(x_{k+1}-x_k)^2}{\Delta\beta}+\hbar V(x_k)\Delta\beta\right)} \prod_{k=0}^{P-1} dx_k & (2.101)\\ &= \frac{1}{Q(V,T)} \lim_{P\to\infty} \mathscr{R}_P^P \underset{\mathcal{V}^P}{\int \cdots \int} \\ &\quad \times \left(\frac{1}{P}\sum_{j=1}^{P} A(x_j)\right) e^{-\frac{1}{\hbar}\sum\limits_{k=0}^{P-1}\left(\frac{m}{2\hbar}\frac{(x_{k+1}-x_k)^2}{\Delta\beta}+\hbar V(x_k)\Delta\beta\right)} \prod_{k=0}^{P-1} dx_k. & (2.102)\end{aligned}$$

The average function

$$A_P(x_1, ..., x_P) := \frac{1}{P}\sum_{j=1}^{P} A(x_j) \tag{2.103}$$

is therefore an estimator (or microstate function) for the observable $\hat{A}$. Since by equation (2.90) we have $Q = \lim_{P\to\infty} Q_P$ equation (2.102) can be rewritten as

$$\begin{aligned}\langle \hat{A} \rangle &= \lim_{P\to\infty} \frac{1}{Q_P(V,T)} \mathscr{R}_P^P \underset{\mathcal{V}^P}{\int \cdots \int} \\ &\quad \times A_P(x_1, ..., x_P) e^{-\frac{1}{\hbar}\sum\limits_{k=0}^{P-1}\left(\frac{m}{2\hbar}\frac{(x_{k+1}-x_k)^2}{\Delta\beta}+\hbar V(x_k)\Delta\beta\right)} \prod_{k=0}^{P-1} dx_k. & (2.104)\end{aligned}$$

If we now define the function $g_P(x_1, ..., x_P)$ to be the probability density which occurs in the last equation, i.e.

$$g_P(x_1, ..., x_P) := \frac{1}{Q_P(V,T)} \mathscr{R}_P^P \, e^{-\frac{1}{\hbar}\sum\limits_{k=0}^{P-1}\left(\frac{m}{2\hbar}\frac{(x_{k+1}-x_k)^2}{\Delta\beta}+\hbar V(x_k)\Delta\beta\right)}, \tag{2.105}$$

then equation (2.104) can be written as

$$\langle \hat{A} \rangle = \lim_{P\to\infty} \underset{\mathcal{V}^P}{\int \cdots \int} A_P(x_1, ..., x_P) g_P(x_1, ..., x_P) \prod_{k=0}^{P-1} dx_k. \tag{2.106}$$

This can be seen as a limit of finite ensemble averages of the observable $\hat{A}$ via the estimator A_P (when the function g_P is interpreted as the phase space distribution function of the finite system), which is given by

$$\langle \hat{A} \rangle_{g_P} = \underset{\mathcal{V}^P}{\int \cdots \int} A_P(x_1, ..., x_P) g_P(x_1, ..., x_P) \prod_{k=0}^{P-1} dx_k. \tag{2.107}$$

Using this expression in equation (2.106) allows us to express $\langle \hat{A} \rangle$ compactly as

$$\langle \hat{A} \rangle = \lim_{P \to \infty} \langle A \rangle_{g_P}. \tag{2.108}$$

How such expectation values of systems modeled by the path integral formalism as above can be computed numerically will be the central topic of the next section.

2.4 Path Integral Molecular Dynamics

In the last section central quantities of statistical mechanics were expressed using the path integral formulation of quantum mechanics, such as the CDM, the CPF, and a large class of expectation values. And moreover, it was possible to represent them as limits of sequences which are indexed by the number of imaginary time points P. This seems to be particularly favorable for numerical simulations, since they can take into account only a finite number of degrees of freedom. The question on how such system can be simulated will be shown in this section. This will lead us to a technique generally referred to as PIMD.

2.4.1 Derivation

We start again with a single particle in a one-dimensional bounded and connected volume $\mathcal{V}$, and consider the following generic expression based on the path integral formalism used in the last section:

$$C \underset{\mathcal{V}^P}{\int \cdots \int} A_P(x_1, ..., x_P)\, \mathscr{R}_P^P \, e^{-\frac{1}{\hbar} \sum\limits_{k=0}^{P-1} \left(\frac{m}{2\hbar} \frac{(x_{k+1} - x_k)^2}{\Delta\beta} + \hbar V(x_k) \Delta\beta \right)} \prod_{k=0}^{P-1} dx_k, \tag{2.109}$$

where C is an arbitrary constant. The above formula contains as specific cases both

(1) the finite CPF (introduced in equation (2.89) for N-particle systems), by setting $C = 1$ and A_P to the identity function,

(2) the expectation value $\langle A \rangle_{g_P}$ of the microstate function A, by setting $C = 1/Q_P(V, P)$ (see equation 2.107).

At first we reformulate formula (2.109) by changing the prefactor $1/\hbar$ in the exponent with β, and define the constant γ_P by

$$\gamma_P := \frac{\sqrt{P}}{\beta\hbar}, \quad P \in \mathbb{N}. \tag{2.110}$$

This leads to the following form of the above expression:

$$C \int\limits_{\mathcal{V}^P} \cdots \int A_P(x_1, ..., x_P)\, \mathscr{R}_P^P\; e^{-\beta \sum\limits_{k=0}^{P-1} \left(\gamma_P^2 \frac{m}{2}(x_{k+1}-x_k)^2 + V(x_k)\frac{1}{P}\right)} \prod_{k=0}^{P-1} dx_k. \tag{2.111}$$

By transforming the constant $\mathscr{R}_P^P$ into a Gaussian integral over the momentum space variables $p_0, ..., p_{P-1}$, which is here done by letting

$$\tilde{m} = \frac{mP}{(2\hbar\pi)^2}, \tag{2.112}$$

one can obtain the representation

$$C \int\limits_{\mathbb{R}^{dP}} \int\limits_{\mathcal{V}^P} A_P(x_1, ..., x_P) e^{-\beta \sum\limits_{k=0}^{P-1} \left(\frac{p_k^2}{2\tilde{m}} + \gamma_P^2 \frac{m}{2}(x_{k+1}-x_k)^2 + V(x_k)\frac{1}{P}\right)} \prod_{k=0}^{P-1} dx_k dp_k. \tag{2.113}$$

Moreover, the constant $\tilde{m}$ can principally be freely chosen in the current context because it will not affect any thermodynamic equilibrium properties. It follows that the partition function of the system can be written as

$$Q_P(V,T) = \int\limits_{\mathbb{R}^{dP}} \int\limits_{\mathcal{V}^P} e^{-\beta \sum\limits_{k=0}^{P-1} \left(\frac{p_k^2}{2\tilde{m}} + \gamma_P^2 \frac{m}{2}(x_{k+1}-x_k)^2 + V(x_k)\frac{1}{P}\right)} \prod_{k=0}^{P-1} dx_k dp_k. \tag{2.114}$$

What is truly remarkable regarding this representation is that this is the CPF of a classical system, namely the one with the Hamiltonian

$$H_{\text{class}}(\boldsymbol{x}, \boldsymbol{p}) := \sum_{k=0}^{P-1} \left(\frac{p_k^2}{2\tilde{m}} + \gamma_P^2 \frac{m}{2}(x_{k+1}-x_k)^2 + V(x_k)\frac{1}{P}\right). \tag{2.115}$$

It belongs to a system of P particles, and the first term in the sum involving the momenta is the kinetic energy. The last term is the potential energy divided by the number of particles, and the second part is a quadratic potential between two particles with adjacent indices. In other words, the Hamiltonian is of a system resembling a circular chain consisting of P beads which are connected to each other by springs with spring constant $\gamma_P^2 m/2$. The particles are also called beads since the system has some similarity to necklaces with P beads. Therefore the rings are also called necklaces [296, 194], or alternatively also as ring polymers since they consists of multiple connected identical elements (which can be seen as monomers when the connections are included and evenly split among the beads).

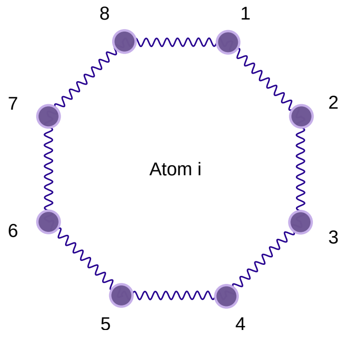


Figure 2.6: Path integral necklace of a single atom, for the special case where number of beads P is equal to 8. The beads having the numbers 1 to 8 which are present in the figure correspond to the possible values of the index $k \in \{1, ..., P\}$ in equation (2.114), where $k = 0$ and $k = P$ denotes the same bead.

For systems of N noninteracting particles in d dimensions it follows in a similar way that the finite CPF shown in (2.89) can be expressed as

$$Q_P(N,V,T) = \left(\prod_{i=1}^{N} \mathscr{R}_{P,i}^{dP}\right) \int\limits_{\mathbb{R}^{NdP}} \int\limits_{\mathcal{V}^{NP}} \tag{2.116}$$
$$\times\, e^{-\beta \sum\limits_{k=0}^{P-1} \left(\sum\limits_{i=1}^{N} \frac{\left(\boldsymbol{p}_k^{(i)}\right)^2}{2\tilde{m}_i} + \sum\limits_{i=1}^{N} \gamma_P^2 \frac{m_i}{2} \left(\boldsymbol{r}_{k+1}^{(i)} - \boldsymbol{r}_k^{(i)}\right)^2 + \frac{1}{P} V\left(\boldsymbol{r}_k^{(1)}, ..., \boldsymbol{r}_k^{(N)}\right) \right)} \prod_{i=1}^{N} \prod_{k=0}^{P-1} d\boldsymbol{r}_k^{(i)} d\boldsymbol{p}_k^{(i)},$$

where $\boldsymbol{r}_0^{(i)} = \boldsymbol{r}_P^{(i)}$ for $i \in \{1, ..., N\}$. What is most notable is that only beads of different atoms are interacting witch each other which have the same index, as illustrated in Figure 2.7.

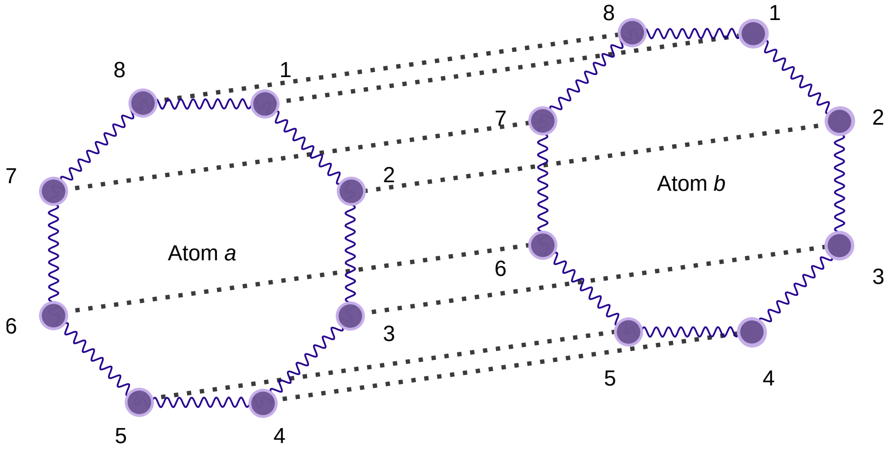


Figure 2.7: Path integral representation of two atoms for the special case where number of beads P is equal to 8.. Only beads of the same imaginary time are interacting with each other, represented by dotted lines. The atom names a and b correspond to different values of the variable $i \in \{1, ..., N\}$ in equation (2.116) which runs over all the atoms. The beads having the numbers 1 to 8 which are present in the figure correspond to the possible values of the index $k \in \{1, ..., P\}$ in equation (2.116), where $k = 0$ and $k = P$ denotes the same bead.

It follows that in the context of statistical mechanics, there exists a mapping between the systems of classical ring polymers and quantum mechanical systems modeled by the path integral formalism if the finite versions of their partition functions are used (cf. equation (2.89)). This mapping is also called the *classical isomorphism* [51], and it is of monumental importance for carrying out simulations of systems modeled by the path integral formalism.

From the viewpoint of the quantum mechanical path integral side of the classical isomorphism, which is given by equation (2.89), each bead represents the same atom, but at different points in imaginary time. For this reason the multiple beads per atom are sometimes also called replicas, and since they are present in different imaginary time points it is seems intuitively logical that only those beads are interacting which belong to the same imaginary time. However, it is remarkable that on the classical side of the isomorphism all of the imaginary time slices collapse to the same moment in real time. That the beads are connected by harmonic potentials corresponds to the fact that it is more likely to find a particle (in a thermodynamic system in equilibrium) after a certain (imaginary) time closer to its previous position than farther away.

We now come back to the question on how to simulate quantum mechanical systems represented by the finite CPF in the path integral formalism. According to the classical isomporhism, to the finite CPF path integral formalism (i.e. equation (2.89)) there corresponds a classical system with the same CPF. And in this section we have seen that expectation values of microstate functions which only depend on the configuration of the system are therefore identical for both the finite path integral and the corresponding classical system. This implies that for obtaining expectation values of canonical ensembles most of the machinery can be used which is available for classical systems. Very common are classical MD simulations which simulate the quantum mechanical path integral systems via the classical isomorphism, they are generally referred to as PIMDs. The equations of motion are readily derived from the Hamiltonian specified in equation 2.115.

2.4.2 Enhancements and Further Developments

The derivation above leads us to the naive form of the classical Hamiltonian corresponding to path integral systems. PIMD simulations which are based on this naive form often exhibit very slow convergence because of the significantly different time resolutions needed for the various degrees of freedom in such systems. A naive implementation would therefore need to use very small time steps to accommodate all of them appropriately, while only a few of them might really require it. Two basic methods which have been devised to facilitate this problem are based on normal mode transformations [66] or so-called staging variables [109].

Based on the fundamental path integral simulation techniques there have been developed a large set of additional tools and methods, and the theory has also been extended to allow for instance the simulation of certain time-dependent properties. Two common methods for computing for instance real-time quantum correlation functions are centroid molecular dynamics (CMD) [86, 97, 46] and ring polymer molecular dynamics (RPMD) [66].

Chapter 3

QM/MM Models

Philosophy is written in that great book which ever lies before our eyes — I mean the universe — but we cannot understand it if we do not first learn the language and grasp the symbols in which it is written.

Galileo Galilei
The Assayer

Contents

3.1 Introduction

Fundamental physical models of molecular systems can be broadly classified as either being of classical nature or of quantum mechanical (QM) nature. A broad class of classical models in which atoms are represented by single point particles are the so called molecular mechanics (MM) models. Both MM and QM models have their own advantages. One of the major strength of MM models is that they are computationally often much faster than QM models, while a major advantage of QM models is that they allow a more accurate description of molecular systems by explicitly considering their quantum nature to a certain extend.

The QM/MM methodology tries to combine the best of both worlds, MM and QM. It is not a fundamental model type, rather it is a hybrid model which arises when merging the other two models into a new combined method.

The QM/MM methodology can also be used in the newly developed FEM laid out in Chapter 4, and we have developed a new QM/MM scheme which is particularly suitable for our primary applications involving biomolecular systems.

Historical Notes

The QM/MM methodology was developed in the 1970s. In 1972 a predecessor of the method was published by Arieh Warshel and Martin Karplus for the simulation of conjugated molecules [318], and in 1976 Warshel and Michael Levitt demonstrated the full method by simulating an enzyme reaction [319].

While it might seem trivial to combine two existing methods into a new hybrid model, it might only do so in hindsight after its invention. To discover and develop a joint method from already existing individual models requires that one can see beyond them and perceive the bigger picture. And moreover it requires the development of the interlinkage part which allows to join the individual components seamlessly together into a single model in a way which is useful and practical. This is what Warshel, Karplus and Levitt have accomplished, and in 2013 the Nobel Prize in Physics was awarded to them *for the development of multiscale models for complex chemical systems* [284, 9].

The method was soon applied to many systems of interest, in particular biomolecular systems, and further development continued to evolve actively to this date.

Applications

The QM/MM methodology has a wide area of application. In general it can be used to study the dynamical behavior or thermodynamic properties of the system in question [219].

In computational physics (including computational chemical- and biophysics) it can for instance be used to study physical properties of solutes in condensed systems such as fluorescence [218, 129], or properties which involve more than one environment such as solvation free energies and effects [288, 225, 242, 279, 152, 287, 4] or distribution coefficients (such as the well widely used logP value between water and octanol) [153]. Is also an appropriate method for computing spectra of various types, such as vibrational IR and Raman spectra, as well as electronic absorption or emission spectra [21, 131].

In computational chemistry the simulation of reactions of organic and inorganic compounds is of central importance, and the QM/MM method is often quite suitable for them since it allows the formation and breaking of bonds [122, 123, 276, 23, 215]. Moreover, they have been used to study such subtle properties such kinetic isotope effects [222]. The computation of $\mathrm{pK_a}$ values has also been carried out [175].

The method is also used in computational biology when studying binding processes involving biological macromolecules such as proteins, RNA or DNA. Also enzyme reactions, and/or the effects of metals in proteins are often seen applications of QM/MM simulations in this field [225, 333, 95, 122, 204, 138].

In the field of computational pharmacy, in particular CADD, QM/MM can be of great value during the hit identification and hit/lead optimization steps when computing the free energy of binding by docking or simulation based approaches [52, 53, 234, 341, 316, 116, 59, 343, 151, 18]. But also during the target preparation step it can prove useful when refining or sampling the target structure [200].

In computational material science and nanotechnology it is for example sometimes used to study systems which involve metals because metals often exhibit profound quantum properties and it can be challenging, or even impossible, to model the desired properties of such systems with pure MM methods only [125, 102, 182].

Chapter Overview

In this chapter we will at first review the fundamental techniques of the QM/MM method (section 3.2). After that we will in 3.3 on page 46 come to advanced methods for diffusive systems in general. And finally the newly developed Quantum Adaptive Sphere Assembly Restraints (QUASAR) method, an extended QM/MM scheme for diffusive systems, will be described in 3.4 on page 50.

3.2 Fundamentals

When modeling a system by the QM/MM method, it is partitioned into different regions. In this text we define them as follows:

Definition 3.2.1 *Given a molecular system which is modeled by the QM/MM method, we define*

(i) the set $\mathcal{S}$ as the set of all atoms of the system,

(ii) the set $\mathcal{Q} := \{a \in \mathcal{S} \mid a \text{ is modeled on the QM level}\}$,

(iii) the set $\mathcal{M} := \{a \in \mathcal{M} \mid a \text{ is modeled on the MM level}\}$,

(iv) the set $\mathcal{B}$ as the set of all special atoms needed at the QM/MM boundary, which are either link atoms or dual role atoms (cf. 3.2.2),

(v) the set $\mathcal{A} := \mathcal{S} \cup \mathcal{B}$ representing the augmented system.

It follows that $\mathcal{S} = \mathcal{Q} \cup \mathcal{M}$, but this does not necessarily have to be a disjoint union because in some QM/MM schemes there can be atoms belonging to both regions. $\mathcal{B}$ does not need to contain atoms. Special boundary atoms are only needed when covalent bonds are cut[1] at the boundary, and this can in certain cases be avoided. The partitioning is illustrated in Figure 3.1 on the next page.

QM/MM methods can vary widely in how exactly they couple the MM and QM regions. These characteristics can also be used to QM/MM classify the methods in various ways, and the most important of them are listed below.

- The type of the **coupling quantity scheme.** The type of physical quantity which is coupled/mixed between the different subsystems to obtain a quantity for the entire system.
- The type of the **boundary atom scheme.** The way in which atoms connected by covalent bonds which are cut at the boundary are treated.
- The type of the **elecrostatic coupling scheme.** The way in which the MM and the QM system are coupled electrostatically.

The above properties are of a basic nature in the sense that all QM/MM methods possess them. There are optional features which these methods can have in addition, and these can be seen as advanced features. One particular example of such an advanced feature is the ability to model diffusive systems in a desirable way (see 3.3 on page 46).

[1] The bond is in not most cases not literally cut, even though it is commonly said that a covalent bond is cut if the QM/MM boundary is running across this bond.

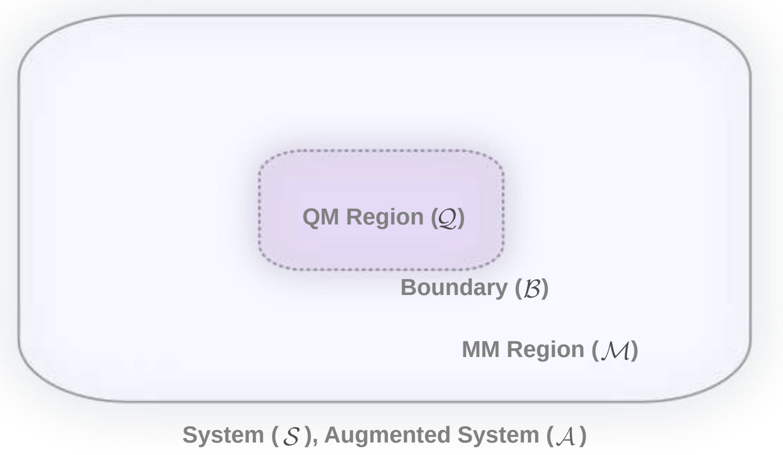

QM Region ($\mathcal{Q}$)
Boundary ($\mathcal{B}$)
MM Region ($\mathcal{M}$)
System ($\mathcal{S}$), Augmented System ($\mathcal{A}$)

Figure 3.1: Partitioning of the system into QM, MM, and boundary regions within QM/MM methods. The partitioning does not have to be disjoint, and the type of the boundary atoms depends on the boundary scheme.

3.2.1 Coupling Quantity Schemes

Each QM/MM method needs to have a central coupling quantity, a physical quantity which is computed for the entire system via its individual parts. There are two primary quantities which are normally coupled, the potential energy and the forces. In energy-coupling schemes it is possible to define a global potential energy function from which the forces follow naturally. In force-mixing schemes it is however not possible to formulate a global potential energy function (if it would the coupling could have been done via the potential energy). The circumstance that there is no total energy in force-mixing QM/MM methods renders it suitable only for applications which do not require the knowledge of this quantity. However, nearly all free energy simulation methods depend on it, including the newly developed method (QSTAR) presented later in this text (in section 4.3 on page 69).

Regarding energy-mixing schemes there exist two types, additive and subtractive schemes, and we will briefly look at both of them.

Additive Energy-Mixing Schemes

In additive energy coupling schemes the total potential energy of the entire (augmented) system $\mathcal{A}$ is defined to be the sum of the potential energies from the individual parts, i.e.

$$V^{\text{aug}}_{\text{QM/MM}}(\mathcal{Q},\mathcal{M},\mathcal{B}) := V_{\text{Q}}(\mathcal{Q}\cup\mathcal{B}) + V_{\text{M}}(\mathcal{M}) + V_{\text{M,Q}}(\mathcal{M},\mathcal{Q}), \tag{3.1}$$

where

(i) $V_{\text{Q}}(\mathcal{Y})$ is the QM potential energy of the atoms in the set $\mathcal{Y}$,

(ii) $V_{\text{M}}(\mathcal{Y})$ is the MM potential energy of the atoms in the set $\mathcal{Y}$,

(iii) $V_{\text{M,Q}}(\mathcal{Y}_1,\mathcal{Y}_2)$ is the QM/MM coupling energy between the regions $\mathcal{Y}_1$ and $\mathcal{Y}_2$, where the first argument is treated as the MM region and the second argument as the QM region.

If the set of special boundary atoms $\mathcal{B}$ is non-empty, then the total energy $V_{\mathrm{QM/MM}}(\mathcal{A})$ of the augmented system will contain the energy of more atoms than were present in the original system $\mathcal{S}$. While the special boundary atoms are needed to compute the energy of the quantum system $V_{\mathrm{Q}}(\mathcal{Q} \cup \mathcal{B})$, one can correct for their contribution on the MM level by subtracting the term

$$V_{\mathrm{corr}}(\mathcal{B}, \mathcal{Q}) := V_{\mathrm{M}}(\mathcal{B}) + V_{\mathrm{M,M}}(\mathcal{B}, \mathcal{Q}), \tag{3.2}$$

where the function $V_{\mathrm{M,M}}(\mathcal{V}_1, \mathcal{V}_2)$ is defined to be the coupling energy between the two regions $\mathcal{V}_1$ and $\mathcal{V}_2$ on the MM-level. This leads to the definition of

$$V_{\mathrm{QM/MM}}(\mathcal{Q}, \mathcal{M}, \mathcal{B}) := V_{\mathrm{Q}}(\mathcal{Q} \cup \mathcal{B}) + V_{\mathrm{M}}(\mathcal{M}) + V_{\mathrm{M,Q}}(\mathcal{M}, \mathcal{Q}) - V_{\mathrm{corr}}(\mathcal{B}, \mathcal{Q}). \tag{3.3}$$

However, in practice the term $V_{\mathrm{corr}}(\mathcal{B}, \mathcal{Q})$ is often neglected as $V_{\mathrm{M}}(\mathcal{B})$ is usually very small and $V_{\mathrm{M,M}}(\mathcal{B}, \mathcal{Q})$ nearly constant, and therefore often the uncorrected version $V^{\mathrm{aug}}_{\mathrm{QM/MM}}(\mathcal{Q}, \mathcal{M}, \mathcal{B})$ is used instead.

Subtractive Energy-Mixing Schemes

When the electrostatic interactions between the QM and MM region are modeled on the MM level (i.e. using the mechanical embedding), then there is an alternative to the additive scheme. Instead of adding the energies of the individual parts, one can compute the MM-energy of the entire non-augmented system $\mathcal{S}$, and then replace the MM-energy of the augmented QM region by its QM-energy:

$$V_{\mathrm{QM/MM}}(\mathcal{Q}, \mathcal{M}, \mathcal{B}) := V_{\mathrm{M}}(\mathcal{S}) - V_{\mathrm{M}}(\mathcal{Q} \cup \mathcal{B}) + V_{\mathrm{Q}}(\mathcal{Q} \cup \mathcal{B}). \tag{3.4}$$

For QM/MM schemes with mechanical embedding it can be shown that the additive scheme is equivalent to the subtractive scheme [253]. Using the expansions

$$V_{\mathrm{M}}(\mathcal{S}) = V_{\mathrm{M}}(\mathcal{Q}) + V_{\mathrm{M}}(\mathcal{M}) + V_{\mathrm{M,M}}(\mathcal{M}, \mathcal{Q}) \tag{3.5}$$

and

$$V_{\mathrm{M}}(\mathcal{Q} \cup \mathcal{B}) = V_{\mathrm{M}}(\mathcal{Q}) + V_{\mathrm{M}}(\mathcal{B}) + V_{\mathrm{M,M}}(\mathcal{Q}, \mathcal{B}) \tag{3.6}$$

we obtain when starting from the subtractive energy $V^{\mathrm{sub}}_{\mathrm{QM/MM}}$ as defined in equation (3.4):

$$\begin{aligned}
V^{\mathrm{sub}}_{\mathrm{QM/MM}}(\mathcal{Q}, \mathcal{M}, \mathcal{B}) &= V_{\mathrm{M}}(\mathcal{S}) - V_{\mathrm{M}}(\mathcal{Q} \cup \mathcal{B}) + V_{\mathrm{Q}}(\mathcal{Q} \cup \mathcal{B}) && (3.7)\\
&= V_{\mathrm{M}}(\mathcal{Q}) + V_{\mathrm{M}}(\mathcal{M}) + V_{\mathrm{M,M}}(\mathcal{M}, \mathcal{Q}) \\
&\quad - V_{\mathrm{M}}(\mathcal{Q}) - V_{\mathrm{M}}(\mathcal{B}) - V_{\mathrm{M,M}}(\mathcal{Q}, \mathcal{B}) + V_{\mathrm{Q}}(\mathcal{Q} \cup \mathcal{B}) && (3.8)\\
&= V_{\mathrm{M}}(\mathcal{M}) + V_{\mathrm{M,M}}(\mathcal{M}, \mathcal{Q}) - V_{\mathrm{M}}(\mathcal{B}) - V_{\mathrm{M,M}}(\mathcal{Q}, \mathcal{B}) \\
&\quad + V_{\mathrm{Q}}(\mathcal{Q} \cup \mathcal{B}) && (3.9)\\
&= V_{\mathrm{Q}}(\mathcal{Q} \cup \mathcal{B}) + V_{\mathrm{M}}(\mathcal{M}) + V_{\mathrm{M,M}}(\mathcal{M}, \mathcal{Q}) - V_{\mathrm{corr}}(\mathcal{B}, \mathcal{Q}) && (3.10)\\
&= V^{\mathrm{add}}_{\mathrm{QM/MM}}(\mathcal{Q}, \mathcal{M}, \mathcal{B}) && (3.11)
\end{aligned}$$

according to equation (3.1) since in the case of electrostatic embeddings we have

$$V_{\mathrm{M,M}}(\mathcal{M}, \mathcal{Q}) = V_{\mathrm{M,Q}}(\mathcal{M}, \mathcal{Q}). \tag{3.12}$$

It follows that substractive schemes are equivalent to additive schemes for systems with mechanical embedding, but additive schemes allow much more flexibility regarding their components, in particular in relation to the electrostatic coupling scheme. Yet there is a number of well-known QM/MM schemes available which are based on the subtractive scheme, e.g. the IMOMM method [195], the IMOMO method [311, 126], as well as the famous ONIOM method [276, 310] which have all been developed by Morokuma *et al.*

3.2.2 Boundary Atom Schemes

The boundary between the QM and the MM region does ideally not run through covalent bonds. Sometimes this can be avoided, but not always, for instance when the QM region contains a small region of biological macromolecule such as a protein. MM models do not have a big problem when some bonds are cut or removed. However, this is generally not true for the QM models, and special techniques need to be applied to treat the truncated bond in a suitable way.

Many techniques have been developed for this purpose, and they mostly fall into one the three following categories.

(1) Link-atom schemes.
(2) Dual-atom schemes.
(3) Frozen localized orbital schemes.

Generally only single bonds are cut to avoid further complications, which is usually possible since one has often some freedom to decide which bond exactly is to be cut. The nomenclature of atoms near the cut bond is illustrated in picture 3.2. The atoms are labeled according to their distance to the covalent bond which was cut.

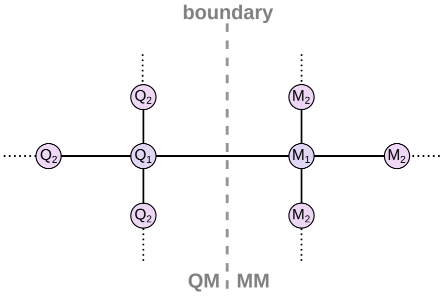


Figure 3.2: Naming conventions of atoms near the QM/MM boundary.

In link-atom schemes additional atoms are added to the system to engage the free valence electrons again which were created by the cutting of the bonds. The added link-atoms are usually hydrogen atoms, but also other types of atoms can be used as long as they are monovalent. In dual-atom schemes, no new atoms are added. Instead the M_1 atom is used in both the MM, as well as in the QM region in some modified way which allows to preserve the covalent bonds which runs through the boundary. Finally, one can also modify the first atom on the QM side (Q_1) by keeping the orbitals of the free valence electron constant, which is the key idea of the frozen localized orbital approach.

3.2.3 Electrostatic Schemes

The coupling term $V_{\mathrm{M,Q}}$ was introduced for additive schemes in equation 3.1 on page 42. For two regions $\mathcal{Y}_1$, $\mathcal{Y}_2$ it can usually be expressed as

$$V_{\mathrm{M,Q}}(\mathcal{Y}_1,\mathcal{Y}_2) = V^{\mathrm{bonded}}_{\mathrm{M,Q}}(\mathcal{Y}_1,\mathcal{Y}_2) + V^{\mathrm{vdw}}_{\mathrm{M,Q}}(\mathcal{Y}_1,\mathcal{Y}_2) + V^{\mathrm{es}}_{\mathrm{M,Q}}(\mathcal{Y}_1,\mathcal{Y}_2), \tag{3.13}$$

where

- $V^{\mathrm{bonded}}_{\mathrm{M,Q}}(\mathcal{Y}_1,\mathcal{Y}_2)$ accounts for the bonded interactions between region $\mathcal{Y}_1$ and $\mathcal{Y}_2$,
- $V^{\mathrm{vdw}}_{\mathrm{M,Q}}(\mathcal{Y}_1,\mathcal{Y}_2)$ represents all attractive and repulsive van der Waals forces between region $\mathcal{Y}_1$ and $\mathcal{Y}_2$,
- $V^{\mathrm{es}}_{\mathrm{M,Q}}(\mathcal{Y}_1,\mathcal{Y}_2)$ is the electrostatic energy between region $\mathcal{Y}_1$ and $\mathcal{Y}_2$.

The second argument in the above functions is always assumed to be the QM region, but how the interactions are computed depends on the precise scheme. The bonded interactions and the van der Waals forces are usually modeled simply on the MM-level by standard force fields such as CHARMM or AMBER. However, for the electrostatic interactions there is a vast variety of methods available.

Electrostatic schemes can be classified into three major categories, ordered below with increasing sophistication:

(1) Mechanical embeddings.
(2) Electrostatic embeddings.
(3) Polarized embeddings.

In mechanical embeddings, the electrostatic interactions between the MM region and the QM region are modeled solely on the MM-level. Among the advantages of this type are that it is computationally efficiency and relatively simple to implement. However, it has several drawbacks as well. To mention only two of them, the QM density is not polarized by the charges of its environment, which can have a major influence on it, and the absence of this polarization can lead to significant errors. Also, for the entire system MM point charges are needed while the QM region frequently contains special molecules (such as ligands) for which no high-quality MM parameters are available. This circumstance can even have been the primary reason for using QM or QM/MM methods rather than MM methods.

In electrostatic embeddings, the point charges of the atoms in the MM region are explicitly included in QM computations. This solves the two major advantages mentioned above for mechanical embeddings: No point charges for the QM atoms are required, and the QM density is polarized by the environment. One major disadvantage of this type of embedding is that at the boundary MM atoms which are very close to the QM density can easily overpolarize the latter, which is one of the most common problems. However, many techniques have been developed over the past decades to handle this problem. Another disadvantage is that it is computationally more expensive than mechanical embeddings.

Polarized embeddings represent the most realistic category because they allow not only the polarization of the QM density by the MM region, but also the polarization in the opposite direction. For this to be possible the MM region requires to be modeld by a polarizable force field, such as the CHARMM polarizable forcefield [177, 14, 305]. The two regions can then mutually polarize each other in a self-consistent manner. However, it is still often non-trivial to run such simulations, and moreover the self-consistent cycles can computationally be very demanding since not only one but multiple QM evaluations are needed per time step.

3.3 Methods for Diffusive Systems

The previous section dealt with basic properties of QM/MM methods, but they can have also additional features which go beyond the basic principles. In this text they are referred to as *advanced* QM/MM methods. A major class among them are the methods for diffusive systems. In diffusive systems, i.e. systems in which some or all of the molecules can move around, special additional techniques are usually required when the QM region is supposed to be localized while containing a part of the diffusive phase. Without a special treatment the molecules would be able to veer away from the primary QM region.

There are two primary subclasses of methods for diffusive systems:

(1) Adaptive methods.
(2) Restrained methods.

In adaptive systems, the regions $\mathcal{Q}$ and $\mathcal{M}$ (which were introduced in definition 3.2.1) change as the simulation evolves in time, i.e. these regions are adapted dynamically to the microstates of the system. In other words, as molecules diffuse away from/close to the central QM region, they are reclassified as either QM or MM atoms. An alternative to diffusive systems are restraining potentials, which prevent the initial QM molecules in the diffusive phase to wander off too far from the QM region.

In relation to methods for diffusive systems, the entire system $\mathcal{S}$ can generally be divided into several regions:

- The *core quantum region* $\mathcal{Q}_c$, which contains all atoms of the system which are permanently treated on the QM level.
- The *adaptive quantum region* $\mathcal{Q}_a(t)$, which contains all atoms of the system which are adaptively treated on the QM level at time t
- The *transition region* $\mathcal{T}(t)$, which contains all atoms of the system which have a partial QM and a partial MM character at time t
- The *classical region* $\mathcal{M}(t)$, which contains all atoms of the system which are treated on the MM level at time t.
- The *quantum region* $\mathcal{Q}(t)$, which contains all atoms of the system which are fully treated on the QM level at time t, i.e. $\mathcal{Q}(t) = \mathcal{Q}_c \cup \mathcal{Q}_a(t)$.

Not all methods make use of the transition region, in which case one can see it as being an empty, infinitely small surface.

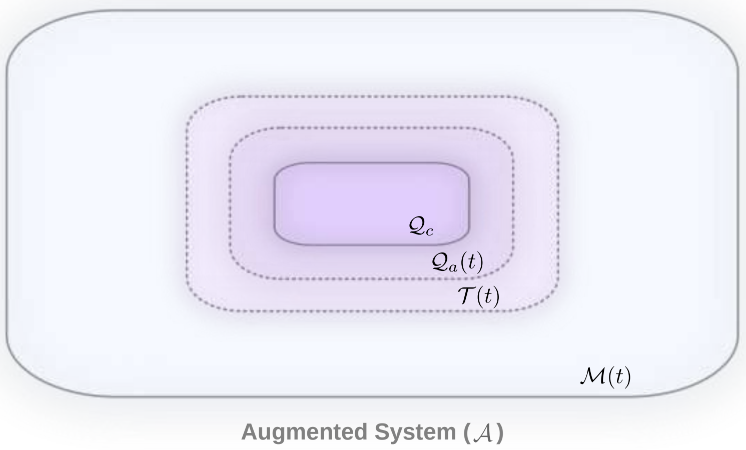


Figure 3.3: Partitioning of the system in diffusive QM/MM methods. The core region $\mathcal{Q}_c$ is time independent and contains all the atoms which are always treated on the QM level. The $\mathcal{Q}_a(t)$ region on the other hand contains the QM atoms which can pass from the QM region to the MM region and vice versa via the time-dependent transition region $\mathcal{T}(t)$.

3.3.1 Adaptive Methods

In adaptive methods, the dynamic repartitioning can be carried out in a number of ways. The by far most widely used approach is based on mere distance criteria, but also other methods exist such as number-based repartitioning or density-based repartitioning. In adaptive methods the transition region $\mathcal{T}$ allows to smooth the forces on the atoms when a molecule migrates from region to the other, which without smoothing can cause numerical problems and unnatural behavior of the system near the boundary.

Force-Mixing Methods

Many of the available adaptive schemes are force-mixing methods. One of their major advantages is that they allow to smooth the forces in the transition region without increasing the computational demand very much, often requiring at most one additional QM/MM energy evaluation per conformation. However, the major disadvantage is that either no potential energy is defined or it is not conserved, and thus they are not suitable for most FEM. Among the major methods which are available in this category are the following:

- **QMCF.** The *quantum mechanical charge field* method (QMCF) [239] is based on the *Hot-Spot* method (HS), which was the first adaptive method and has been developed in 1996 [143, 251].
- **ONIOM-XS.** The *ONIOM-XS* is an extension of the ONIOM method to diffusive systems, allowing the exchange of solvents (XS). It includes a transition region [144]. It conserves the energy only when a single molecule is in the transition region, which is rarely the case in practical applications. It was published in the year 2002.

- **Buffered Force.** The *Buffered Force* (BF) method was published by Bernstein in 2012 [24]. It avoids several problems at the QM/MM boundary by computing the forces of the QM region $\mathcal{Q}$ by including an additional buffer region (not corresponding to $\mathcal{T}$ in our notation).
- **NA.** The *Number-Adaptive* (NA) method was suggested in 2012 [280]. Instead of using only the distance as the central criterion for deciding which atoms have to be reclassified, it uses in addition also the number of molecules and keeps the constant in both the regions $\mathcal{Q}_a$ and $\mathcal{T}$.
- **ABRUPT.** The *ABRUPT* method is the one of the conceptually most simple methods, published in 2013. Atoms are simply reclassified instantly after leaving a rigid sphere, making the smoothing region non-existent and the transition *abrupt*. [44]
- **DBA.** In the *Density-Based Adaptive* (DBA) method of the year 2014, the dynamic partitioning of the system is based on the presence of non-covalent interactions between molecules in the regions $\mathcal{Q}$ and $\mathcal{M}$ [312, 355].
- **SCMP.** In the *Size-Consistent Multipartitioning* (SCMP) method (published 2014) the system is divided into a number of partitions of equal size, which can speed up the computations when parallelized. [320]
- **mPAP.** Also from the year 2014 is the *modified Permuted Adaptive Partitioning* (mPAP) method, which is based on the PAP method (which is an energy-mixing method, *vide infra*). [224]
- **TA.** The *Time-Adaptive* (TA) method of the year 2015 is unique in the sense that the smoothing of the forces of transiting atoms depends on the time rather than on the location. Unfortunately also here energy conservation is not possible for the physical system. [30]

Energy-Mixing Methods

The major advantage of energy-mixing methods is that the total energy of the system is conserved. But this advantage comes at a very high computational price, because in order to smooth the transition of molecules from one region to another, multiple QM/MM energy evaluations are required per conformation. Thus the major advantages and disadvantages of energy-mixing and force-mixing schemes are rather the inverse of each other. Among the energy-mixing schemes which have been developed are the following:

- **PAP.** The *Permuted Adaptive Partitioning* (PAP) method, published in 2007 [119], is possibly the most exact method which is available to date. But it is computationally also the most demanding method as it scales as 2^N, where $N = |\mathcal{T}|$. Based on the PAP method are the *Adaptive Partitioning Redistributed Charge* (APRC) and the *Adaptive Partitioning Redistributed Charge and Dipole* (APRCD) methods published in 2011 by Pezeshki. These allow that not only the entire molecule can be reclassified, but also smaller parts of it. Therefore individual molecules can principally be present simultaneously in all four types of regions ($\mathcal{Q}_c$, $\mathcal{Q}_a$, $\mathcal{T}$ and $\mathcal{M}$).
- **SAP.** The *Sorted Adaptive Partitioning* (SAP) method is a relative of the PAP method (published in the same article in 2007 [119]), but it is computationally

much cheaper. The number of QM/MM energy evaluations scales linearly with the number of molecules $N = |\mathcal{T}|$ in the transition region.

- **DAS.** The *Difference-based Adaptive Solvation* (DAS) method, which was published in 2009 [45], has some commonalities with the SAP method, and scales linearly as well. However, it uses a bookkeeping mechanism and is numerically more stable and efficient [356].

3.3.2 Restrained Methods

An alternative to reclassifying the atoms in diffusive systems when they migrate during the dynamics is to instead confine them to their respective regions in which they were initially located. Among the disadvantages of restraint methods is that the impermeable walls alter the motion of the particles significantly, and therefore they are not suitable for certain dynamical properties and analyses. However, the restraining walls have the advantage that they prevent a net flow of particles from one region to the other which can take place if the chemical potential in the MM and the QM region are different from each other. Naturally no smoothing mechanism is needed, since the particles cannot transit to the other region. There is only one method available so far in this categerory, viz. the Flexible Inner Region Ensemble Separator (FIRES) method [242] (published in 2012). In the FIRES method the regions $\mathcal{Q}$ and $\mathcal{M}$ are defined by spheres around a chosen central atom $a_c \in \mathcal{Q}$, and there is no transition region (i.e. $\mathcal{T} = \emptyset$). At any given moment in time the closest atom of the adaptive QM region $\mathcal{Q}_a = \mathcal{Q}$ defines the inner radius, and when an MM atom comes closer to the center of the sphere than the inner radius the restraining potential reflects the particle which has crossed the sphere (as well as its QM counterpart since the reflection has to happen on both sides of the boundary). The method is illustrated in Figure 3.4. An advantage of this method is that it is computationally cheap and that the total energy is defined.

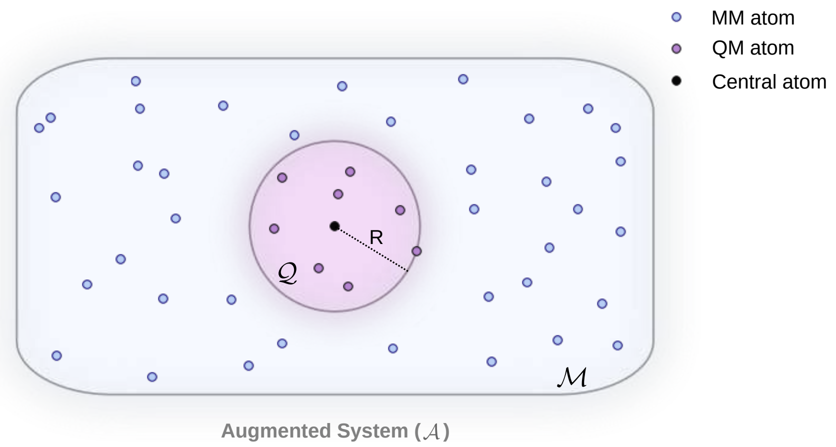


Figure 3.4: Illustration of the FIRES QM/MM scheme. The adaptive boundary region of the FIRES method is present in form of a spherical potential centered around a core atom (black). The QM atoms are non-adaptive, and the one with the farthest distance to the core atom defines the radius of the restraining potential.

3.4 The QUASAR Method

3.4.1 Introduction

Most free energy simulation methods (like the one developed later in this thesis) can principally be applied to systems which are modeled by the QM/MM method. However, only those QM/MM schemes are suitable in which the mixing quantity is the energy. While our new free energy method can be used for to a wide range of problems, for us the primary application area of interest are biomolecular simulations involving macromolecules such as proteins. For such systems it can in many cases be very favorable, and sometimes even necessary, to have the possibility to include a QM region of arbitrary shape, tailored to the specific case. However, most available QM/MM methods for diffusive systems which use energy-mixing (summarized in section 3.3 on page 46) employ single spheres centered around the central QM core region $\mathcal{Q}_c$ [356]. If for instance one would like to model a relatively shallow protein-protein interface (PPI) together with a small molecule (possibly a ligand/inhibitor) and very close solvent molecules on the QM level, a spherical QM region would not be suitable. A sphere centered at the middle of the PPI would result in a dome-like QM region as illustrated in figure 3.5, which would include many more layers of solvent molecules than necessary. Also, centering the sphere more deeply within the protein might only work in certain favorable cases since proteins vary greatly their shape, which can also lead to the sphere protruding at the other side of the protein. The additional solvent molecules in the QM region would in many cases lead to a great computational expense since QM methods have scaling behaviors as worse as $O(N_e^7)$ where N_e is the number of electrons (e.g. MP4). Therefore in QM/MM simulations it is usually of high priority to reduce the number of molecules in the QM region as much as possible (and reasonable) since this can speed up the sampling dramatically.

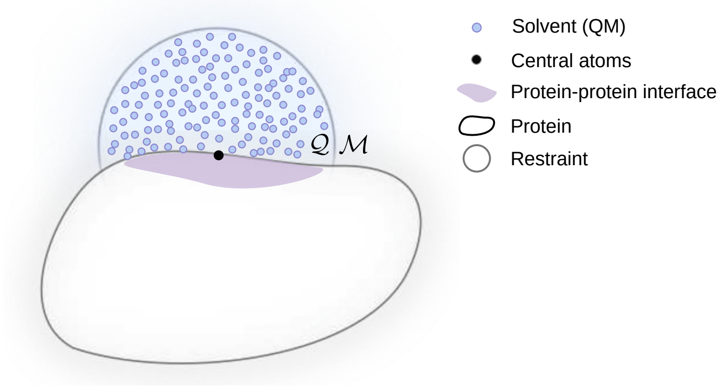


Figure 3.5: A spherical QM region is used in a system containing a protein-protein interface which should be modeled on the QM level together with nearby solvent molecules when using a QM/MM method. As can be seen in this example, when the QM region of the diffusive phase is defined by a single sphere, the resulting shape can be quite unfavorable.

The situation is even worse for instance when one intends to model a region of a DNA sequence on QM level along with the first two or three layers of surrounding solvent molecules (see Figure 3.6). In this light QM regions of arbitrary shape seem to

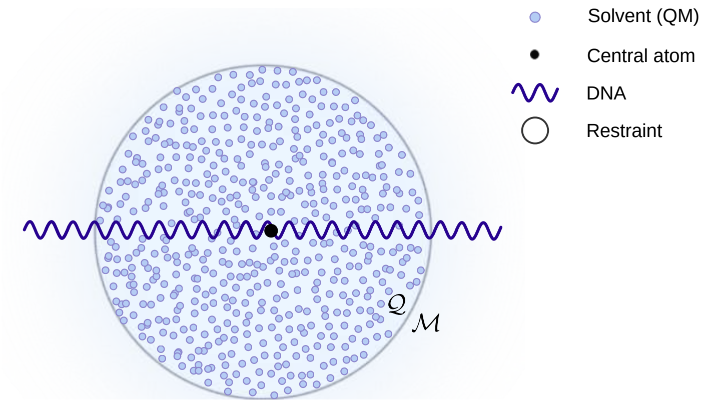


Figure 3.6: A system containing a solvated DNA segment is modeled by the QM/MM approach with the goal of modeling the DNA and nearby solvent molecules on the QM level. Using a single sphere would result in a QM region which is highly unfavorable due to the large amount of solvent molecules included which are not in direct proximity of the DNA segment.

be extremely favorable for certain biomolecular applications, but possibly also in other areas such as computational material science.

An ideal QM/MM method for our applications in the context of FEMs would have (to some extend due to the above considerations) the following properties:

(1) The mixing-quantity is the total potential energy.

(2) The shape of the QM/MM can be arbitrary.

(3) It is computationally cheap.

(4) The forces on all the atoms are continuous.

(5) No net flow of particles from one region to another due to a possible difference in the chemical potentials between the MM and the QM regions.

A QM/MM scheme all these properties was not available, and therefore we have designed a new scheme which meets all of them.

3.4.2 General Formulation

Regarding the class of the QM/MM scheme, a constrained scheme seems to be most favorable for our intended use cases since we do not require dynamical properties but only equilibrium properties.

Form of the Boundary

One of the key features which the new scheme is supposed to have is a QM region of arbitrary shape. For practical applications it is sufficient if the desired shape is sufficiently similar to one of the supported shapes. This could be done in theory by closed smooth surfaces, such as Bézier surfaces. Alternatively, one could allow all finite covers of spheres, which will be more convenient in most ways (in theory, implementation and application), and this approach was used for the new scheme.

At first we define the set of all finite covers of spheres by

$$\mathfrak{S} := \left\{ \{S(R_s, \boldsymbol{r}_s) \,|\, s \in \{1, ..., N_S\}\} \,\middle|\, R_s \geq 0, \boldsymbol{r}_s \in \mathbb{R}^3, s \in \{1, ..., N_S\}, N_S \in \mathbb{N} \right\}, \quad (3.14)$$

where $S(R, \boldsymbol{r})$ is the sphere of radius R and center $\boldsymbol{r}$, and N_S is the number of spheres in a specific cover.

In order to preserve the thermodynamic ensemble, the boundary needs to be flexible, determined by the location of the outermost and innermost particles in the regions. Therefore the spheres in the cover will not be fixed, but rather dynamically adjust to the configuration of the system, and therefore an adjustable cover can be written as

$$\mathscr{C}(t) = \{S_s(t) = S(R_s(t), \boldsymbol{r}_s(t)) \mid s \in \{1, ..., N_S\}\} \in \mathfrak{S}^{\mathbb{R}_0^+}, \; t \geq 0. \quad (3.15)$$

If the cover would consist only of a single sphere, then the boundary would be equal to the boundary of the FIRES method. If there is more than one sphere in the cover $\mathscr{C} \in \mathfrak{S}$ which determines the boundary region, the question arises if one particle should be associated to only one sphere or to multiple of them, since the association to a sphere center is needed to define the boundary (i.e the radii of the spheres in the cover). Another related question is if a particle should be confined to its initial sphere or not. Ideally, each particle should be free to move within the entire QM region, otherwise we would have multiple sub-ensembles within the QM region. To facilitate the free migration within the QM region at each time step each particle is reassigned to its closest sphere.

To formalize this principle we define for each sphere $S_s(t) \in \mathscr{C}(t)$ the set $\mathcal{S}_s(t) \subseteq \mathcal{Q}_a$ as the collection of all adaptive QM atoms which belong to the sphere S_s at the time t, where an atom belongs to a sphere when the sphere is the one with the closest center to the atom. Moreover, the time-dependent radius $R_s(t)$ of a sphere $S_s(t) = S(R_s(t), \boldsymbol{r}_s(t))$ is given by

$$R_s(t) := \max\left(\{0\} \cup \{R_a(s,t) \mid a \in \mathcal{S}_s(t)\}\right), \quad (3.16)$$

where $R_a(s,t)$ is the distance of atom a from the center of the sphere $S_s(t) = S(R_s(t), \boldsymbol{r}_s(t))$, i.e.

$$R_a(s,t) := \|\boldsymbol{r}_a(t) - \boldsymbol{r}_s(t)\|, \quad (3.17)$$

in which $\boldsymbol{r}_a(t)$ are the coordinates of the atom a. In equation (3.16) the number 0 is included in the maximum for the case that the sphere does not contain any atoms at some time point, in which the radius is set automatically to 0.

In the above way, if the spheres are sufficiently much overlapping, the particles can pass freely from sphere to sphere as they are reassigned to the neighboring sphere within the overlap region (see also Figure 3.7).

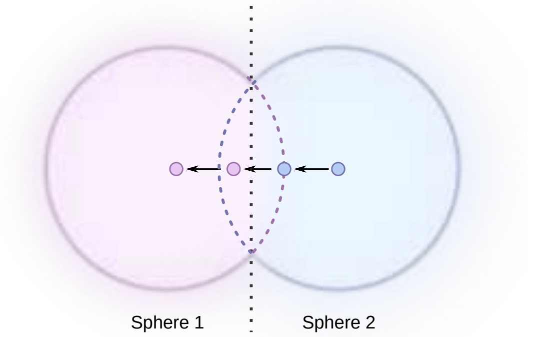


Figure 3.7: Transition of a particle between different spheres in the QUASAR method. At each time point each particle is assigned to the closest sphere, effectively resulting in a transition plane between the centers of overlapping restraining spheres. Shown in the figure is a single particle which is moving from sphere 2 to sphere 1 (having the same color of the sphere to which it is assigned to).

If it should happen that a molecule is partially assigned to one sphere and partially to another sphere while being in the middle of both, this should not lead to any problems. One reason is that there is normally no restraining force acting on the atoms of the molecule during the transition, as each atom is always assigned to its nearest sphere during each time step. However, it can happen that one of the two spheres of the transition exert restraining forces on the transiting molecule if the overlap of the two sphere is not large enough at that region. (For the unexpected case that this should cause problems one can also use the option to treat the molecules as single particles, as described below).

One can show that the separation of the system by the QUASAR method via an adjustable cover of spheres $\mathscr{C}(t) \in \mathfrak{S}^{\mathbb{R}_0^+}$) preserves the ensemble in many cases, i.e. that the thermodynamic equilibrium properties or the partition function do not change[2]. We assume for simplicity that we have only one type of solvent molecule (e.g. water) in the system consisting of M solvent molecules within a canonical ensemble of volume $\mathcal{V}$, and we treat each molecule as a single particle. At first we define the function $\mathcal{V}_{\mathcal{Q}}(\boldsymbol{r}_Q(t))$ to be the volume of the QM region $\mathcal{Q}$, i.e.

$$\mathcal{V}_{\mathcal{Q}}(\boldsymbol{r}_Q(t)) := \bigcup_{S \in \mathscr{C}} S(t), \quad t \geq 0 \tag{3.18}$$

$$= \bigcup_{s=1}^{N_S} S_s(R_s(t), \boldsymbol{r}_s(t)), \tag{3.19}$$

where

$$\boldsymbol{r}_Q(t) = \big(\boldsymbol{r}_1(t), ..., \boldsymbol{r}_{M_Q}(t)\big) \tag{3.20}$$

are the coordinates of the molecules in the QM region and $M_Q = |\mathcal{Q}|$ is the number of molecules in the QM region. Closely related we define $\mathcal{V}_{\mathcal{M}} \subseteq \mathcal{V}$ to be the volume of the MM region by

$$\mathcal{V}_{\mathcal{M}}(t) := \mathcal{V}(t)/\mathcal{V}_{\mathcal{Q}}(t), \quad t \geq 0. \tag{3.21}$$

[2] The situation is the same for the FIRES method, and the idea of derivation shown here is based on the one shown in [242].

It follows that $\mathcal{V}_Q \cup \mathcal{V}_\mathcal{M} = \mathcal{V} \quad \forall t \geq 0$.

The classical configurational CPF Z of the adaptive phase of the original unpartitioned system is given by

$$Z(M,V,T) = \frac{1}{M!} \int\limits_{\mathcal{V}} e^{-\beta H} \prod_{i=1}^{M} d\boldsymbol{r}_i, \tag{3.22}$$

where M is the normalization factor accounting for the indistinguishability of the M solvent molecules of the adaptive phase. Since all the molecules are of the same type they are indistinguishable, one can at first integrate over the M_Q coordinates within the QM region, and only afterwards over the remaining coordinates in the MM region (since one can exchange the identical particles). If this is done the factorial constant needs to be adjusted, leading to

$$Z(M,V,T) = \frac{1}{M_M!} \int\limits_{\mathcal{V}_\mathcal{M}(\boldsymbol{r}_Q)} \int\limits_{\mathcal{V}} \frac{1}{M_Q!} e^{-\beta H} \prod_{i=1}^{M_Q} d\boldsymbol{r}_i \prod_{j=M_Q+1}^{M_M} d\boldsymbol{r}_j. \tag{3.23}$$

Thus it is possible to consider the inner subsystem as being impermeable, and carry out the simulations accordingly by confining the QM atoms in the volume $\mathcal{V}_Q(\boldsymbol{r}_Q)$.

Restraining Potential

If a particle crosses the boundary of its current sphere, it should be reflected from the wall. This can be effectively implemented by using a strong restraining potential which imitates the reflection. The direction of the force of the restraining potential has to be parallel to the normal vector at the point of the sphere where a particle tries to cross it.

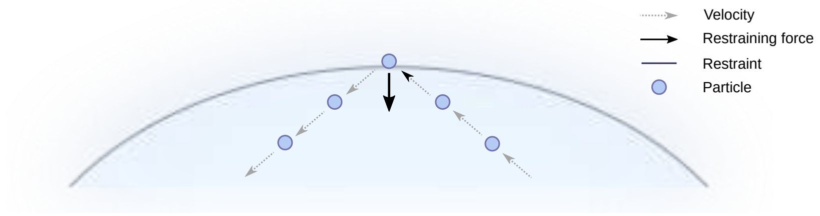


Figure 3.8: Restraining force at the QM/MM boundary pointing in the normal direction of the restraining sphere.

If individual atoms of molecules would try to cross the boundary this could in a few cases lead to convergence problems in the QM region if not all atoms of the molecule are experiencing the relatively strong restraining force, since the geometry of the molecule might be disturbed for a short time (and because a few QM methods can be sensitive to such distortions). To prevent such problems, a variant of the method was devised in which the constraining potential treats the molecules as a single particle, and therefore all atoms of the molecule will experience the same force. We refer to it as the Uniform Force Distribution (UFD) variant. This mechanism also serves as an additional protection while molecules which are currently in the process of transiting from one sphere to another sphere should experience non-uniform restraining forces (which only happens if the overlap of the regions is not large enough at that location).

We will now define the basic restraining potential of the QUASAR method for a given system with cover

$$\mathscr{C}(t) = \{S_s(t) = S(R_s(t), \boldsymbol{r}_s(t)) \mid s \in \{1, ..., N_S\}\} \tag{3.24}$$

at timepoint t by

$$V_{\mathrm{QUASAR}}(t) := \frac{1}{2}\,\omega \sum_{s=1}^{N_S} \sum_{a \in \mathcal{M}} \big(\min\{0, R_a(s,t) - R_s(t)\}\big)^2, \tag{3.25}$$

where $\omega > 0$ is the force constant. In the case that the UFD variant is used, the potential looks similar, but instead of atoms the center of masses of the molecules have to be used.

3.4.3 Applications

The QUASAR method has many potential applications. We start with the case mentioned in the introduction where the QM region is supposed to contain a PPI, a potential inhibitor, and a few layers of solvent molecules. With the QUASAR method one can for instance assign a constraining sphere to each atom of the PPI, place all the atoms of the PPI which are supposed to be in the QM region in the QM core region $\mathcal{Q}_c$ as well as the inhibitor, and add as many solvent molecules as desired into the adaptive QM region $\mathcal{Q}_a$ (see also Figure 3.9).

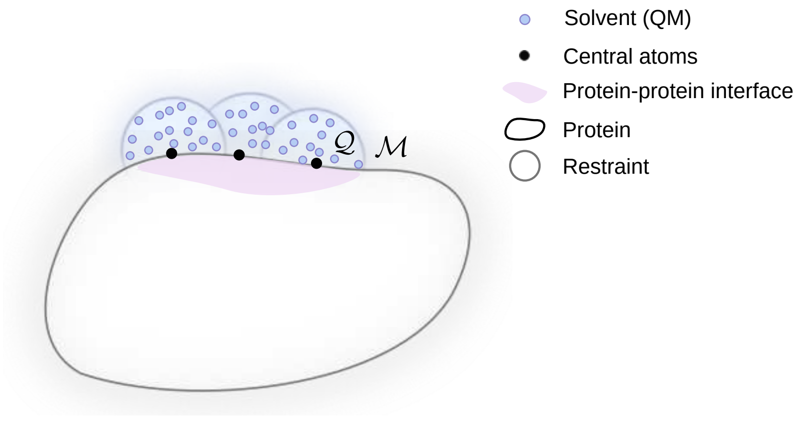


Figure 3.9: The QUASAR method applied to a system containing a protein-protein interface. QUASAR spheres can be placed throughout the interface to completely cover it, allowing to include a few layers of solvent on top of it.

Regarding segments of DNA or RNA, the QUASAR method seems to be nearly ideal, as one can place constraining spheres along the sequence, and solvent molecules in the adaptive QM region. Since the cover of spheres is adaptive, the interaction of DNA/RNA with itself is still possible as spheres from different sites can merge dynamically (see Figure 3.10).

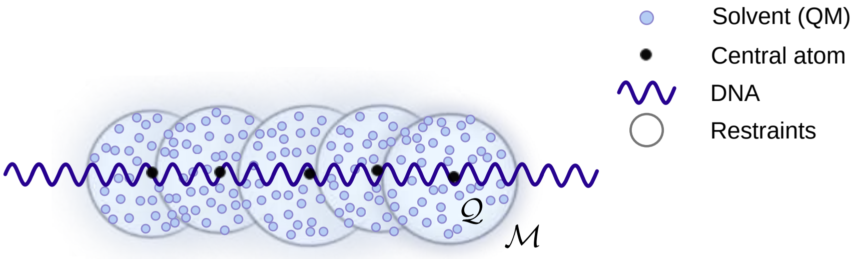


Figure 3.10: The QUASAR method applied to a system containing a DNA fragment. QUASAR spheres can be placed along the DNA strand, allowing to form a joint elongated, adaptive QM region which only includes as many solvent molecules as desired.

The adaptive nature of the cover and its re-merging capabilities render the method also suitable for modeling a number of small solutes in a solvent. For instance an amino acid or salt ions could be dissolved in water, and each of them placed in a single QM sphere (as illustrated in Figure 3.11). As two of these atoms/molecules approach each other, their water spheres will merge and allow for an interaction between the two objects on the QM level.

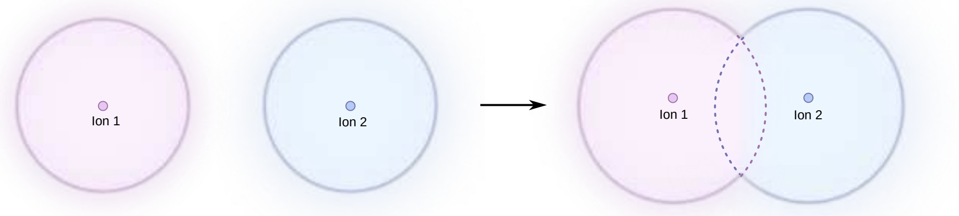


Figure 3.11: Two ions surrounded by a shell of QM solvent atoms, where each ion serves as the anchor/core atom of its restraining sphere. As two ions approach, their spheres merge adaptively. The larger the overlap between the two adaptive spheres the larger their radius will get on average.

Also in computational material science and nanotechnology the method might be useful for certain applications.

3.4.4 Discussion and Outlook

The QUASAR method has all the properties which seemed to be important for our intended field of application (i.e. FESs). It is based on energy-mixing, allows arbitrary shapes of the QM region, and could not be cheaper with one single QM/MM energy evaluation per timestep. In addition it brings along the advantages of the class of constraint methods for diffusive system, and therefore there is no need to smooth the forces due to the impermeable wall, which in addition prevents an unnatural net flow of particles from one region to another due to differences in the chemical potentials. Most importantly however, the great flexibility of the adaptive QM region renders the method suitable for a vast array of applications.

The method could be further extended in the future to allow for additional shaping options when defining the QM region. For instance, the concept of planes could be introduced to allow for even boundary surfaces. These could for instance be used to divide constraint spheres into two halves, which could be useful in specific cases. Alternatively, one could also introduce other basic shapes such as cubes or cylinders (which could for instance be useful when modeling carbon nanotubes).

The method was implemented in the software package i-QI, which is introduced in chapter 6 on page 105.

Chapter 4

Free Energy Methods

We must be clear that when it comes to atoms, language can be used only as in poetry. The poet, too, is not nearly so concerned with describing facts as with creating images and establishing mental connections.

Niels Bohr

Contents

4.1 Introduction

Free energies are of fundamental importance in many natural sciences, from physics over chemistry to biology. The free energy is the quantity which determines which processes are happening spontaneously, from chemical reactions to binding affinities of small ligands to their macromolecular molecular target.

Despite their central role in many sciences, significant research efforts, and dramatically increased computational power over the last decades, for many applications the ability to predict binding affinities by computational methods with sufficient accuracy and precision remains elusive. Therefore in most cases these methods still do not represent a replacement for laboratory experiments, but to this date often rather play a supportive role (if they are used at all).

The need for better methods seems clear, and the potential impact on fields such as pharmaceutical research enormous. Better free energy methods can be a central means to aid pharmaceutical research and development to shift more and more towards computational methods. And with that contribute to solving many problems which this field is facing these days.

Sources of Errors

Free energy simulation methods are a major class of FEMs, and many of them can in theory be exact. Yet in practice they rarely deliver the accurate free energy, due to different sources of errors which can affect the predicted value (see also [202]).

1. **The Conformational Sampling.** Free energy simulation methods generally depend on the phase space distribution function (or its restriction to the configuration space). However, to obtain the exact distribution requires infinitely long simulation times, and therefore the phase space distribution can only be approximated, which can lead to significant errors, in particular if some parts of the system (e.g. a protein) exhibit conformational changes only on large time scales. These errors are of random nature.

2. **The Force Field.** The force field (with which we refer to any modeling method, including quantum mechanical ones) is usually represented by the Hamiltonian, and the accuracy of the computed free energies by methods using the Hamiltonian depends also on the accuracy of it. The resulting errors which are caused by inaccurate force fields are of systematic nature.

3. **The System Representation.** The representation of the system which is used (i.e. the molecular model) for the simulations might be inaccurate, e.g. protonation states might be incorrect (when using MM based methods) or the initial conformation might be biologically irrelevant and the system trapped in an unrealistic minimum on the potential energy surface (PES). This can happen for instance if the model was obtained by homology modeling or if it is a low resolution nuclear magnetic resonance (NMR) structure. The resulting errors are again of a systematic nature.

4. **The Free Energy Method.** The method which is used to compute the free energy from the simulation data can be exact in theory, but it does not have to be. Free energy simulation methods can also be approximate in nature (e.g. if they neglect certain terms by purpose to accelerate the computations).

5. **The Code.** The implementation which is used to run the simulations and to carry out the free energy computations can contain bugs which lead to additional errors in the predicted free energies (on top of the other possible error sources).

The last three sources of errors can be eliminated in many cases already today. This is certainly true for the code, but also for the FEM as one can choose a method which is in theory exact. Regarding the system representation, it is not always possible to obtain a fully accurate one since not always the precise structure, composition or protonation states are known. But often it is possible to obtain a relatively accurate representation (for small molecules for example by computational methods, and for macromolecules if a high resolution structure is available). The two remaining sources, the sampling and the force field, generally represent the major sources of errors. And these two often reciprocally reinforce each other, as the more accurate the force field the more slow normally the sampling. However, a great number of methods have been developed to accelerate the the conformational sampling, such as umbrella sampling [291], metadynamics [15], methods belonging to the replica exchange family [216], coarse graining [293], adaptive resolution methods [230, 231, 315], or multiple time step algorithms [128, 205]. Still, conformational sampling remains a challenge for systems in which motions on very large time scales happen, as often found in biological systems. However, with sufficient time and computing power to converge free energy simulations with high accuracy quantum mechanical Hamiltonians it is already possible for certain types of systems. Yet also for these the predicted free energies are still far from being reliable enough. Therefore it seems that the force fields which are used are in general not accurate enough, and the new method which is presented in section 4.3 on page 69 is trying to make improvements in this regard.

Historical Notes

In the introduction of Chapter 2 it was mentioned that Max Born was one of the founding fathers of quantum mechanics. Most remarkably, he is also one of the pioneers regarding free energy methods [58]. In an article of the year 1920 the first traces of the pertubative approach for computing free energies can be discovered [35]. Work related to the pertubative approach has also been published by John Kirkwood in 1935 [147] as well as by Théophile De Donder in 1938 [58]. The Free Energy Pertubation (FEP) method, probably the most well known and widely used FEM still today, was published by Robert Zwanzig in 1954 [357]. However, there is another method still popular today which was published even earlier, viz. Thermodynamic Integration (TI) by Kirkwood in 1935 in the same article where he wrote about the pertubative approach [147]. In the decades following those early developments the methods have been used in many applications, and various new methods have been developed.

Chapter Overview

We will at first in Section 4.2 look at some of the basics of FEMs, and more closely in particular at alchemical free energy simulation methods. After that QSTAR, the newly developed FEM which is also of alchemical nature, will be introduced in Section 4.3.

4.2 Fundamentals

4.2.1 Classification of Methods

There is already a great variety of FEMs available, and it can be a significant challenge for newcomers in the field to navigate their way through the wide array of available approaches, or even get an overview of them. What contributes to this circumstance is that there is no unique way of classifying the methods, rather there are several ways to do so, and available reviews often give little information in this regard. Some of the most relevant characteristics which can be used to classify FEMs are the following:

- **Relativity:** Whether absolute or relative free energies are computed. While only relative free energies (i.e. free energy differences) carry a useful physical meaning, some methods compute the absolute free energy of the different states before subtracting them, while others compute the free energy difference directly without the need for absolute free energies as an intermediate step.
- **Empiricism:** Whether the method is based on empirical knowledge or on the laws of physics (i.e. an *ab initio* method).

- **Sampling:** Whether the method is based on one or a few conformations only or if it takes entire trajectories as input. In the latter case the method is called a FES method.
- **Transformation Type:** Whether the transformations are geometrical (meaning here physically possible) or alchemical (i.e. not observable in the real world).
- **Force Field:** The force field which is used to model the systems. Some methods require certain types of force fields, while other are independent of the modeling method.
- **Equilibrium Type:** Whether the method is an equilibrium method or a non-equilibrium method.
- **Number of States:** The number of thermodynamic states which the method is considering and for which conformations need to be provided. This can be 1-point methods, 2-point methods (also called end-point methods), n-point methods, or continuous methods (i.e. in theory an infinite number of intermediate states). 1-point methods which are not simulation methods are often called scoring functions, which are used in molecular docking.
- **Free Energy Type:** There are many different types of free energies (for instance solvation free energies, free energies of reaction, and many more). Some methods are designed for specific types of free energies (e.g. for binding free energies), and a subset of these are standalone methods, while the other part are working on top of other more general methods.

Some of the above characteristics are applicable to every method, while others are only relevant for certain types, which is helpful when creating a hierarchical classification scheme. However, to consider every possible characteristic in such a scheme is hardly practical. One possible (non-extensive) classification schemes which takes into account several of the above properties is shown in Figure 4.1 on the facing page (details on some of the simulation-based classes can be for instance be found in [57]). Also shown is a second classification scheme which is based on the type of free energy, which can be used to complemented the first scheme since some FEMs are designed for special types.

4.2.2 The Free Energy Difference Equation

Free energy simulation methods depend on the phase space distribution of the system, which can be approximated by MD or Monte Carlo (MC) based simulations. These methods are in most cases based on the *free energy difference equation* which serves as a starting point for their formal derivation [6].

To derive this equation, we assume we have given two canonical ensembles representing two states, the initial state A and the final state B. When we assume that they have the number of particles N and the same temperature T, and that the Hamiltonians are of the form

$$H(\boldsymbol{r}, \boldsymbol{p}) = K(\boldsymbol{p}) + K(\boldsymbol{r}), \tag{4.1}$$

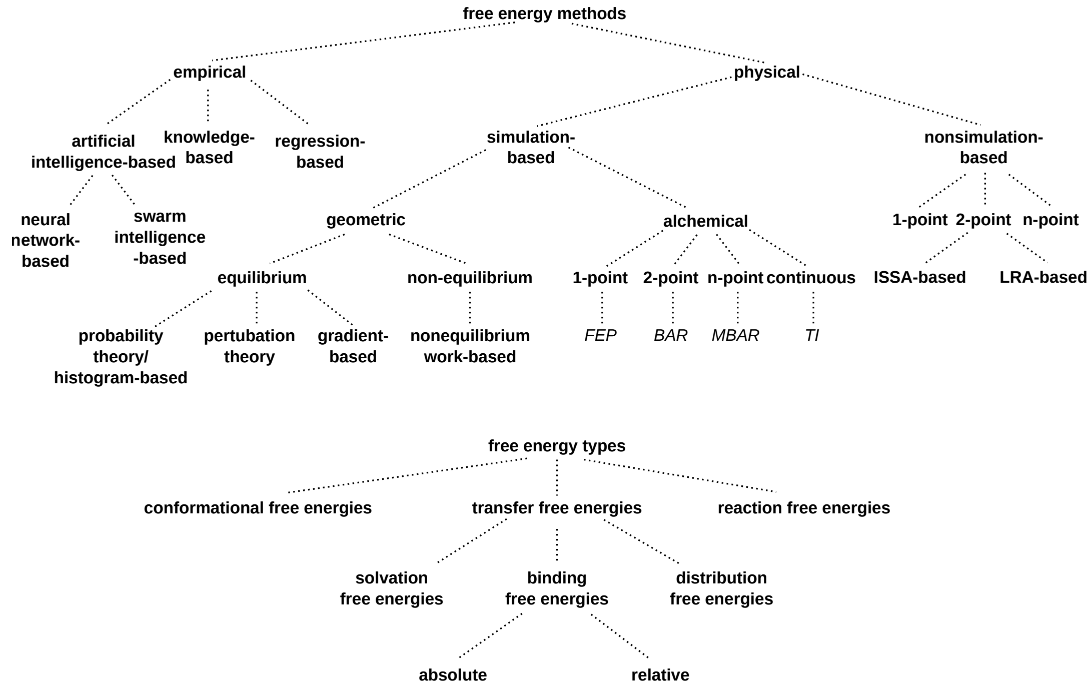


Figure 4.1: Upper tree: Major classes of free energy methods. ISSA stands for implicit solvent surface area and LRA for linear response approximation. Classes of free energy methods are in bold, while for the alchemical free energy methods several important example are included in italics. Lower tree: Several important types of free energies.

where $\boldsymbol{r} = (\boldsymbol{r}^{(1)}, ..., \boldsymbol{r}^{(N)})$ and $\boldsymbol{p} = (\boldsymbol{p}^{(1)}, ..., \boldsymbol{p}^{(N)})$, then we obtain for the *total* Helmholtz free energy difference $\Delta A^{\mathrm{t}}_{A\to B}$ [58]:

$$\begin{aligned} \Delta A^{\mathrm{t}}_{A\to B} &= A_B(N,V,T) - A_A(N,V,T) && (4.2)\\ &= -k_B T \ln Q_B(N,V,T) - kT \ln Q_A(N,V,T) && (4.3)\\ &= -k_B T \ln \frac{Q_B(N,V,T)}{Q_A(N,V,T)} && (4.4)\\ &= -k_B T \ln \frac{C_B Z_B(N,V,T)}{C_A Z_A(N,V,T)}, && (4.5) \end{aligned}$$

where Z is the configurational integral given by

$$Z(N,V,T) = \int\limits_{\mathcal{V}^N} e^{-\beta V\left(\boldsymbol{r}^{(1)},...,\boldsymbol{r}^{(N)}\right)} \prod_{i=1}^{N} d\boldsymbol{r}^{(i)}. \tag{4.6}$$

The last transformation for arriving at equation (4.5) is possible because

$$\begin{aligned} Q(N,V,T) &= C_1 \int\limits_{\mathbb{R}^{dN}} \int\limits_{\mathcal{V}^N} e^{-\beta\left(K\left(\boldsymbol{p}^{(1)},...,\boldsymbol{p}^{(N)}\right)+V\left(\boldsymbol{r}^{(1)},...,\boldsymbol{r}^{(N)}\right)\right)} \prod_{i=1}^{N} d\boldsymbol{r}^{(i)} d\boldsymbol{p}^{(i)} && (4.7)\\ &= C_1 C_2 \int\limits_{\mathcal{V}^N} e^{-\beta\left(K\left(\boldsymbol{p}^{(1)},...,\boldsymbol{p}^{(N)}\right)+V\left(\boldsymbol{r}^{(1)},...,\boldsymbol{r}^{(N)}\right)\right)} \prod_{i=1}^{N} d\boldsymbol{r}^{(i)}, && (4.8) \end{aligned}$$

where

$$C_2 = \int\limits_{\mathbb{R}^{dN}} e^{-\beta K\left(\boldsymbol{p}^{(1)},\ldots,\boldsymbol{p}^{(N)}\right)} \prod_{i=1}^{N} d\boldsymbol{p}^{(i)}. \tag{4.9}$$

The kinetic contributions which are included in the constants C_A and C_B in equation (4.5) can be expressed analytically, and are also referred to as the *ideal gas* contributions [91, 58]. Because they can be computed analytically, and because what can be measured experimentally are the configurational contributions, the quantity of interest is the so-called *excess free energy difference* [58]:

$$\Delta A_{A\to B} = -k_B T \ln \frac{Z_B(N,V,T)}{Z_A(N,V,T)}. \tag{4.10}$$

The fundamental equation (4.10) is the *free energy difference equation.* In the remaining text ΔA will always refer to the excess free energy difference.

4.2.3 Alchemical Free Energy Methods

One important class of FEMs are represented by the *alchemical* approaches. These methods have earned their name due their ability to compute free energies between states which contain different elements and different chemical species, and they can do so by transforming the initial state into the final state in a way which is unnatural (i.e. it would not happen in reality) [6]. However, the theory behind these methods is rigorous and exact. As an example, it is possible to transform lead into gold, a procedure on which mankind has worked on for millenia (as it is known that already the Hellenistic Greeks have pursued this endeavor [179]).

The ability of these methods to perform alchemical transformations has at least two major advantages. At first, it allows to compute free energy differences not only between states which can be physically and chemically compatible with each other. But it allows to consider free energy differences between states with different elements and chemical species. This circumstance can be of great value for instance in several applications such as in drug design where the difference of the binding affinity between different small molecules/ligands to a common receptor is often of paramount importance (e.g. in hit/lead optimization). This general applicability of alchemical methods is one reason why they are relatively widely used. Another advantage is that they provide an alternative means to compute free energies differences which can principally also be carried out by geometrical methods, but might be able to do so more efficiently.

Thermodynamic Pathways

As mentioned in the previous section, only free energies between two different states (here denoted as A and B) are of relevance. However, the computations do not have to be carried out directly between these two states. It is also possible to insert any number of intermediate states between the two states, compute the free energy differences between the pairs of adjacent states, and sum the free energy differences up to obtain the total free energy difference between the states A and B (see also [6] for the basic theory). If in mathematical terms we denote with S_i, $i \in \{1, ..., M\}$ all the states of the series the endstates are given by $S_1 = A$ and $S_M = B$, then we have

$$\Delta A_{\mathrm{A} \to \mathrm{B}} = \sum_{i=1}^{\mathrm{M}-1} \Delta A_{\mathrm{S}_i \to \mathrm{S}_{i+1}}. \tag{4.11}$$

It is often convenient to use the shorter notation ΔA_i instead of ΔA_{S_i}. The series of transformations between the two states A and B as described above is also called a *thermodynamic pathway*, and illustrated in Figure 4.2.

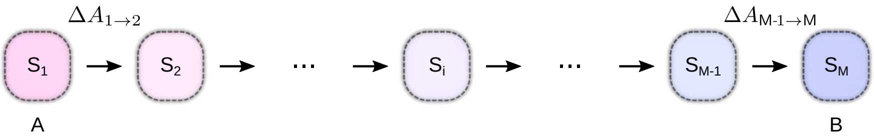
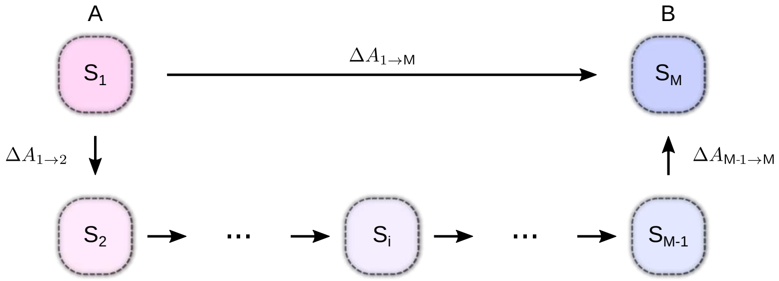


Figure 4.2: Thermodynamic pathway in which system A (initial state) is connected to system B (final state) by a series of intermediate states.

If the transformations which take place between the different states contain alchemical (i.e. non-physical/non-chemical) transformations, then the thermodynamic pathway is also called an *alchemical pathway.* Thermodynamic pathways also called thermodynamic cycles, because they can be represented as cyclic graphs as shown in Figure 4.3. Computing the free energy difference between the intermediate steps might involve more steps and therefore simulations, but it can still dramatically increase the overall efficiency, for instance due to faster convergence rates or reweighting potentials.



A
B
S1
$\Delta A_{1\to \text{M}}$
SM
$\Delta A_{1\to 2}$
$\Delta A_{\text{M-1}\to\text{M}}$
S2
···
Si
···
SM-1

Figure 4.3: Thermodynamic pathway arranged in a cycle in which system A (initial state) is connected to system B (final state) by a series of intermediate states.

Corresponding to the two types of pathways, the intermediate states can also be either physical or non-physical (alchemical). In the latter case they are often a linear combination of the two end states parametrized by some coupling parameter λ [58]. The *alchemical potential* (also called *meta potential* [297]) $V(\boldsymbol{r}_1, ..., \boldsymbol{r}_N; \lambda)$ of these states is in such cases normally of the form

$$V(\boldsymbol{r}; \lambda) = f(\lambda) U_{\mathrm{A}}(\boldsymbol{r}) + g(\lambda) U_{\mathrm{B}}(\boldsymbol{r}), \tag{4.12}$$

where $\lambda \in [0, 1]$ and the functions f and g are called the *coupling* or *switching functions* satisfying

$$f(0) = g(1) = 1 \tag{4.13}$$

$$f(1) = g(0) = 0. \tag{4.14}$$

However, intermediate states which are physical rather than alchemical also play an important role in applications, for instance when computing solvation free energy differences or binding free energies, and for such special types of free energy there also exist tailored thermodynamic cycles (see also subsection 7.2.2 on page 120 for more details on such examples).

Alchemical Topologies

At first the basic concept of dummy atoms is introduced (as for instance also described in [6]). Dummy atoms of molecules are atoms which do not interact with their environment, i.e. have no non-bonded interactions (e.g. electrostatic or van der Waals interactions). However, if they are part of a molecule they are usually still connected to the molecule via bonded terms to keep their position close to the location where the physical counterpart of the dummy atom would be located. If the states A and B have a different number of atoms then the standard approach is to add dummy atoms to these systems to match their atom numbers. This is important because in the fundamental free energy difference equation (4.10) it is assumed that the two systems possess the same number of degrees of freedom. Moreover, alchemical potentials (as for instance the version given in equation (4.12)) are functions of a single unified/mixed system, and therefore the same number of atoms is needed.

In alchemical transformations the atoms of the molecules which are transformed into each other can be classified in different ways, the most important of which are defined below.

Definition 4.2.1 (Special Atom Types)
Given two molecules (or, more generally systems) A and B which should be transformed into each other. Then we define

(i) $\mathcal{S}_\mathrm{A}$ *and* $\mathcal{S}_\mathrm{B}$ *to be the sets of atoms of the molecules (systems) A and B*

(ii) $\mathcal{A}_\mathrm{c}$ *the atoms in the atomic maximum common substructure (AMCS) of the two molecules (i.e. atoms types and the molecular connectivity graph have to match)*

(iii) $\mathcal{A}_\mathrm{a}$ *and* $\mathcal{A}_\mathrm{b}$ *to be the remaining atoms of A and B which are not in the atomic common molecular substructure, i.e.*

$$\mathcal{A}_\mathrm{a} := \mathcal{S}_\mathrm{A} \setminus \mathcal{A}_\mathrm{c} \tag{4.15}$$

$$\mathcal{A}_\mathrm{b} := \mathcal{S}_\mathrm{B} \setminus \mathcal{A}_\mathrm{c} \tag{4.16}$$

(iv) $\mathcal{T}_\mathrm{c}$ *as the set of atoms in the topological maximum common substructure (TMCS) of the two molecules (i.e. only the molecular connectivity graph has to match, not the atom types)*

(v) $\mathcal{T}_\mathrm{t}$ *as the set of atoms in the TMCS of the two molecules which are not in the AMCS (and therefore are only topologically shared), i.e.*

$$\mathcal{T}_\mathrm{t} := \mathcal{T}_\mathrm{c} \setminus \mathcal{T}_\mathrm{c} \tag{4.17}$$

(vi) $\mathcal{T}_\mathrm{a}$ *and* $\mathcal{T}_\mathrm{b}$ *to be the atoms of A and B which are not in the TMCS, i.e.*

$$\mathcal{T}_\mathrm{a} := \mathcal{S}_\mathrm{A} \setminus \mathcal{T}_\mathrm{c} \tag{4.18}$$

$$\mathcal{T}_\mathrm{b} := \mathcal{S}_\mathrm{B} \setminus \mathcal{T}_\mathrm{c}. \tag{4.19}$$

How exactly the molecular connectivity graph, the TMCS and the TMCS are determined can vary, depending on the precise method. Usually bond orders are ignored, and it is recommended to avoid ring opening/breakage [180]. Often they are created manually when one has to deal only with one or a few systems. But software tools have been developed such as FEW [121], LOMAP [181], or FESetup [186], which can be helpful when dealing with a large number of systems.

The augmented versions of the above sets are denoted with the superscript *aug*. Since the added dummy atoms are usually added such that the augmented versions of A and B have the same topology, we have

$$\mathcal{S}_{\mathrm{A}}^{\mathrm{aug}} = \mathcal{S}_{\mathrm{B}}^{\mathrm{aug}}. \tag{4.20}$$

Regarding the transformation of one molecule into another via alchemical transformations, there are different ways on how this can be accomplished. The major approaches are described below.

Dual-Topology Approach. In the dual-topology approach, the atoms in $\mathcal{A}_{\mathrm{a}}^{\mathrm{aug}}$ and $\mathcal{A}_{\mathrm{b}}^{\mathrm{aug}}$ are treated as separate atoms in each state of the alchemical pathway. When starting the alchemical transformations at the initial state A the atoms in $\mathcal{A}_{\mathrm{a}}$ are gradually decoupled from their environment, in other words transformed into dummy atoms. At the same time the atoms in $\mathcal{A}_{\mathrm{b}}$ are re-coupled into their environment, i.e. transformed from dummy atoms into fully interacting atoms. This means that the potential of the atoms in $\mathcal{T}_{\mathrm{t}}$ are computed two times, one time by the potential of the state A and one time of the potential of state B, even though the atoms in $\mathcal{T}_{\mathrm{t}}$ are replacing each other one-to-one (and thus can be seen as being transformed into each other). This double evaluation (or treatment) serves as the definition of the dual-topology approach [1].

In the standard dual-topology approach, the atoms in $\mathcal{A}_{\mathrm{a}}$ and $\mathcal{A}_{\mathrm{b}}$ have all their own coordinates . This approach can therefore be also called the *dual-topology dual-coordinate* approach [176, 124].

A non-standard dual-topology method, called *dual-topology single-coordinate* approach, has been described by Li *et al.* in [176].

Single-Topology Approach. An alternative to the dual-topology approach is the single-topology approach. When transforming an atom of the set $\mathcal{T}_{\mathrm{t}}$ into an atom of $\mathcal{S}_{\mathrm{b}}$, the transformation is carried out by representing them by a single joint atom whose energies and forces are obtained by evaluating a single potential function (which usually is representing a mixture of the two atoms which are transformed into each other).

The single-topology approach has a number of advantages over the dual-topology approach.

- **Computational Speed.** In the single-topology approach the atoms in the set $\mathcal{T}_{\mathrm{t}}^{\mathrm{aug}} \left(= \mathcal{A}_{\mathrm{a}}^{\mathrm{aug}} = \mathcal{A}_{\mathrm{b}}^{\mathrm{aug}}\right)$ are only evaluated by one potential, while in the dual-topology approach they are evaluated by two different potentials. Depending on the number of atoms in this set, and the modeling method which is used, this can make a significant difference regarding the computational expense.

[1] Some authors use alternative different definitions, which are often equivalent but not always.

- **Fewer Endpoint Problems.** Endpoint problems are problems which occur during simulations of alchemical states near the endpoints. With the single-topology approach there can be considerably fewer endpoint problems than the standard dual-topology approach has, as there are less dummy atoms which could spatially overlap with atoms of the environment, which can for instance cause problems for the free energy computations as the potential energies of the cross evaluations (which are used in many free energy methods) can become extremely high in such cases.
- **Faster Convergence.** The single-topology approaches can converge faster than the standard dual-topology approaches, in particular when the simulation times are not very long or if the number of thermodynamic windows is not very large [217, 176].
- **More Realistic Endstates.** In the single-topology approach there are fewer dummy atoms needed than in the standard dual-topology approach which requires for endstates dummy atoms for all the atoms in $\mathcal{T}_{\mathrm{t}}^{\mathrm{aug}}$. Since dummy atoms have no physical counterpart and not present in the real systems, they represent an unnatural deviation of the original systems, which can have an effect on properties like the free energy.

On the other hand, a major disadvantage of the single-topology approach is that it cannot be readily applied when QM methods are used, as QM-potentials make it difficult to merge the potentials of the two states into a single common potential. This gives rise for instance to fractional numbers of electrons which are difficult to handle and apply in general, even though such a method has been tested on some test cases [176, 264].

One of the advantages of the standard (dual-coordinate) dual-topology approach is that it can increase the phase space overlap between different states in certain cases, and thus enhance convergence. For instance to avoid the opening of rings. Therefore when the single-topology approach is used, it is often combined with specific uses of the dual-topology approach, giving rise to mixed topologies. However, not all simulation programs allow this, many only support one type of topology. A major advantage of dual-topology approaches in general is that they can be used readily with QM and QM/MM methods.

The non-standard dual-topology variant which uses single-coordinates (as mentioned above) is able to solve some of the disadvantages which the standard dual-topology approach has, but not all of them. Foremost it still needs to evaluate the potential energy two times (which is the very definition of the dual-topology approach), which can computationally be expensive, in particular for QM or QM/MM simulations. In addition it looses one of its main advantages, the improved phase space overlap between different states when functional groups of very different geometry are transformed into each other.

4.3 The QSTAR Method

4.3.1 Introduction

As outlined in the introduction of this chapter on page 59, the two most challenging sources of errors in free energy simulations are in most cases related to the force field and the conformational sampling. And the primary goal of this subproject was to develop a novel, next generation free energy method which

(1) holds the highest promise of all conceived ideas for new methods to increase the reliability (accuracy and precision) of free energy simulations
(2) is suitable for large biomolecular systems
(3) can be carried out at a reasonable computational expense.

An extensive investigation with the goal of identifying a new approach which satisfies the above points was carried out, and at the end it seemed clear to the author which idea is holding the most promise and potential. As it turned out it was possible to develop a new free energy method based on this idea, and it has several additional advantages (besides the ones listed above) to which we will come later in this chapter.

The new method, QSTAR, has several interesting analogies to the QM/MM approach which was introduced in Chapter 3:

(1) **Multiple Components.** Both of them are based on existing methods, and merge these as elementary parts into a new multi-component method.
(2) **Component Flexibility.** They are both very flexible regarding the individual components, allowing for a vast range of choices and combinations.
(3) **Schemes.** The new principles and parts which they add on top of the individual components allow for a large variety of options and additional developments. Therefore these sub-methods are in both cases also called *schemes*.

Another interesting connection between the two methods will become clear after the central principle of the method was introduced in Subsection 4.3.2.

The Overall Idea

In the development of the QSTAR method the force field was in the focus. Improving free energy methods in terms of their force field seems to be a daunting task as there exist already a large number of highly sophisticated potentials, both on the MM and the QM level. On the MM level are generic force fields such as the CHARMM General Force Field (CGenFF) [304, 306, 307] or the General Amber Force Field (GAFF) [313], but also specialized force fields for specific biomolecular systems such as the CHARMM protein force field [25] or its AMBER counterpart ff14sb [188]. Also on the QM level a wide range of methods have been developed for modeling the electronic structure of atoms and molecules, from extremely accurate ab initio methods such as post-Hartree-Fock methods based on Møller–Plesset perturbation theory (e.g. MP3 or MP4 [156]) or the coupled cluster technique [303], over sophisticated DFT functionals such as B3LYP [16, 267, 290], to advanced semi-empirical methods such as DFTB3 [93].

However, all these methods model either the entire particles or its nuclei as single points in space, which means they do not model the quantum nature of the nuclei explicitly. This approximation can be derived via the Born-Oppenheimer approximation, which is generally assumed to cause an negligible error for most applications [67]. But is this really the case? The nuclear delocalization is by trend the more profound the lighter the atom is, implying that hydrogen atoms are among the atoms affected most strongly. And indeed it has been shown that the acidity of certain amino acid side chains can lead to a 10 000-fold increase when nuclear quantum effects (NQEs) are modeled explicitly in comparison to nuclei represented by classical particles [317].

When now looking at biomolecular systems (our primary target application), hydrogen atoms are found nearly everywhere, being the most abundant atom in any living organism due to its role in water molecules. But in addition most organic molecules including biological macromolecules such as proteins also contain a large number of hydrogen atoms. Since these hydrogen atoms are not all buried somewhere, but found on the surfaces of the molecules, they are expected and known to directly take part in reactions and non-covalent interactions. For the above reasons it can therefore be expected that the quantum nature of the nuclei can play a significant role in many biological processes, such as certain biochemical reactions and non-covalent interactions, which are of central importance in drug discovery. Even more, it has been suggested that the error caused by treating light nuclei on the classical level is as least as large as the error of approximating the electrons by electronic structure theory [50]. Since heroic efforts have been made in the past decades to improve electronic structure methods, further progress in this area is generally difficult and a time consuming process. However, if the error due to the neglect of NQEs is expected to be significant for light nuclei, then it seems that the force field can be improved substantially in a single step by simply modeling the quantum nature of the nuclei. This is for instance possible via the path integral formalism (which was introduced in Chapter 2) which allows to model the quantum nature of particles, including atomic nuclei. In this thesis was therefore been attempted to develop a novel free energy method which is intrinsically based on the path integral formulation of quantum mechanics by making specific use of its unique structure.

4.3.2 General Formulation

The Principle

Molecular systems can be represented on the quantum level by the path integral formalism in real time, as was shown in Section 2.2. Moreover, certain thermodynamic properties can be expressed via the path integral formulation in imaginary time (as described in Subsection 2.3.2). This is true for instance for all expectation values of observables which only depend on the configuration of the system, and thus also for most alchemical FEMs such as the Bennett acceptance ratio (BAR) method. The classical isomorphism now allows us to map the finite path integral-based partition functions as specified in equation (2.89) of the states of interest to their classical counterparts given in equation (2.116). In this way one can approximate the states of interest as a system of classical ring polymers (or necklaces), in which each particle is represented by $P \in \mathbb{N}$ beads.

The key idea is now to transform individual beads of different thermodynamic states into each other, as illustrated in Figure 4.4. Physically such states are not possible from the quantum mechanical point of view, as each bead represents the same atom only different imaginary time, and thus they have to be identical from the quantum mechanical point of view. However, the classical isomorphism has provided us with a system of ring polymers which is well-defined, and the malleability of the potential energy function allows us to carry out such transformations of individual spheres (or beads), as well as to compute the free energy difference between the resulting intermingled alchemical states. Due to these transformations the new method was given the name Quantum Sphere Transformation Alchemical Route (QSTAR) approach, and it has has several distinct features and advantages as will be described later in this chapter.

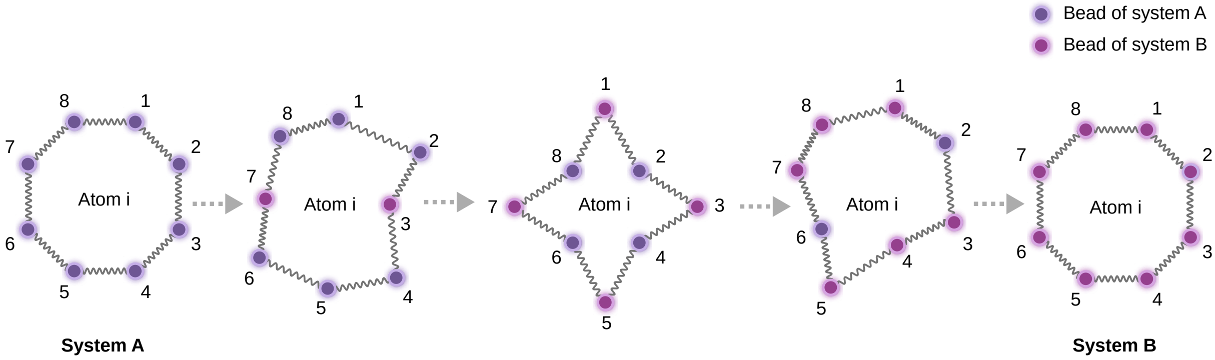


Figure 4.4: The principle of the QSTAR method. System A is transformed into system B by gradually exchanging the path integral beads of the necklaces of each atom from one system to the other. This will result in a number of alchemical intermediate states, depending on the number P of path integral beads. The different configurations of the ring polymers shown in the figure in the different states indicate that the ring polymers are constantly in motion during the time evolution.

The General QSTAR Potential

We will now formulate the potential of the QSTAR approach. For this purpose we consider two states A and B of N particles with potentials U_A and U_B. Moreover we let $P \in \mathbb{N}$ refer to the number of imaginary path integral time slices, and $P' \in \mathbb{N}$ such that $P' \geq P$. We now let V_A and V_B be potentials which

(i) are associated to states A and B

(ii) are parametrized by a variable $\kappa \in \{0, ..., P'\}$, i.e. $V_\mathrm{A} = V_\mathrm{A}(\kappa)$ and $V_\mathrm{B} = V_\mathrm{B}(\kappa)$, which represent κ-dependent modifications of U_A and U_B

(iii) satisfy $V_\mathrm{A}(P') = U_\mathrm{A}$ and $V_\mathrm{B}(0) = \mathrm{B}$, i.e. they recover the original potentials each at one endpoint of the κ-domain.

These potentials are called the *modified QSTAR potentials*. The parameter P' is introduced to allow the consideration of more different states than $P + 1$, the usefulness of which will become more clear later in this section, as will the motivation behind the endpoint conditions (iii). The *total QSTAR potential* in its general form is now defined by

$$W(\boldsymbol{r},\kappa) := \sum_{k=1}^{P}\sum_{i=1}^{N}\gamma_P^2\frac{m_i}{2}\left(\boldsymbol{r}_k^{(i)}-\boldsymbol{r}_{k-1}^{(i)}\right)^2 \tag{4.21}$$
$$+\sum_{k=1}^{\min\{\kappa,P\}}\frac{1}{P}V_A\left(\boldsymbol{r}_k^{(1)},...,\boldsymbol{r}_k^{(N)},\kappa\right)+\sum_{k=\kappa+1}^{P}\frac{1}{P}V_{\mathrm{B}}\left(\boldsymbol{r}_k^{(1)},...,\boldsymbol{r}_k^{(N)},\kappa\right),$$

where $\kappa \in \{0,...,P'\}$. In this definition we have

$$\boldsymbol{r}=\left(\boldsymbol{r}^{(1)},...,\boldsymbol{r}^{(N)}\right), \tag{4.22}$$

where each $\boldsymbol{r}^{(i)}$ denotes the coordinates of all the beads of particle i, and

$$\boldsymbol{r}^{(i)}=\left(\boldsymbol{r}_1^{(i)},...,\boldsymbol{r}_P^{(1)}\right) \tag{4.23}$$

are the coordinates of the P beads of the path integral necklaces representing the particle with index i. Thus there are NdP degrees of freedom, where d is the dimension of the space (naturally in most cases $d=3$). The above notation is compatible with the notation in Chapter 2 as we here have only cyclic paths, and thus $\boldsymbol{r}_0^{(i)}=\boldsymbol{r}_P^{(i)}$. And altogether there exist $P'+1$ different $W(\kappa)$-states. The parameter κ specifies how many beads of each state should be used for the intermediate alchemical states for the formation of entangled path integral necklaces, and it parametrizes the modified QSTAR-potentials V_{A} and V_{B}. If $P'>P$ then for $\kappa\in\{P+1,...,P'\}$ the total potential $W(\kappa)$ only consists of beads of system A and may vary only if the potential V_{A} varies with κ. In summary, the potential is a three-level potential involving the original state potentials U_A, U_B (indirectly), the modified state potentials V_A, V_B, and the total potential W representing the entangled ring polymers.

The above definition of the potential is the *naive* version because it is based on the primitive PIMD formulation (see equation (2.116) for the associated finite PIMD partition function). Enhanced PIMD versions have a slightly modified potential (which affect mainly the quadratic term connecting the beads), but the above QSTAR potential can be straightforwardly adapted to such variations as the concept does not change. In practice only the enhanced versions are used as they are considerably more efficient.

QSTAR States and Pathways

The QSTAR potential gives rise to $P'+1$ different thermodynamic states T_i, $i\in\{0,...,P'\}$. These states are also referred to as the $W(\kappa)$-states. The recovery condition (iii) on the modified potentials (specified on page 71) V_{A} and V_{B} ensures that $W(0)$ and $W(P')$ represent the pure end states A and B in their ring polymer representation, i.e.

$$W(\boldsymbol{r},P')=\sum_{k=1}^{P}\sum_{i=1}^{N}\gamma_P^2\frac{m_i}{2}\left(\boldsymbol{r}_k^{(i)}-\boldsymbol{r}_{k-1}^{(i)}\right)^2+\sum_{k=1}^{P}\frac{1}{P}U_A\left(\boldsymbol{r}_k^{(1)},...,\boldsymbol{r}_k^{(N)}\right) \tag{4.24}$$

and

$$W(\boldsymbol{r},0)=\sum_{k=1}^{P}\sum_{i=1}^{N}\gamma_P^2\frac{m_i}{2}\left(\boldsymbol{r}_k^{(i)}-\boldsymbol{r}_{k-1}^{(i)}\right)^2+\sum_{k=1}^{P}\frac{1}{P}U_B\left(\boldsymbol{r}_k^{(1)},...,\boldsymbol{r}_k^{(N)}\right). \tag{4.25}$$

Therefore only the intermediate states T_i, $i \in \{1, ..., P' - 1\}$ can indeed be true hybrid states.

When computing the free energy difference between the end states A and B via the $W(\kappa)$-states T_i, one has the freedom to choose which of these states should be included in the alchemical pathway which connects the states A and B, which means that not all of these states might be used but only some of them, the number of which we denote with $L \in \{2, ..., P' + 1\}$. To account for this circumstance the *state selection function*

$$h : \{1, ..., L\} \to \{0, ..., P'\} \tag{4.26}$$

is employed, which has to be strict monotonically decreasing (and thus injective), and to satisfy the endpoint conditions

$$h(1) = P', \tag{4.27}$$
$$h(L) = 0. \tag{4.28}$$

This function allows us to denote the states of the alchemical pathway as

$$Q_l := T_{h(l)}, \quad l \in \{1, ..., L\}. \tag{4.29}$$

These states are referred to as the *QSTAR states*. Because of the conditions on the function h the endpoints of the pathway coincide with the original endstates A and B, i.e

$$Q_1 = T_{h(1)} = T_{P'} \mathrel{\widehat{=}} W(P') \mathrel{\widehat{=}} \mathrm{A}, \tag{4.30}$$
$$Q_L = T_{h(L)} = T_0 \mathrel{\widehat{=}} W(0) \mathrel{\widehat{=}} \mathrm{B}. \tag{4.31}$$

When starting from the original potentials U_A and U_B, the states Q_l can be seen as the fifth level of the conceptual QSTAR hierarchy as illustrated in Figure 4.5.

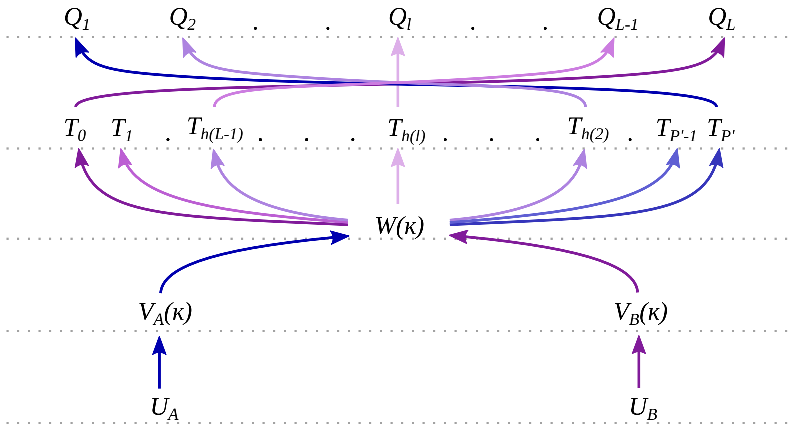


Figure 4.5: Conceptual hierarchy of the QSTAR method. Based on the original potentials U_A and U_B the modified $V_\mathrm{A}(\kappa)$ and $V_\mathrm{B}(\kappa)$ states are created, which in turn are used to define the total QSTAR potential $W(\kappa)$. The total QSTAR potential gives rise to the $W(\kappa)$-states T_κ. Not all of them are necessarily used during for the final alchemical pathway, and the state selection function h defines which of them are employed. The blue and berry colors in the figure are used to indicate the influence of states A and B, respectively, while the colors in between them indicate the degree to which both states have contributed.

Extended QSTAR Method

The QSTAR scheme can be generalized to include multiple different $W(\kappa)$-states for the same value of κ by parametrizing the modified QSTAR potentials V_A and V_B not only by κ but also by the alchemical pathway index $l \in \{1, ..., L\}$ (i.e. the QSTAR state index). In this case, the dependence of V_A and V_B on these two variables can be effectively reduced to the dependence of l since the κ-values of the $W(\kappa)$-states which are used for the alchemical cycle depend on l via the state selection function h.

4.3.3 Characterization

The QSTAR approach in its generic form as defined in the previous subsection has a variety of inherent properties and features.

Potential Flexibility

The QSTAR method can use virtually any atom-based modeling method (for the potentials U and $V(\kappa)$ introduced above). The necklaces arising from the path integral formulation which approximate the quantum nature of the particles can either represent entire atoms or only their nuclei. In the former case any of the available MM force fields can be used, and in the latter case any electronic structure method. Moreover hybrid methods such as the QM/MM method are also directly applicable, in which case the beads of the necklaces represent both atoms and nuclei depending on the region of the system.

Alchemical Topologies

The alchemical transformations of the QSTAR method are intrinsically of a single-topology nature, and thus bring along all its advantages: Fewer endpoint problems, faster convergence, less computational expense and more physically realistic endstates.

Single Topologies for QM Methods. As described earlier in this text (in Subsection 4.2.3), the standard single topology approach does not work in a practical way when QM methods are used in ordinary simulations (i.e. when the particles are represented by single points).

However, the inherent single-topology nature of the QSTAR method provides for the first time a complete, readily applicable, single topology approach for QM methods. And it does so not only in one way, but in two ways.

First, in the QSTAR method the particles are generally modeled on the quantum mechanical level due to the path integral representation. When the entire atoms are treated as elementary particles (i.e. when MM models are used for the potentials U and V above), then the PIMD simulations will be on the MM level. And the alchemical single-topology transformation of the QSTAR, while appearing to happen on the MM level, is in fact on the QM level since the ring polymers model the quantum nature of the particles.

Secondly, the potentials on U and V can be any QM and QM/MM potentials to model the electronic structure explicitly on the QM level as well. In this case the PIMD simulations are not on the classical level anymore, but on the QM level as well. This means that the QSTAR method provides a readily available single topology approach for electronic structure methods as well.

Single-Topology Dual-Coordinate Approach. Moreover, the QSTAR method not only supports one type of single topology, it provides two of them, viz. *single-coordinate* and *dual-coordinate* variants. Previously mainly three types of topologies have been available (as introduced in Subsection 4.2.3):

- Dual-topology dual-coordinate
- Dual-topology single-coordinate
- Single-topology single-coordinate.

When using the standard single topology approach for the QSTAR transformations, the transformations become automatically single-coordinate as is the case with ordinary alchemical pathways. But if the dual-topology paradigm is applied within the QSTAR framework, one obtains essentially a fourth type of topology, the *dual-coordinate single-topology* approach. The reason is that while with the QSTAR method the atoms in $\mathcal{T}_\mathrm{t}$ (as defined in Subsection 4.2.3) are also present two times when using dual topologies, this is only the case for beads with different indices when looking at associated atom pairs (from system A and B). This means that for these atoms there exist exactly P beads per atom-pair which are not treated as dummies. Therefore the corresponding atom pairs of the molecules A and B in $\mathcal{T}_\mathrm{t}$ the full potentials V_A and V_B are evaluated only P times, which is the same number as if the atoms pairs were treated via the single-topology approach, i.e. corresponding to one full potential evaluation per atom. And this is essentially the definition of the single-topology approach as stated in Subsection 4.2.3. Moreover, the ring polymers of these atom pairs have their individual coordinates, and hence it is a *dual-coordinate* approach. The single-topology dual-coordinate approach combines the key advantage of the dual-topology dual-coordinate approach in that the phase space overlap can be significantly improved when applied for certain functional groups with very different geometries, with the reduced computational costs of the single topology approach. And it can be used seamlessly in combination with the single-topology single-coordinate approach to create hybrid topologies.

Alchemical Pathway Flexibility

The general QSTAR potential as defined in equation (4.24) is very flexible due to its hierarchical structure and allows for a range of different alchemical pathways (also called schemes), which are represented by the QSTAR states $\mathcal{Q}_l$ as defined in equation (4.29). Several possibilities in this regard will be described in the following subsections 4.4.1 on page 77 and 4.4.2 on page 79.

Sampling

The simulations in the QSTAR approach are carried out by the PIMD technique. Several techniques have been developed which are able to speed up PIMD simulations. There is for instance the ring polymer contraction (RPC) method by Tom Markland [190, 189], which provides an efficient mechanism able to dramatically reduce the computational costs of PIMD simulations. This approach also works for QM and QM/MM potentials [134, 191, 139]. Alternatively, colored noise, generalized Langevin equation (GLE) thermostats which were introduced by Michelle Ceriotti can be used to speed up the computations [48, 49].

In addition, many acceleration techniques can be applied to PIMD which have been originally developed for MM or electronic structure simulations. To mention a few examples, metadynamics can be used as shown in [233], and also umbrella sampling naturally works as long as the number of beads remains the same. Also extrapolation methods such as second-generation Car-Parinello molecular dynamics can be readily applied [134].

FEM Flexibility

The QSTAR method is compatible with any alchemical FEM method, such as FEP, BAR, Multistate Bennett Acceptance Ratio (MBAR) or Weighted Histogram Analysis Method (WHAM).

Computational Expense

Even though if the potentials U and V are MM potentials, PIMD simulations are generally more expensive than classical MD simulations as the computational costs of the former scale more or less linearly with the number P of replicas. Therefore, when for example each atom is represented by ten beads, the computational demand will rise by roughly one order of magnitude in comparison to classical MM simulations. However, this rise in computational costs is extremely cheap for methods which model the quantum nature of the particles. For comparison, electronic structure theories (even on the fast semi-empirical level) are usually several orders of magnitude slower than MM methods. Also the scaling behavior of MM force fields is usually better than the one of QM methods. Therefore MM-based (as well QM/MM-based) PIMD simulations can be applied to extremely large systems (for instance biomolecular systems involving hundreds of thousands of atoms) to model the quantum nature of all atoms of the system, which is extremely challenging for even the most efficient linear-scaling QM methods.

In addition accelerated sampling methods as indicated above can be used to further reduce the costs. And the single-topology schemes which were described also decrease the costs since the Hamiltonian only has to be evaluated once per step. And since the single-topology single-coordinate scheme tends to accelerate convergence, it can further reduce the needed simulation time.

4.4 QSTAR Schemes

Within the QSTAR framework it is possible to design a large number of different alchemical intermediate states and schemes. At first we will introduce a classification scheme for the most important types of schemes.

Conventional alchemical pathways normally use an alchemical potential which is parametrized by an continuous variable $\lambda \in [0, 1]$ (as introduced in equation (4.12). We refer to these pathways as λ-potentials. In contrast, the QSTAR general potential $W(\boldsymbol{r}, \kappa)$ specified in equation (4.22) depends on a discrete variable κ. Therefore we refer to such discrete types as κ-potentials. The same holds for the alchemical pathways which are based on these potentials, e.g. pathways based on κ-potentials are referred to as κ-pathways or schemes. The κ-schemes of the QSTAR method are given by the states $\mathcal{Q}_l$, $l \in \{1, ..., L\}$ as defined in equation (4.29).

The κ-schemes can be classified primarily by the modified QSTAR potentials $V_\mathrm{A}(\kappa)$ and $V_\mathrm{B}(\kappa)$ (which were described on page 71), in several ways. These potentials modify the molecules A and B (given by their original potentials U_A and U_B), and can now be classified according to whether they transform the entire molecule uniformly or only a part (or fraction) of it. This classification gives therefore rise to

(1) *holotransformation schemes*, abbreviated as *η-schemes*,

(2) *semitransformation schemes*, abbreviated as *σ-schemes*.

Alternatively, the QSTAR schemes can be classified as

(1) *κ-invariant schemes*, abbreviated as *ι-schemes*,

(2) *κ-variant schemes*, abbreviated as *ν-schemes*.

QSTAR-schemes are intrinsically κ-dependent as follows by the very definition of the total QSTAR potential $W(\kappa)$. But the modified QSTAR potentials $V_\mathrm{A}(\kappa)$ and $V_\mathrm{B}(\kappa)$ do not have to depend on κ, and the above classification refers to these potentials rather than to $W(\kappa)$. All σ-schemes are necessarily κ-dependent since they change the molecules in some partial way, and hence they are always *ν-schemes*. But the situation is different for η-schemes in which the potentials $V_\mathrm{A}(\kappa)$ and $V_\mathrm{B}(\kappa)$ can be κ-invariant. For these schemes the it follows that

$$V_\mathrm{A}(\kappa) = U_\mathrm{A} \tag{4.32}$$

and

$$V_\mathrm{B}(\kappa) = U_\mathrm{B} \tag{4.33}$$

since by their definition on page 71 they have to satisfy $V_\mathrm{A}(P') = U_\mathrm{A}$ and $V_\mathrm{B}(0) = U_\mathrm{B}$. The above classification of schemes is included in Figure 4.6 on the following page showing the hierarchical structure.

4.4.1 Holotransformation Schemes

In the previous subsection two classes of η-schemes (holotransformations) were introduced, *ι-schemes* (κ-invariant transformations) and *ν-schemes* (κ-variant transformations). We will now look at both of them more closely.

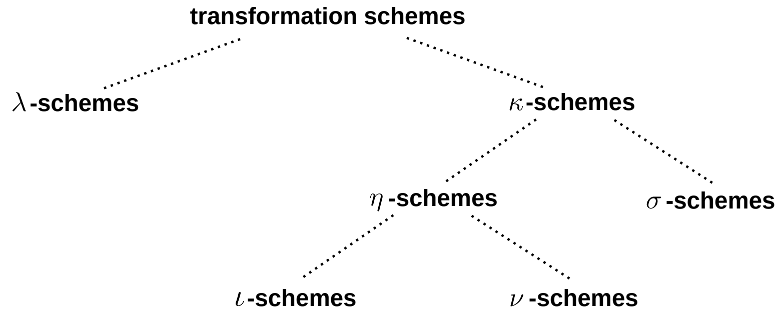


Figure 4.6: Overview of major classes of elementary alchemical transformation schemes.

ι-Schemes (κ-Invariant Schemes)

In ι-schemes the potentials V_{A} and V_{B} are κ-independent and equal to U_{A} and U_{B}, respectively. Even though these potentials are predetermined, there are still several choices such as

(1) $P \in \mathbb{N}$, the number of beads per path integral necklace

(2) $P' \in \mathbb{N}$, $P' \geq P$ the number total $W(\kappa)$-states

(3) $L \in \{2, ..., P+1\}$, the number of QSTAR states, i.e. the number of states which constitute the alchemical pathway

(4) $Q_l = T_{h(l)}$, $l \in \{1, ..., L\}$, the intermediate states of the alchemical pathway which are determined via the function h.

This scheme is illustrated in Figure 4.7 on the next page. Of course it is possible to simply include all the $P+1$ possible intermediate states, but one usually tries to minimize the number of states as much as possible since this will reduce the computational expense. However, if one reduces the number of states too much, the phase space overlap between the states will become too small, which again increases the required simulation time per state. Finding the optimal alchemical pathway between states remains one of the challenges in the field of alchemical FESs.

Even though ι-schemes are the most simple type of QSTAR schemes, they can be very useful for certain types of systems. In particular when there are no dummy atoms or only dummy atoms of small distance. The distance of a dummy atom is defined to be the number of covalent bonds (graph edges) of the shortest path to the common substructure (CSS) of the molecule. This can for instance be the case if the two molecules are very similar, but also if one molecule is transformed into a different version of itself. For instance when the systems are modeled on the MM level it can be favorable to switch off at first the electrostatic interactions of the molecule. In this case the initial state A would be the original system, and the final state B would be the same system except that in the molecule of interest the charges are set to zero.

ν-Schemes (κ-Variant Schemes)

Of a more complex nature than the ι-schemes are the ν-schemes, the other major type of η-schemes, since they allow the modified QSTAR potentials V_{A} and V_{B} to vary with κ. This principle is illustrated in Figure 4.8 on page 80. As an example on the MM level, the potentials V_{A} and V_{B} can be defined such that certain types of interactions of

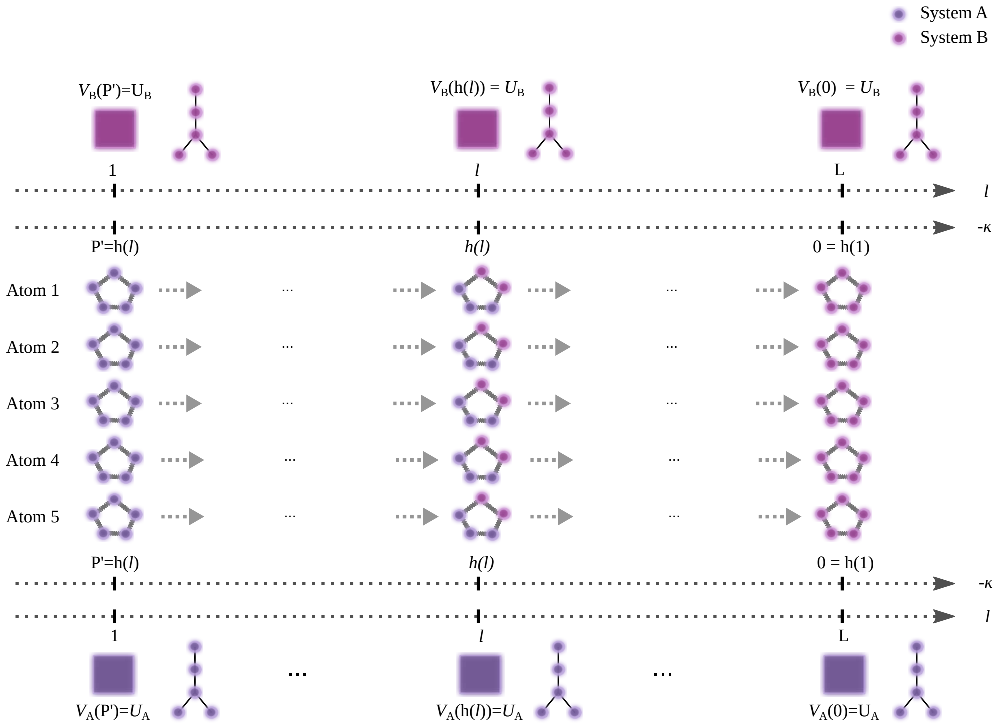


Figure 4.7: Illustration of the ι-scheme within QSTAR. A five-atomic molecule of system A is transformed into a five-atomic molecule of system B, the topology such that each atom can be mapped onto each other between the two system (i.e. there exist no dummy atoms). Since the potentials V_{A} and V_{B} are invariant under κ, these potentials which are intermingled within the $W(\kappa)$-states are equal to the original potentials U_{A} and U_{B}.

specific atoms of the molecule gradually switch off (for instance electrostatic or van der Waals interactions). This can for instance be accomplished by using a continuous order parameter $\lambda \in [0,1]$ which gradually switches off the desired components from $\lambda = 1$ (full interactions) to $\lambda = 0$ (no interactions).

4.4.2 Semitransformation Schemes: PEARL

In contrast to η-schemes in which the alchemically transformed molecules are only modified as a whole, in σ-schemes the molecules are modified non-uniformly.

σ-schemes can be very useful for instance when certain atoms of one state (or molecule) are to be transformed into dummy atoms during the alchemical pathway independent of the mixing with the other endstate.

In alchemical simulations with dummy atoms at the endpoints (i.e. at the initial state A and the final state B) we often have the problem that those atoms spatially overlap with the environment since there is no interaction with it (and thus no repulsive forces between the atoms). And this in turn can lead to serious problems for the FEMs since the potential cross-evaluations (i.e. the evaluation of the coordinates of one state at the potential of the another state) which are required for most alchemical FEMs lead either to extremely high energies or are not evaluable at all (since for instance numerical problems might occur or QM methods not converge).

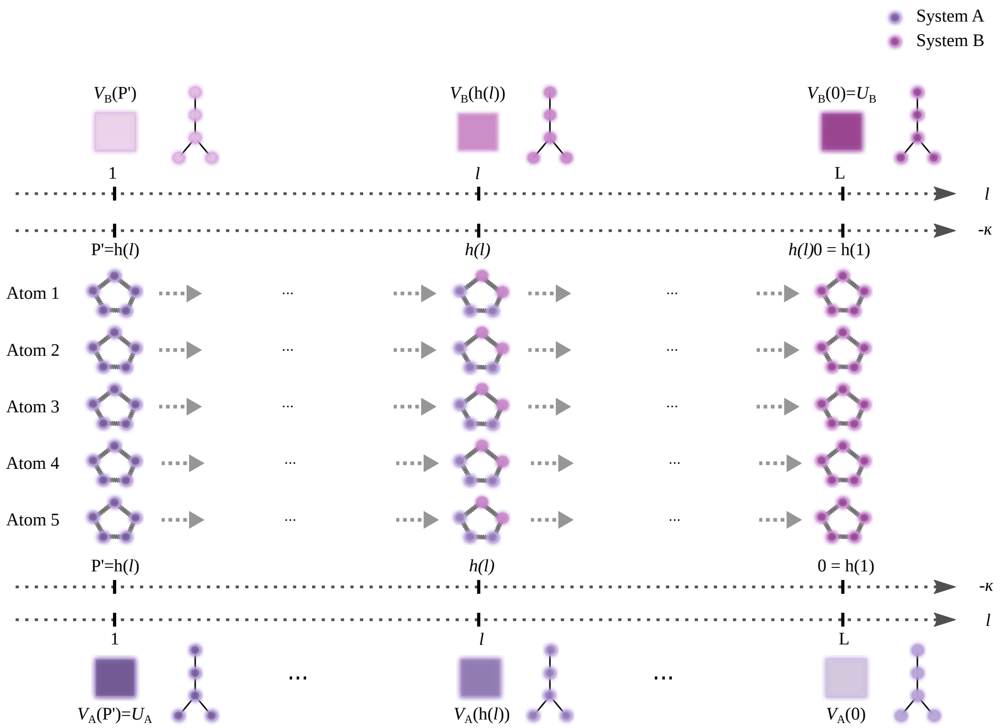


Figure 4.8: Illustration of the ν-scheme within QSTAR. A five-atomic molecule of system A is transformed into a five-atomic molecule of system B, the topology such that each atom can be mapped onto each other between the two system (i.e. there exist no dummy atoms). The potentials V_A and V_B are dependent of the value of κ, these potentials which are intermingled within the $W(\kappa)$-states change during the alchemical pathway.

The most common approach to solve these issues when the system is treated on the MM level is to use so-called soft-core potentials, in which the Lennard-Jones potential is replaced by more gentle functions which do not have a singularity at 0. For a single particle the following function is an example of a typical softcore potential [227, 258, 266]:

$$U(r,\lambda) = 4\epsilon\lambda^b \left(\left(C(1-\lambda)^a + \left(\frac{r}{\sigma}\right)^6 \right)^{-2} - \left(C(1-\lambda)^a + \left(\frac{r}{\sigma}\right)^6 \right)^{-1} \right), \tag{4.34}$$

where $a, b \in \mathbb{N}$, $\lambda \in [0,1]$, $r \geq 0$ is the distance between the two particles, and σ as well as ϵ correspond to the parameters of the standard Lennard-Jones potential. While this approach often seems to work well, it also has a number of disadvantages.

(1) The softcore potentials are physically more unrealistic than the Lennard-Jones potential, which might effect the properties of interest.

(2) Softcore potentials are only supported by some codes, in particular codes which have a native support for alchemical free energy simulations. However, there are many simulation programs available which do not support softcore potentials, such as ACEMD[114, 226] (an highly efficient code fully dedicated to running on graphics processing units), to mention only one example.

(3) This approach is not applicable for QM simulation, since there is no replaceable Lennard-Jones interaction.

The second problem is particularly relevant for PIMD simulations since there are not many programs available yet which are able to run them. Moreover, this approach lacks support for electronic structure methods. This is a practically relevant problem for the QSTAR approach since electronic structure methods represent a major class of potentials which can be used with it.

An alternative to softcore potentials is to transform only one or very few dummy atoms per alchemical step so that the appearing atoms still fit relatively well into the free space between the already fully interacting part of the molecule and the environment. This procedure is called the *serial insertion* approach and has been demonstrated by Stefan Boresch in [32]. It is able to solve all of the above mentioned problems which softcore potentials suffer from. However, the method was only presented in a rudimentary manner for alchemical transformation in which entire molecules are annihilated (as done frequently for instance when computing solvation free energies).

The PEARL Scheme

Introduction. Due to the above mentioned reasons the serial insertion approach seems generally to be the most flexible option when using the QSTAR method, and therefore we have developed a precisely defined family of alchemical schemes which is based on it. The new scheme is aimed to have the following features:

(1) **Arbitrary Alchemical Transformations.** In the original article [32] only the complete annihilation of molecules has been described. However, support for general alchemical transformations would allow the scheme to be used to a much larger range of applications. In particular also for relative binding free energies in CADD, which is one of our primary applications.

(2) **Precise Definition.** In the original article [32] the atoms which are transformed in each step were manually selected. A well-defined procedure which defines which atoms are added in which step allows the scheme to be implemented and applied automatically. This is for instance very helpful when one has to process a large number of different molecules at the same time.

(3) **Selectable Pathway Lengths.** The number of states in the alchemical pathway should be freely selectable.

(4) **Intelligent Atom Selection.** The atoms which are transformed in each step should be selected in a favorable way, which is a non-trivial task.

Condition (3) implies when the scheme is used within the QSTAR approach that the parameter $L \in \{2, ..., P' + 1\}$ can be freely chosen for any pairs of systems. The new scheme is called the Parallel Endpoint Atom Removal and Location (PEARL) approach (for reasons becoming clear further below in this subsection), and it will now be defined more precisely.

General Formulation. Given two molecules A and B which are to be transformed into each other, we denote with $\mathcal{M}_\mathrm{A}$ and $\mathcal{M}_\mathrm{B}$ the sets consisting of the atoms of A and B, respectively (the same definitions can be given A and B being arbitrary systems instead of single molecules only, but it is more accessible if the concept is introduced with the latter). Moreover we let $\mathcal{C}_{\text{css}}$ be the atoms of the CSS of the two molecules (i.e. the atoms which are transformed into each other), and we let

$$\mathcal{C}_{\text{A}} := \mathcal{M}_{\text{A}}/\mathcal{C}_{\text{css}} \tag{4.35}$$

$$\mathcal{C}_{\text{B}} := \mathcal{M}_{\text{B}}/\mathcal{C}_{\text{css}} \tag{4.36}$$

be the atoms of the two molecules which are transformed into dummy atoms during the alchemical pathway.

The atoms in $\mathcal{C}_{\text{A}}$ and $\mathcal{C}_{\text{B}}$ will now be located into layers in some way, meaning that each atom is assigned an integer which specifies its layer. Based on these integers we define

$$\mathcal{L}_S = \mathcal{L}_S(i) \subseteq \mathcal{C}_S,\ i \in \{0, ..., R_S\} \tag{4.37}$$

for $S \in \{\text{A}, \text{B}\}$ to be the corresponding sets of layers (e.g. $\mathcal{L}_{\text{A}}(i)$ represents the atoms of $\mathcal{C}_{\text{A}}$ which were assigned to layer i), where R_{A}, $R_{\text{B}} \in \mathbb{N}$ are the maximum number of layers of the sets $\mathcal{L}_{\text{A}}$ and $\mathcal{L}_{\text{B}}$, respectively. $\mathcal{L}_S(0)$ is defined to be the empty set, i.e. $\mathcal{L}_S(0) = \emptyset$. If the alchemical pathway is supposed to consist of $L \in \mathbb{N}$ states, then there will be $L - 1$ alchemical steps or transformations. Using these numbers we now define the *standard layer stepsizes* by

$$\Lambda_S^{\text{std}} := \frac{R_{\text{S}}}{L-1}, \quad S \in \{\text{A}, \text{B}\}, \tag{4.38}$$

and furthermore we let

$$\Lambda_S^{\text{rem}} := R_S \bmod (L-1), \quad S \in \{\text{A}, \text{B}\} \tag{4.39}$$

be the number of remaining layers (i.e the number of layers which would be left if in every alchemical transformation only the standard layer step size would be applied). These notions allow us to define the *individual step sizes*, which are dependent on the (alchemical) step index, by

$$\Lambda_S^{\text{ind}}(a) := \Lambda_S^{\text{std}} + \chi_{[1,\Lambda_S^{\text{rem}}]}(a), \quad a \in \{1, ..., L-1\},\ S \in \{\text{A}, \text{B}\}, \tag{4.40}$$

where χ is the characteristic function which takes the value 1 in the interval $[1, \Lambda_S^{\text{rem}}]$, and 0 otherwise. The individual step sizes take into consideration the Λ_S^{rem} remaining layers, and distribute them equally among the first Λ_S^{rem} steps. Based on these individual step sizes the *cumulative step sizes* are defined by

$$\Lambda_S^{\text{cum}}(a) := \sum_{i=1}^{a} \Lambda_S^{\text{ind}}(i), \quad a \in \{0, ..., L-1\}, \tag{4.41}$$

for $S \in \{\text{A}, \text{B}\}$ (where $\Lambda_S^{\text{cum}}(0) = 0$). They allow us to define the partial molecular atom sets of A and B corresponding to each alchemical state by unifying their layers with the CSS according to the cumulative step sizes, i.e.

$$\mathcal{P}_S(l) := \mathcal{C}_{\text{css}} \cup \left(\bigcup_{i=1}^{\Lambda_S^{\text{cum}}(l-1)} \mathcal{L}_S(i) \right), \quad l \in \{1, ..., L\}, \tag{4.42}$$

for $S \in \{\mathrm{A}, \mathrm{B}\}$. The above definition implies that the original atom sets can be recovered by

$$\mathcal{M}_S = \mathcal{P}_S(L) \tag{4.43}$$

and

$$\mathcal{C}_{\mathrm{css}} = \mathcal{P}_S(1) \tag{4.44}$$

for $S \in \{\mathrm{A}, \mathrm{B}\}$.

We are now able to introduce the notion of the *partial molecular potential functions*

$$\tilde{U}_{\mathcal{P}_S(l)}, \quad l \in \{1, ..., L\}, \tag{4.45}$$

which are defined as having the same form as the original potentials U_S, $S \in \{\mathrm{A}, \mathrm{B}\}$, except that only the atoms in $\mathcal{P}_S(l)$ are fully interacting particles while the remaining atoms (i.e. $\mathcal{M}_S/\mathcal{P}_S(l)$) are present only as dummy atoms. How exactly this transformation of a part of the system into dummy atoms is carried out is not predefined. With MM-potentials the transformation of atoms into dummy atoms is relatively unproblematic, but when the potential is QM level then the situation is more intricate. In this case QM/MM potentials can be one way to allow the transformation of fully interacting atoms into dummy atoms.

So far the scheme is of a general nature and independent of the topology to be used, and it can now be adopted for both single- and dual-topology approaches. For single-topologies one would merge the sets $\mathcal{P}_\mathrm{A}$ and $\mathcal{P}_\mathrm{B}$ using counter-rotating indices.

Adoption to QSTAR. We will now further develop the PEARL approach into a complete QSTAR scheme. For this purpose we only need to define the modified QSTAR potentials V_A and V_B, which (with the notation introduced in this subsection) can be simply done by

$$V_\mathrm{A}(h(l)) := \tilde{U}_{\mathcal{P}_\mathrm{A}(L+1-l)}, \quad l \in \{1, ..., L\}, \tag{4.46}$$

and

$$V_\mathrm{B}(h(l)) := \tilde{U}_{\mathcal{P}_\mathrm{B}(l))}, \quad l \in \{1, ..., L\}. \tag{4.47}$$

Only those modified QSTAR potentials were defined which are needed for the alchemical pathway, and therefore the same holds for the $W(\kappa)$-states. The reverse direction for the state A is needed since this molecule is reduced to the CSS during the alchemical pathway, while path molecule B is grown from the CSS to its fully interacting version. The complete scheme is illustrated in Figure 4.9 on the next page. The name of this method, PEARL, is now more easily comprehensible. From the endstate (endpoints) of the alchemical pathway (i.e. the states A and B), atoms are gradually removed from the initial state A and added in parallel to the state B to grow it to its full form found at the endpoint.

Layer Assignment Schemes

What is still open is the definition of how the atoms of the sets $\mathcal{C}_\mathrm{B}$ $\mathcal{C}_\mathrm{B}$ are located to the layers.

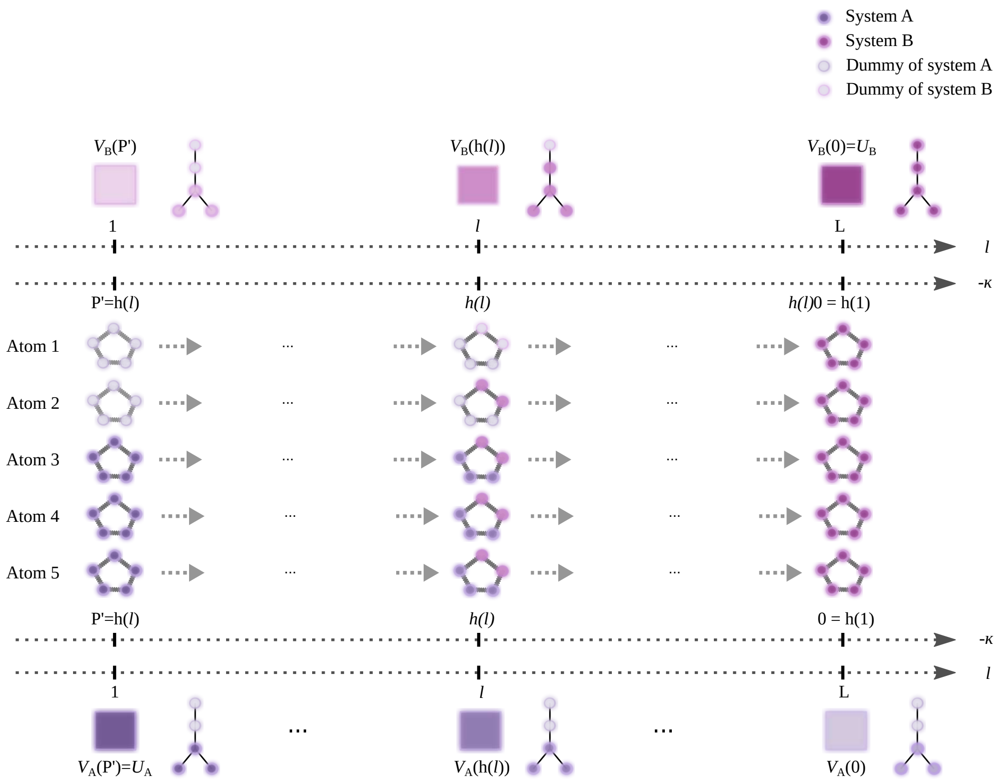


Figure 4.9: Illustration of the PEARL-scheme as a special case of a σ-scheme within QSTAR. A three-atomic molecule of system A is transformed into a five-atomic molecule of system B, resulting in two dummy atoms of in the augmented system A (i.e. QSTAR state $l = 1$). During the alchemical pathway V_B, the modified QSTAR potential of system B, gradually transforms the dummy atoms into fully interacting atoms.

PEARL-P. The most straightforward approach would be compute the distance from each atom in the above sets to the CSS, and to place the atoms into the layers corresponding to these distances. This approach is shown in Figure 4.10, and it is referred to as PEARL Primitive (PEARL-P). This scheme can work well in certain cases, but it also has several weaknesses. For instance if a carbon atom is connected to three atoms with a large van der Waals radius which are all located in the next layer (while the fourth bonded atom connects the carbon to the CSS, i.e. is located one layer below itself), then we might still experience exactly the problem which this approach is supposed to avoid (i.e. the spatial overlap of the appearing atoms with atoms of the environment) since if all three of these large atoms appear besides each other they will together consume a relatively large volume in space. This problem is not only of hypothetical nature, but was encountered by the author (and actually identified in the first place) in test simulations which we carried out.

PEARL-N. To solve the described problem a more advanced scheme was devised. If the distance of the atoms which have to be assigned to layers is l, then they are not automatically all added to the new layer as it would have been with the primitive mode of Parallel Endpoint Atom Removal and Location (PEARL). Instead, for each atom of the current distance l the common parent atoms are determined, i.e. the atoms of distance $l - 1$ which are bonded to the atoms of the current distance l. The parent atoms

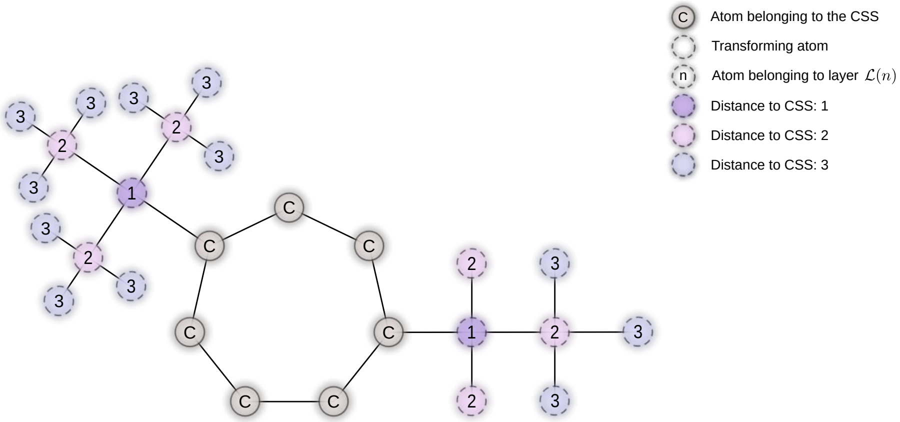


Figure 4.10: Illustration of the PEARL-P scheme by an example involving a small organic molecule (representing either of the two molecules, A or B, which are transformed into each other). The atoms are colored by their (shortest) distance to the CSS. The numbers within the non-CSS atoms represent layer indices to which they were assigned to by the PEARL-P scheme, which by definition of this scheme equals their distance.

allow the clustering of the current atoms (i.e. their children) into groups of *neighboring atoms* which are denoted by $\mathcal{G}_i$, $i \in \{1, ..., G\}$. If the group $\mathcal{G}_i$ consists of M_i atoms, then they are distributed among the next M_i layers, one per layer. This can happen simultaneously with all the neighbor groups $\mathcal{G}$ which exist of the current distance, so that each of the next layers can contain multiple atoms. Therefore, if $M = \max\{M_1, ..., M_G\}$, then M new layers will be created and contain at least one atom. This algorithm starts at the CSS and progresses outwards to the atoms which are farthest away. After all atoms of a particular distance were assigned to the layers, the procedure proceeds in the same way with the atoms of the next-higher distance. The scheme is called the PEARL Neighbor-separation (PEARL-N) mode, and it is illustrated in Figure 4.11. The

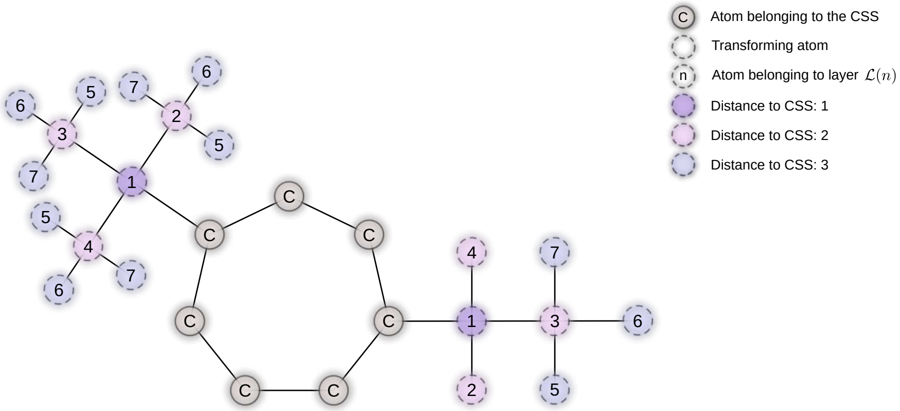


Figure 4.11: Illustration of the PEARL-N scheme by an example involving a small organic molecule (representing either of the two molecules, A or B, which are transformed into each other). The atoms are colored by their (shortest) distance to the CSS. The numbers within the non-CSS atoms represent the layer indices to which they were assigned to by the PEARL-N scheme, which considers groups of neighboring atoms (i.e. atoms bound to the same parent atom, where a parent has to be closer to the CSS than the child atoms). In contrast to the PEARL-P scheme, neighboring atoms are assigned to successive layers rather than to the same layer.

approach was tested, and it solves the numerical and convergence problems indeed as intended.

PEARL-B. While PEARL-N seems to work relatively well and solves the problems it was designed for, the scheme can in some cases increase the number of layers significantly, which in turn can lead to longer alchemical pathways (if for instance one layer should be added per alchemical pathway step), and hence to increased computational costs. While for some cases the layer assignment seems to be nearly optimal (such as for the molecule illustrated in Figure 4.11), there are other cases for which the scheme does not work optimally, but for which a more intelligent assignment procedure is possible. This is for instance the case with the molecule shown in Figure 4.12 on which the PEARL-N scheme was applied. As can be seen in this example, in the right branch of the molecule the atoms of distance 4 (light blue) are assigned only after all the atoms of distance 3 were assigned, which causes a delay for the right branch atom assignment because the left branch contains more than three atoms of distance 3 while the right branch only contains one. The problem in this example is that the atoms of the neighboring group in one branch are blocking the distribution of other neighboring groups which are smaller but have longer chains present in their remaining subtrees. It therefore seems favorable if individual branches and all subbranches would be considered during the assignment procedure, which can significantly reduce the number of layers (as shown in Figure 4.13) and thus length of the alchemical pathways. This scheme is referred to as the PEARL-B scheme. Since this scheme also separates the neighboring atoms, it can be seen as an advanced neighbor-separation scheme, while PEARL-N is a more basic variant.

PEARL-XH. The PEARL Branch-consideration (PEARL-B) scheme seems already nearly optimal. However, further enhancements are possible. PEARL-B separates its neighbors, but this is in most cases only needed if the atoms which are separated are not hydrogen atoms. Hydrogen atoms have a significantly smaller van der Waals radius

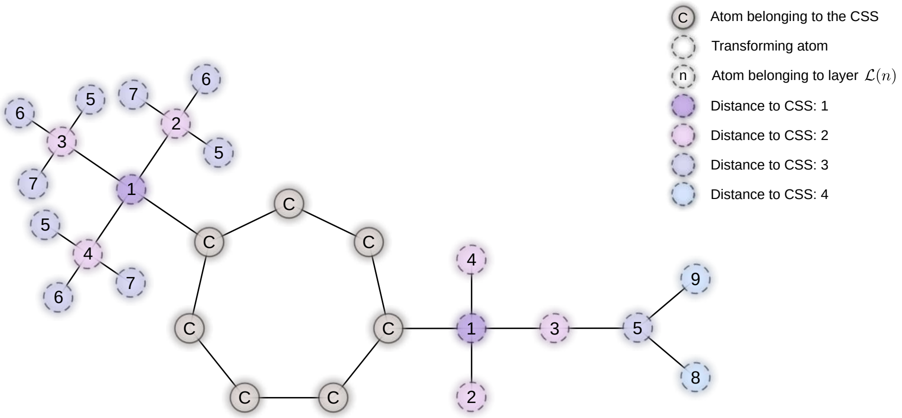


Figure 4.12: Application of the PEARL-N scheme on another pair of example molecules (referred to as example 2, while example 1 is the molecule pair which was used in the previous illustrations). In the current example a weakness of PEARL-N becomes visible. Because this assignment scheme is based on the CSS-distance of the atoms only, and atoms of the next higher distance are assigned only if all the atoms of the current distance are allocated, more layers are created than in fact really needed to achieve the desired neighbor separation.

than the larger atoms of organic molecules (which are primarily carbon, nitrogen or oxygen). Test simulations have shown that hydrogen atoms are so small that it is usually no problem if there appear multiple of them bonded to the same atom, as is often the case in small organic molecules. Therefore the PEARL schemes were equipped with an *hydrogen-single-step* option. When activated then all the hydrogen atoms are not considered during the layer assignment procedure. This feature works with all the previously described schemes, and is referred to as PEARL-X Hydrogen-single-step (PEARL-XH), where X is a placeholder for any of the available, previously introduced PEARL schemes. It is illustrated in Figure 4.14. As can be seen in the figure, the number of layers in this example is 5, while with the basic neighbor-separation scheme PEARL-N the number of layers is 9. This is a reduction of nearly 50 %, and in real applications the reduction can be even higher due to the vast diversity among organic molecules.

The number of layers does not determine the number of steps in the alchemical cycle of the PEARL scheme, since the layer assignment was completely decoupled from the number of pathway steps L. As described in the definition of PEARL, if there are more layers than steps then multiple layers are united and added for single steps. The PEARL method and all of the above schemes were implemented in the HyperQ suite (see Chapter 7).

4.4.3 Macroschemes

Multiple alchemical pathways can generally be joint together to give combined (macro) cycles. There are several ways of how this can be done in a useful way with QSTAR schemes.

This can for instance be done with QSTAR schemes only. To mention some examples, on the MM level there are different types of interactions: bonded terms, van der

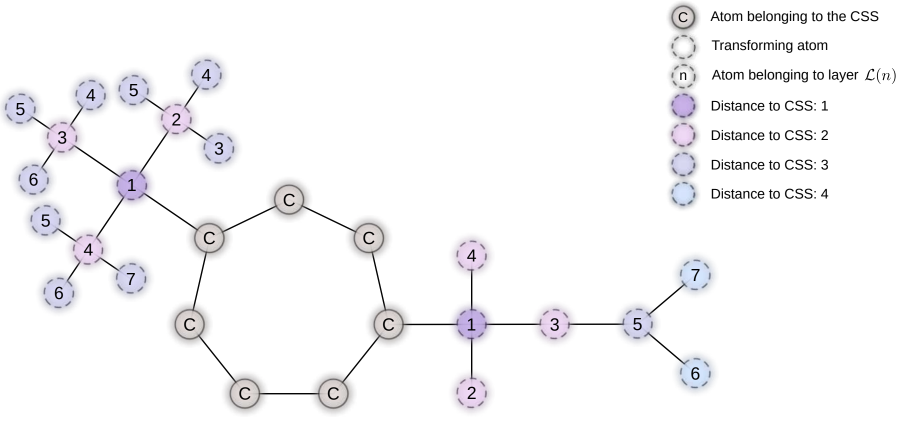


Figure 4.13: Illustration of the PEARL-B scheme by application to example 2. The atoms are colored by their (shortest) distance to the CSS. The numbers within the non-CSS atoms represent the layer indices to which they were assigned to by the PEARL-B scheme. In contrast to PEARL-N, during the allocation procedure in which the neighbors are separated, PEARL-B also considers the branches (and all subbranches) individually, which prevents unnecessary blockages and therefore reduces the numbers of layers.

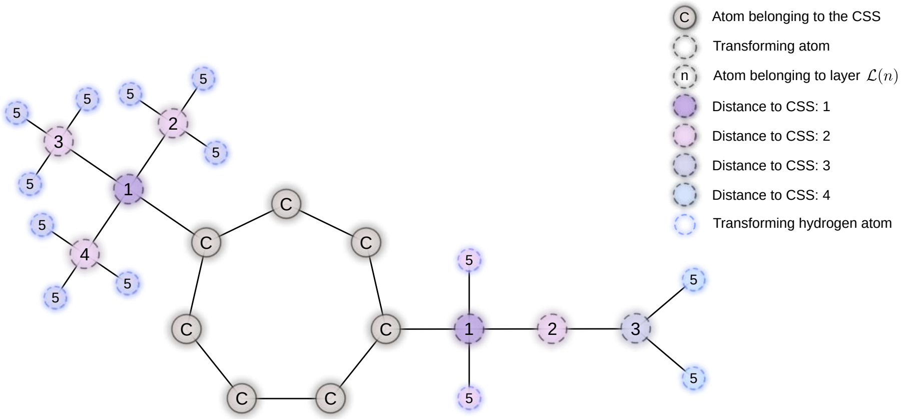

C Atom belonging to the CSS
Transforming atom
n Atom belonging to layer $\mathcal{L}(n)$
Distance to CSS: 1
Distance to CSS: 2
Distance to CSS: 3
Distance to CSS: 4
Transforming hydrogen atom
C
1
2
3
4
5

Figure 4.14: Illustration of the PEARL-BH scheme (i.e. the hydrogen-single-step option of the PEARL-XH scheme applied to PEARL-B scheme, therefore the character X is replaced with by B). In contrast to the previous PEARL schemes, PEARL-XH takes into account the atom type, and transforms all hydrogen atoms in a single unified step due to their relatively small size in comparison to the other atom types found in organic molecules. The molecule shown is the same molecule as in the previous two examples, except that all monovalent atoms are assumed to be hydrogen atoms. These atoms are shown with reduced size and their stroke is colored in blue.

Waals interactions, and electrostatic interactions. It can be favorable during alchemical transformations to switch off at first a certain type of interaction, such as the electrostatic interaction, for different reasons (such as to increase the phase space overlap between different alchemical states). This leads for example to macroschemes of the form

$$A \xrightarrow{\Delta A_{A\rightarrow A'}} A' \xrightarrow{\Delta A_{A'\rightarrow A'}} B' \xrightarrow{\Delta A_{B'\rightarrow B}} B. \tag{4.48}$$

This in turn allows to apply different combinations of QSTAR schemes, such as η schemes for the intramolecular alchemical transformations and ν schemes such as PEARL for the intermolecular changes to accommodate the possibly present dummy atoms:

$$A \xrightarrow{\eta-\text{scheme}} A' \xrightarrow{\nu-\text{scheme}} B' \xrightarrow{\eta-\text{scheme}} B. \tag{4.49}$$

Alternatively, one can combine the discrete QSTAR schemes (which are κ-schemes) with λ-schemes (i.e. schemes with a continuous alchemical potentials), a promising combination which might take the shape of

$$A \xrightarrow{\lambda-\text{scheme}} A' \xrightarrow{\nu-\text{scheme}} B' \xrightarrow{\lambda-\text{scheme}} B. \tag{4.50}$$

These are only few examples, there are many other ways of constructing additional types of thermodynamic macrocycles involving QSTAR. Among these, an interesting possibility involves for instance to switch from an accurate but computationally expensive potential at the endstates to more efficient but less accurate potentials during the intermediate states. Such a switch can occur within the array of electronic structure methods, but also from the QM level to the MM level.

4.4.4 Discussion and Outlook

In this section the QSTAR method was described with its primary purpose to provide a free energy method of a new generation which primarily improves aspects related to the force field and the alchemical pathways. This is achieved by allowing to model not only the electronic structure on a quantum mechanical level, but also the quantum nature of the nuclei. Combining both path integral based approaches with electronic structure methods can in principle provide one of the most accurate simulations for biomolecular systems feasible today. QSTAR does not only achieve its goal by simply including nuclear quantum effects modeled by the path integral formulation, but also by providing a new class of thermodynamic pathways which combine PIMD and with alchemical FES in a promising and symbiotic way, giving rise to an emergent method which is more than the sum of its parts.

Among the advantages and characteristics of the QSTAR approach are the following:

- Support of almost any MM, QM and QM/MM potential energy function.
- Native support of single topologies (and with that their general advantages), in more detail:
    1. Support of the standard single-topology approach (single-topology singe-coordinate).
    2. Introduction of the new single-topology dual-coordinate approach.
    3. Making available the first readily applicable single-topology approach for QM and QM/MM methods.
- Support of most alchemical free energy methods (FEP, BAR, ...).
- Usable with a wide choice of different alchemical pathways (e.g. ι-schemes, ν-schemes or σ-schemes such as PEARL)
- Sampling efficiency. Compatible with many acceleration methods for MM and QM simulations.
- Relatively low computational effort (for including the quantum nature of the particles).
- Arbitrary accuracy on the path integral level since the number of beads P can be freely chosen (as well as the potential energy function).

Details on many of these aspects have already been described in subsection 4.3.3 on page 74.

On page 69 several analogies were mentioned which QSTAR has in common with the QM/MM approach. Now after the method is described, some differences also become clear. The QM/MM method combines QM and MM methods, and tries to provide the advantages of both of both worlds. However, it does so at expense. The QM/MM boundary region can be a major additional error source, which is caused by the circumstance that QM and MM methods are fundamentally different from each other and are by themselves not compatible. Therefore QM/MM can be less accurate than if the QM method would have been applied to the full system. And while it can greatly accelerate the computational speed in comparison to a full QM treatment, it cannot reach the efficiency of MM methods. With QSTAR the situation is different because PIMD and alchemical free energy methods seem to be made for each other. They suit perfectly together and can be combined in a beautiful and emergent way, which rather than being an additional source of error seems to have positive effects on the free energy simulations. Another conceptual difference between QM/MM method and the QSTAR approach is the latter can use the former as one of its elementary components.

The major disadvantage of the QSTAR method is its high computational demand. However, the computer power is increasing continuously, and it will most likely only be a matter of time until PIMD simulations can be routinely applied to larger systems. The quantum nature of the nuclei can have a significant impact in particular on light atoms such as hydrogens, and their negligence can lead to substantial errors in biomolecular systems as indicated earlier [50].

The method was implemented in the free energy suite HyperQ (which is described in chapter 7 on page 117), as well as extensively tested. Moreover QUASAR, the QM/MM method described in Section 3.4, was made available for QSTAR with i-QI (to be described in chapter 6 on page 105).

Chapter 5

Structure-Based Virtual Screenings

> *To those who do not know mathematics it is difficult to get across a real feeling as to the beauty, the deepest beauty, of nature. ... If you want to learn about nature, to appreciate nature, it is necessary to understand the language that she speaks in.*
>
> Richard P. Feynman
> *The Relation of Mathematics to Physics*

Contents

5.1 Introduction

Structure-based virtual screenings are one of the most promising approaches for identifying new hit compounds in CADD projects. When carrying out virtual screenings for a specific target, two primary goals are:

(1) **Binding Strength.** Identification of small molecules which have a binding affinity to the target as strong as possible.

(2) **Multiple Hits.** Obtaining not only one but as many as possible promising hit compounds binding to the target.

A high binding affinity has various advantages of high significance. It reduces the dose required when taking the drug, and hence the side effects which are nearly always present. It provides more freedom during optimization of the drug since one may afford to loose some binding affinity but therefore is able to improve other important properties such as ADMET related properties. These advantages also increase the chance that the compound or its derivations will make it through the preclinical and clinical testing phases, and therefore save considerable amounts of manpower, money, development time, and at the end possibly even many lives. And the benefits of obtaining many hits are of a similar nature. The more hits, the more variety and the more backup compounds one has during the subsequent drug development steps.

However, there are major obstacles in the endeavor to achieve the above mentioned goals. The root cause of these problems is the error of the FEMs which is used during the virtual screening. The major error types of FES methods have already been described in Section 4.1. For virtual screening procedures FES methods are rarely used due to their high computational costs. Instead fast FEMs are normally used which fall into the class of methods called scoring functions (i.e. approaches based on a single conformation, see 4.1 for more complete classification) which are employed within docking programs. One major disadvantage of this class of methods is that the errors are usually significantly higher than for FES methods (provided that the simulations have converged) due to their coarse approximative nature. The errors in the free energy estimates lead in turn to:

(1) **Inaccurate Affinity Rankings.** The order of the compounds in the affinity rankings becomes more inaccurate, which makes it difficult to identify the strongest binders to the target among the screened compounds (and thus to achieve goal (1) from above).

(2) **False Positives.** The incorrect order of the ranking can lead to false positives, which often reduces the true hit rate (success rate) of virtual screenings dramatically. The same is true for false negatives, which implies that many real binders hits might be missed (and also thus goal (2) from above).

When thinking about possible ways on how these problems can be solved (or at least improved), it might appear that developing free energy methods which can provide the sufficient accuracy, precision and speed is the way to go forward as it tackles the root cause of the above mentioned problems. A new class of FES method (QSTAR) has already been developed in this thesis and described in Chapter 4. However, this is (at least in its present form) not suitable to replace scoring functions in virtual screenings due to its high computational demand. One possibility would be to develop a faster (but therefore probably also more inaccurate) version of the method which is suitable for virtual screenings. However, there are at least two alternatives (to improving the FEM) for better achieving the two primary goals described above:

(1) **Upscaling.** Increasing the number of compounds to be screened.

(2) **Multistaging.** Using multiple virtual screenings in several layers.

And these two approaches are explored in more detail in this chapter, upscaling in section 5.2 on page 94 and multistaging in 5.3 on page 99. The combination of these approaches, as well as earlier methods described in this thesis, is the topic of the last section 5.4 on page 100.

Structure-based virtual screenings are most commonly used in CADD, but they can be of use also in other areas. One example is computational material science and nanotechnology, and virtual screening related projects in this field have been reported for instance in relation to the development of organic light-emitting diodes (OLEDs), organic photovoltaics and organic batteries [232].

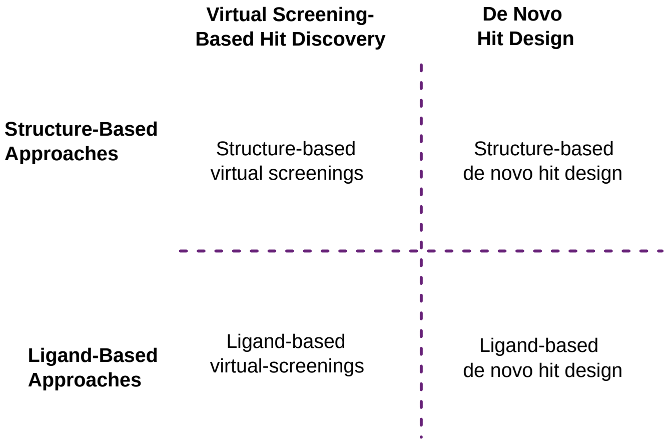
Role in Computer Aided Drug Design

CADD can be divided into three big parts.

(1) Structure preparation,
(2) Hit Identification/Design,
(3) Hit and lead optimization.

Structure preparation refers to the preparation of the receptor structure, which is only relevant in structure-based (i.e. receptor based) drug design. The alternative to receptor-based drug design is ligand-based drug-design, but the former is normally significantly more promising if a receptor structure is available or can be prepared by homology modeling.

The second step, hit identification and design, is all about finding new potential binders to the target of interest. The available approaches can be classified in different ways. There are for instance *virtual screening-based* approaches and *de novo design* approaches. In the former *a priori* prepared compound libraries are screened for their binding affinity, while in the latter completely new compounds are constructed usually on the fly during the screening procedure. Both classification are compatible and their combination is shown in Figure 5.1.



	Virtual Screening-Based Hit Discovery	**De Novo Hit Design**
Structure-Based Approaches	Structure-based virtual screenings	Structure-based de novo hit design
Ligand-Based Approaches	Ligand-based virtual-screenings	Ligand-based de novo hit design

Figure 5.1: Overview of the major classes of the hit identification/design. Structure-based approaches rely on the receptor structure, while ligand-based approaches require only the availability of other known binders for the identification of new potential binders. Structure-based virtual screenings usually rely on molecular docking, but also other free energy methods can in principle be used.

The major advantage of structure-based approaches is that the receptor structure is explicitly taken into account, which in theory allows to obtain more reliable and accurate results than mere ligand-based drug design is normally able to. However, if no receptor structure is available (and cannot be created by homology modeling) then ligand-based approaches are the only possibility, which is their central advantage. De novo based approaches have the advantage that potentially a larger chemical space can be considered as with virtual screening based approaches since normally in each optimization step the compound is further improved by changing single atoms and functional groups. However, the major disadvantage of this approach is that the resulting compounds are often not commercially available, or have even never be synthesized at all. This poses a substantial problem, as the synthesis of a single new compound can be extremely expensive, time consuming, and may at the end even fail. Virtual screening approaches on the other hand don't have to suffer from this circumstance, as one can use as screening collection only commercially available compounds. Therefore structure-based virtual screening procedures provide one of the most promising ways to obtain new hit compounds.

The possibility of carrying out structure-based virtual screenings in order to identify new hit compounds for biomolecular targets was mentioned as early as in the mid 1970s [17, 63], but it has taken a while until sufficient computing power and target structure have become available to make this approach usable for real applications. The turning point at which this has happened might be roughly around the beginning of the current millennium [259].

5.2 The Benefits of Scaling

One variable in virtual screenings is the number of small molecules which are considered. The size of the compound collection (also called compound database) can in theory be any number, even though at least a few thousand compounds are usually included.

One question when considering the upscaling of the compound database which arises is if it does really make sense to go beyond a certain number. For instance one might ask if screening 10 million compounds is expected to have significant benefits over screening 1 million compounds. Possibly 1 million compounds can cover the relevant chemical space already sufficiently so that the remaining space can be explored thoroughly enough during the hit and lead optimization steps to find the best compound possible. To find an answer to these questions it is helpful to consider the entire space of small organic molecules. If we only consider the space of small organic molecules with up to 30 heavy atoms (i.e. non hydrogen atoms) which consist only of the atom types H, C, N and O, then the size of the space is estimated to be at least 10^{60} [31, 146]. If the number of allowed heavy atoms is slightly increased, or other types often found in organic molecules are permitted such as Fl, Cl, Br, or S, then the number will be even larger, possibly reaching 10^{70} or more. This number is hard to imagine, but one might get a feeling on how large it is if one considers the number of atoms in our galaxy, which is estimated to be roughly 10^{67} [47]. In contrast 1 million atoms, or even 10 million, comprises only a tiny fraction of even the smallest bacteria. This comparison helps to realize that even if very large compound sets are considered in virtual screenings, the chemical space explored by them will be vanishingly small. In this light it seems to be favorable to screen as many compounds as possible to find the most favorable candidates attainable, even though it seems out of reach that the truly best compound of the entire chemical space can ever be identified.

5.2.1 Binding Affinities

Given a ligand L and a receptor R, the binding affinity is defined by

$$\Delta A_{\text{bind}} = A_{\text{RL}} - A_{\text{R+L}}, \tag{5.1}$$

where R+L denotes the ligand and the receptor separated from each other and RL denotes the joint ligand-receptor system in the bound state as illustrated in Figure 5.2. In their unbound state the ligand and the receptors are assumed to be so far apart that their interaction can be neglected.

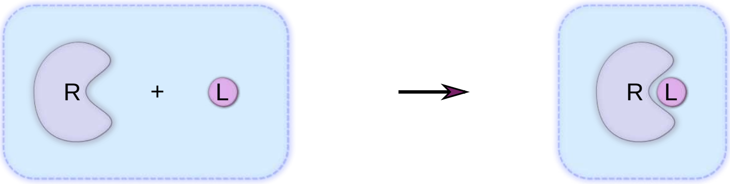


Figure 5.2: The initial and final thermodynamic states relevant in the binding of a ligand to its receptor. In the initial state the ligand and the receptor are assumed to be non-interacting due to their distance.

The more negative the value of the ΔA_{bind} the stronger the binding affinity. We define the *binding strength* ΔB to be the reversed/negative binding affinity, i.e.

$$B := -\Delta A_{\text{bind}}, \tag{5.2}$$

which makes the discussion more simple and naturally. The unit of free energies, and therefore also ΔB_1, is normally expressed as a molar quantity, for instance kcal/mol or kJ/mol (which is an SI derived unit).

The sheer size of the chemical subspace which is screened does not necessarily imply that among its compounds is one which will have a higher binding affinity to a given target. This could for example be the case if the binding strengths would be roughly equally distributed in an interval $[B_{\text{min}}, B_{\text{max}}]$, implying that there would be minimum and maximum values of the affinity due to the laws of physics. However, this is not the case, as becomes for instance clear in [107]. In this article it has been shown that the distribution of the binding affinities of a collection of molecules which have been selected in order to improve the binding affinity of a certain molecule is approximately normally distributed. This indicates that there is no hard limit for the binding affinity, it just gets more and more difficult to further increase the binding strength the higher one goes.

That the above reasoning is not only of theoretical nature but indeed also in reality is beautifully demonstrated by nature itself as described in this paragraph. Many drug design projects start with hit compounds in the micromolar range (i.e. $1\,\mu\text{M} \leq K_{\text{d}} < 1\,\text{mM}$), where K_{d} is the dissociation constant), and the identification of a low nanomolar derivative is often the goal and seen as a major success. Subnanomolar drugs are extremely rare, and this is normally only achieved by antibodies which have a K_{d} normally in the range of medium picomolar to high nanomolar [350]. However, there are notable exceptions. One of them is the biotin-avidin/streptavidin interaction. Biotin is a very small organic molecule of only 244 Dalton (shown in Figure 5.3). But despite its small size, Biotin and several analogs bind noncovalently to their receptors biotin and streptavidin with binding constants of up to 10^{-16} M [168]. In other words in the subfemtomolar range, which is seven orders of magnitudes higher than the binding affinity of low nanomolar binders. The interaction between biotin and avidin is shown in Figure 5.4 on the following page. As can be seen, almost every functional group is engaged in specific interactions, and the molecule is perfectly tailored in almost every

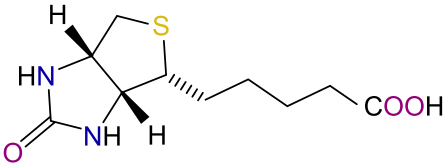


Figure 5.3: Structural formula of biotin, also known as vitamin B_7 or Vitamin H.

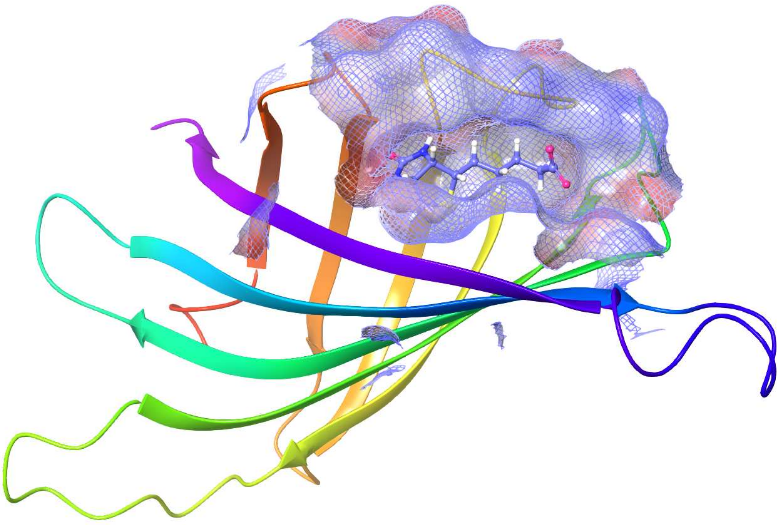


Figure 5.4: Biotin bound to streptavidin, one of the strongest non-covalent interactions known in nature between a small organic molecule and a biological macromolecule. The structure shown represents chain A of the PDB structure 3RY2 which shows the wildtype interaction [170].

aspect for its binding site. However, it is also clear that such an optimal match is very difficult to find.

According to the above discussion the number of small organic compounds is virtually infinite. When using one specific FEM to computationally predict the binding affinity of the compounds in the input collection, we now assume that the predicted binding affinities have the same distribution for each of the compounds in the input collection except for their expectation value μ_i. The more compounds are screened, the higher the predicted binding strengths of the $K \leq N$ highest-scoring compounds will generally be. Since the distribution of the predicted binding strengths does not change (except for its expectation value which becomes higher) with the compound, it is likely that the highest scoring compounds have also in reality a higher binding strength than when less compounds are screened. And in this light it can be expected that the more compounds are screened the higher the real binding strengths will on average be.

Since the probability converges to zero with increasing binding strengths, this means that disproportionally more compounds need to be screened in order to make (on average) the same amount of gain in binding affinity the higher the binding affinity is.

5.2.2 True Hit Rates

The inclusion of more compounds in the input collections in virtual screenings also has positive effects on the true hit rates. For a given virtual screening procedure we define the true hit rate $R_{\text{THR}}(b, K)$ by

$$R_{\text{THR}}(b, K) = \frac{K_N^{\text{real}}}{K}, \tag{5.3}$$

where b is the binding strength threshold which needs to be surpassed for a compound being considered a true hit, K is the number of the highest scoring virtual screening hits which are considered, and K_b^{real} is the number of these which are also hits in reality. With fixed binding strength threshold $b \in \mathbb{R}$ as well as with fixed K the true hit rate will improve with increasing input collection size N. The reason is that the higher N the higher the predicted binding strengths of the top scoring compounds K compounds will be, therefore in turn the larger the distance of estimated binding strengths to the threshold value b. The distribution of the predicted binding strengths of these compounds was assumed to be independent of the precise compound except for the expectation value of the distribution, and the best estimation for the expectation value which we have is the predicted binding strength. Therefore it follows that the higher the predicted binding strength the more likely it is that true binding strength surpasses in reality the threshold b, even if it is lower than the predicted value. Therefore the true hit rate $R_{\text{THR}}(a, K)$ rises on average with the number of compounds screened. This principle is illustrated in Figure 5.5.

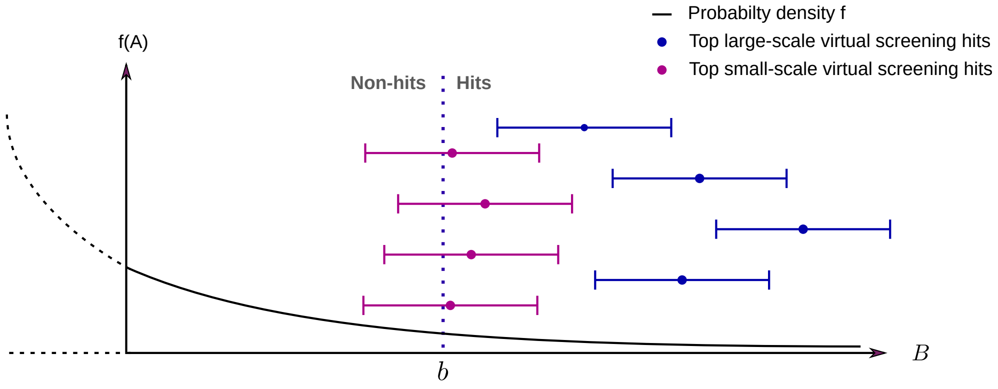

Probabilty density f
● Top large-scale virtual screening hits
● Top small-scale virtual screening hits
f(A)
Non-hits
Hits
b
B

Figure 5.5: Qualitative diagram showing the top hit compounds (dots) and the probability density of obtaining a compound with a certain binding strength (black line). The higher the binding strength the lower in general the probability of finding such a compound. The threshold value defining the above which binding strength the compound is classified as a hit is denoted by b. The four highest scoring compounds are shown are for a virtual screening of smaller scale (berry) and one of larger scale (blue) along with their standard deviation. The larger predicted binding strength the higher the probability that the true binding strength surpasses the threshold b, and therefore also the higher the true hit rate for a fixed number of compounds.

We will briefly look at a more specific example, in which we only consider the single highest scoring compound (i.e. $K = 1$) with predicted binding strength B_1^{virt}. Assuming that the true binding strength B_1^{real} is normally distributed with expectation value B_1^{virt} (i.e. having distribution $\mathcal{N}(B_1^{\text{virt}}, \sigma^2)$) with density f, it follows that the probability of the true binding strength B_1^{real} being higher than the threshold b is in the average given by

$$
\begin{aligned}
R_{\text{THR}}^{\text{ave}}(B_1^{\text{virt}}, b) &= \int_b^\infty f(x; B_1^{\text{virt}}, \sigma^2) dx && (5.4)\\
&= \int_b^\infty \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{\left(x - B_1^{\text{virt}}\right)^2}{2\sigma^2}} dx && (5.5)\\
&= \int_{-\infty}^{B_1^{\text{virt}} + \Delta B_1} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{\left(x - B_1^{\text{virt}}\right)^2}{2\sigma^2}} dx, && (5.6)
\end{aligned}
$$

where

$$\Delta B_1 = \Delta B_1(B_1^{\text{virt}}, b) := B_1^{\text{virt}} - b \qquad (5.7)$$

is the difference between the predicted binding strength and the threshold value. Equation (5.6) follows by the symmetry of the normal distribution. The integral is thus given by

$$
\begin{aligned}
R_{\text{THR}}^{\text{ave}}(B_1^{\text{virt}}, b) &= \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{B_1^{\text{virt}} + \Delta B_1 - B_1^{\text{virt}}}{\sigma\sqrt{2}}\right)\right) && (5.8)\\
&= \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{\Delta B_1}{\sigma\sqrt{2}}\right)\right), && (5.9)
\end{aligned}
$$

where erf is the error function. $R_{\text{THR}}^{\text{ave}}$ is plotted in Figure 5.6 for different values of the standard deviation σ, which clearly shows the favorable dependence of the average true hit rate of the highest scoring compound on ΔB_1. Fast docking functions which are normally used in virtual screenings have generally a high standard deviation, making large ΔB_1 values particularly favorable. The positive effect on the true hit rate of K top compounds rather than only one will be similarly significant.

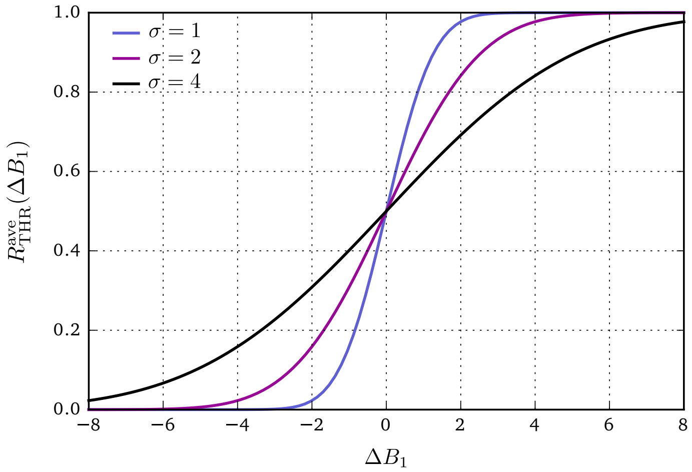


Figure 5.6: The average true hit rate for the single highest scoring compound, $R_{\text{THR}}^{\text{ave}}$, is shown in dependency of $\Delta B_1(B_1^{\text{virt}}, b)$ in kcal/mol for different values of σ.

In conclusion, the more compounds are screened the higher the average expected binding strengths of top compounds, and moreover the higher also the average hit rates. Which in turn means the more true hit compounds (for a fixed number of compounds which we can experimentally verify), which corresponds to goal (2) in the introductory section 5.1 on page 91.

5.3 Multistaging Techniques

The more accurate the FEM which is used for the virtual screening, the more reliable the results naturally are. However, the more accurate the method the more computationally demanding it usually is, and therefore it is mostly far beyond the possibilities to apply these more expensive methods to all compounds which should be screened. However, there is an elegant approach out of this dilemma which allows one to

(1) screen all of the desired input compounds,

(2) obtain free energy estimates by the desired computationally expensive method,

(3) not require substantially more computationally resources than when only the cheap method would be used.

The principle which makes these miraculously sounding combination of features possible is called *multistaging* or *funnel approach* [65, 232].

When using this technique one runs multiple virtual screenings in serial, and only allows a certain percentage of the top scoring compounds to proceed to the next one. The first virtual screening is the *primary virtual screening*, and the other ones are also called *rescoring procedures* since the compounds which made it to these later stages have already passed the first scoring in the primary virtual screening. One can imagine this principle as a multistage funnel, in which less and less compounds pass through each refinement level. The key idea is now to use different FEMs for each refinement procedure, and since with each stage less and less compounds are left one can apply computationally more and more expensive FEMs. The principle is illustrated in Figure 5.7 on the following page. The approach is not fully equivalent to applying the most accurate FEM to all compounds, since in each stage a fraction of the true binders might be filtered out incorrectly. But if the winning sets of each stage, i.e. the compounds which are transferred to the next level, are large enough, then it is often possible to design the procedure that the majority of them lands in the final winning set. As an example, if we screen 10 million compounds, then we could design a three-stage procedure of the form

(1) Primary virtual screening: 10 million compounds,

(2) First rescoring: 200 000 compounds,

(3) Second rescoring: 4 000 compounds.

In this example in each of the first two steps the number of compounds is reduced by a factor of 50, and when this factor is also applied in the last step 80 compounds are left in the final winning set. These could be experimentally tested for their real binding affinities for verification purposes. If we assume that in each step 80 % of the ten top scoring compounds make it to the next level at each step, then in the final winning set

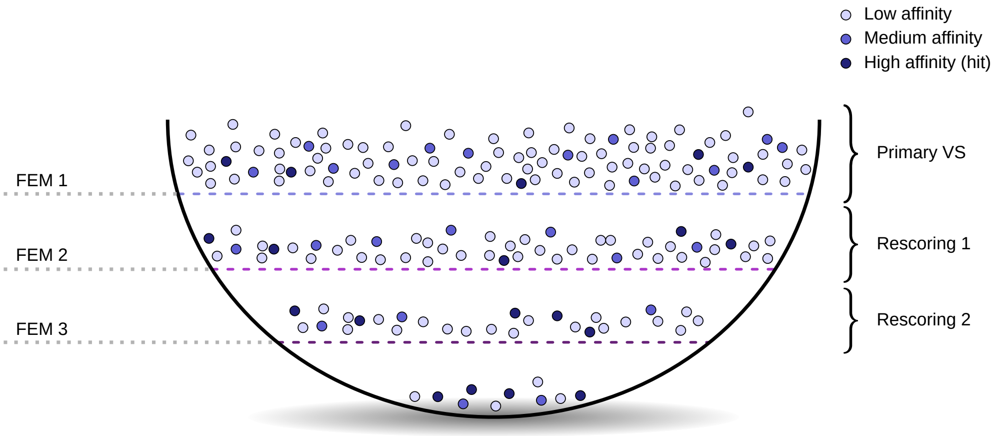


Figure 5.7: Multistage virtual screening layout with three levels. In each level a computationally more expensive but also more accurate/precise method is applied. A certain number of the top scoring compounds is then transferred to the next level, and the reduction of total number allows to carry out a more expensive free energy method. In this way it becomes possible to screen a larger number of compounds at the level of the most expensive free energy method, even though there is the possibility that a certain fraction of the top scoring compounds is filtered out during the refinement steps.

there should be roughly five of them left. In comparison, if we would have only used the primary virtual screening, then approximately 80 compounds would have made it into the top 200 000 compounds, but due to the low quality of this method is not sure that even a single would have made it into the top 80 compounds which are to be tested experimentally. The rank correlation of scoring function can be relatively low, meaning that the compound ranking is strongly deranged from the correct order and that many false positives could have made it into the top region. From this perspective it becomes clear that the utilization of the multistage procedure can provide a substantial increase in the binding affinity of compounds as well as in the true hit rate.

5.4 Combined Approaches

In this chapter we have looked at alternative approaches to improve the quality of structure-based virtual screenings besides improving the FEM itself which is employed during the procedure. The first approach is to scale up the virtual screening with respect to the number of compounds which are tested, and the second approach uses a multistage procedure. A further advantage of these two paradigms is that they are fully compatible with each other, making it possible to apply both at the same time. The higher the number of compounds in the beginning, the more compounds are carried over to the next stages (if the transfer ratio remains the same). And since the upscaling increases the average binding affinity of the top scoring compounds, and with that the true hit rate of them, this will also be the case for the final winning set. We might still loose a similar fraction of the highest scoring compounds during the rescoring-based filterings (similarly as in the example on page on the previous page), but the remaining true hit compounds with the true highest binding will (on average) be stronger binders. Moreover, the other compounds which are in the final winning set which are not the very top scoring compounds have a higher likelihood of still being true hits (since the true hit rate has increased).

5.4.1 Unification with QUASAR and QSTAR

Large-scale multistage structure-based virtual screenings are already combining two of the three major approaches mentioned in the introduction 5.1 on page 91 for achieving the primary goals of high affinity compounds and favorable true hit rates. Besides upscaling and multistaging, the third and most direct approach to achieve the desired goals is to use (or develop) improved FEM as this does essentially tackle the root cause of the arising problems. And this has already been done in this thesis. A new FES method, QSTAR, was developed in section (4.3) which is of extreme force-field accuracy. And while it is not possible to apply QSTAR directly in plain virtual screenings due to the computational expense, it can be used in multistage procedures as the very last and most accurate step, possibly also making use of the PEARL scheme (introduced in Subsection 4.4.2). It is therefore possible to apply all three major approaches to improving the quality of the results at the same time. To complete the picture, in Section 3.4, QUASAR a new QM/MM scheme for diffusive systems, can be used in combination with QSTAR, and therefore become an integral part of multistage virtual screenings as well.

Since QSTAR has not only a single level of accuracy, but a wide range of choices regarding the potential, options and parameters, the method can not only be applied a single time in the last stage, but multiple times in different stages. For instance QSTAR could be applied on the MM level (modeling entire atoms on the path integral level) in the second last step, and on the QM/MM level with QUASAR in the final stage (see also Figure 5.8).

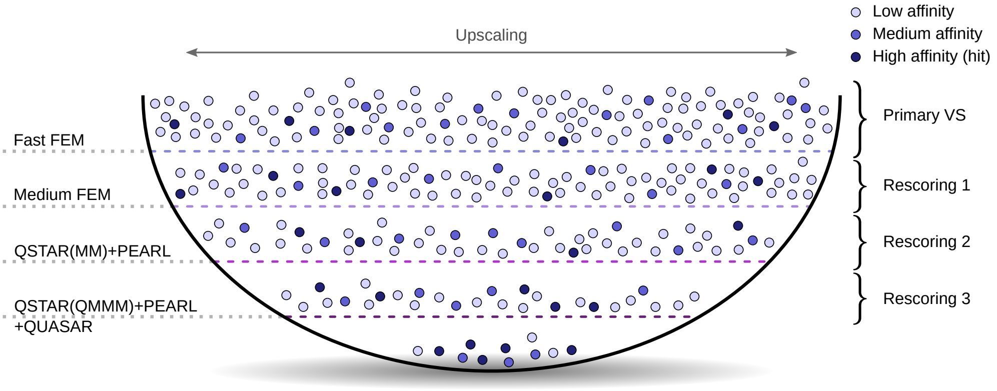


Figure 5.8: Unification of all three major approaches to improving the quality of virtual screenings: Upscaling, multistaging, and utilization of a highly accurate free energy method in the form of QSTAR in combination with PEARL and QUASAR, i.e. the new methods developed earlier in this thesis. The upscaling is represented by the increased horizontal dimension of the funnel. In the first level a fast free energy method based on docking is used. For the second level a more elaborate but less still relatively fast method is most suitable, such as flexible-receptor docking methods. In the third level QSTAR is used at the MM level, possibly with PEARL as the alchemical scheme. And in the fourth level QSTAR is used again on the QMMM level in concert with QUASAR.

5.5 Conclusions and Outlook

In this chapter the benefits of upscaling, multistaging, and improved FEMs within virtual screenings were described, and it was outlined how the newly developed QUASAR QM/MM scheme as well as QSTAR can be used in multistaging procedures. Combining all of these three approaches at the same time might provide one of the most sophisticated and promising approaches for discovering new hit and lead compounds in CADD.

While some software tools already existed for carrying out the first two approaches, the available options seemed to be somewhat limited and unfavorable. Therefore we have developed a new workflow suite, Virtual Flow, which is dedicated to these two approaches. It is described in detail in chapter 8 on page 149.

Also the QSTAR approach was implemented and embedded in a workflow suite called HyperQ, which is able to process large numbers of compounds in an automatic way (see chapter 7 on page 117 for more details).

The QUASAR QM/MM method was implemented in a separate software package called i-QI, which is able to work in concert with HyperQ to carry out the desired FES. It is described in chapter 6 on page 105.

Part II

Model and Method Implementations

Chapter 6

i-QI

Physics is to be regarded not so much as the study of something a priori given, but rather as the development of methods of ordering and surveying human experience.

Niels Bohr
The Unity of Human Knowledge

Contents

6.1 Introduction

In section 3.4 on page 50 the QUASAR method was introduced, which is a QM/MM method for diffusive systems based on restraining potentials. Its key advantages are that the form of the QM/MM region can be nearly arbitrary, and as a restraining method which can preserve thermodynamic ensembles it is particularly suitable in the context of free energy simulations, which was the motivating application for its development.

QUASAR was implemented in a dedicated software package named i-QI. The goal of the new package was not only to provide the possibility to carry out QM/MM simulations with the QUASAR scheme, but in a way which allows to carry out these multiscale simulations in concert with PIMD simulations. The de facto standard code for running PIMD simulations is i-PI, and i-QI has therefore been designed to be a standalone client which can connect to i-PI running as a server.

These aspects are the reason why a new package was created rather than extending one of the existing available code. Among the programs which are already able to carry out QM/MM simulations are CHARMM [332], AMBER [103], QuanPol [289], GROMACS [158], MOPAC [269], NWChem [301], CP2K [166, 165], ChemShell [199], Discovery Studio [27], QMMM [338], QSite/Jaguar [249], Gaussian [94], CPMD [39, 167, 92], LICHEM [155], and since recently also NAMD [210].

i-QI is released under the GNU General Public License v3.0 and freely available under `https://github.com/cgorgulla/i-QI`.

6.2 Overview

6.2.1 i-PI Architecture and Features

The simulation software i-PI is the first sophisticated software which is dedicated to PIMD, published in 2014 in [50]. i-PI was designed from the beginning in a neat client-server layout in which i-PI acts as the server and clients can connect to it during its runtime. This layout is shown in Figure 6.1.

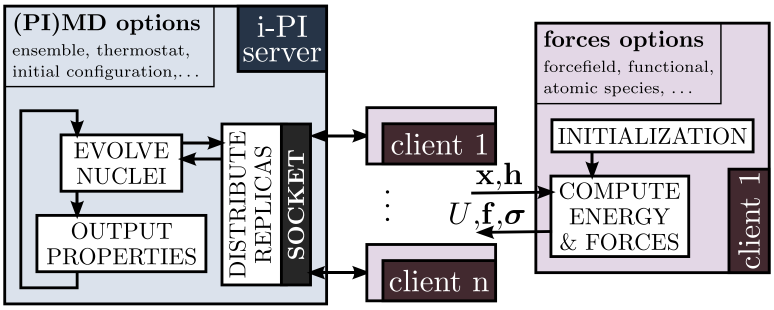


Figure 6.1: Overview of the client-server architecture of i-PI. Multiple clients can connect to the server (i-PI), and the communication happens via sockets in a minimalistic manner. Adapted with permission from [50].

The design is very clean and the roles of the server, as well the clients, clearly predefined. i-PI is handling all the path-integral specific issues, which includes the motion of the particles and the properties of the system. The clients only evaluate the potential energies $V(\boldsymbol{r}_k^{(1)}, ..., \boldsymbol{r}_k^{(N)})$ for the $k \in \{1, ..., P\}$ beads (see equation 2.116) and return these, as well as the forces on each bead , to the server. The harmonic springs which connect the beads are modeled by i-PI. This design has the tremendous advantage that i-PI can principally work together with any existing MM or electronic structure code available. The only modification which is needed on the clients is the implementation of the interface to i-PI and a corresponding runtime mode which outsources the integration of equation of motions, and accepts instead the feed of new coordinates from i-PI.

Another advantage of this design is that it is trivially parallelizable over the number of path integral beads since at each time point there are P of them which are independent (as only beads interact which are present at the same imaginary time). And i-PI makes this parallelization possible by providing the possibility to connect multiple clients at the same time which it distributes among the available beads.

i-PI supports many advanced features related to PIMD. Among them are the following:

- Ring Polymer Molecular Dynamics (RPMD)
- Centroid Molecular Dynamics (CMD)
- Ring Polymer Contraction (RPC)
- Particle Nomentum Distribution Displaced Path Estimator
- Path-Integral Langevin Equation Thermostats
- Path Integrals at Constant Pressure
- Path Integral Generalized Langevin Equations Techniques

- Finite-differences Suzuki-Chin PIMD (experimental)
- Reweighting-based high-order PIMD (experimental)
- Perturbed Path Integrals (experimental)
- Geometry Optimization (experimental)
- Finite-differences Vibrational Analysis (experimental)
- Multiple Time Step Integration (MTS) (experimental)
- Direct Estimators for Isotope Fractionation (experimental)
- Langevin Sampling for Noisy or Dissipative Forces (experimental)

A few minor modifications were made within i-PI to solve some issues such as a problem with the atomic masses when reading in PDB files. For this purpose the i-PI repository was cloned and made available under `https://github.com/cgorgulla/ipi-qmmm`.[1]

6.2.2 i-QI Layout

The QM/MM client i-QI is implement in python (as is the case for i-PI) in an object-oriented design, and the design of the code is inspired to some extend by the structure of i-PI. The class diagram of i-QI is shown in Figure 6.2. The design is such that it can be easily extended to other types of potentials than QUASAR, as well as additional types of constraints within the QUASAR scheme other than spheres.

A few files of i-QI are inspired or based on the corresponding files of i-PI. This is mainly the case for the I/O related files `xml_io.py` and `messages.py`.

6.3 Properties and Implementation

6.3.1 Constraints

The design of the constraint specification and handling is such that it easily allows to extend the software to additional types of constraints. In the current version an arbitrary number of spherical constraints can be specified. Constraints are specified in two steps. In the main configuration file (in XML format) the filename of the constraints file is specified, as shown exemplary in 6.1.

Listing 6.1: Specification of the constraints subconfiguration file within the i-QI main configuration file, both in XML format.

```
<constraints>
  <file type="xml">constraints.xml</file>
</constraints>
```

The spheres can be specified via simple XML tags as well in the subconfiguration file dedicated to the constraints (also in the XML format), in which the only parameter

[1] This separate repository is intended only as a temporary solution and the changes will probably merged back into the original repository of i-PI.

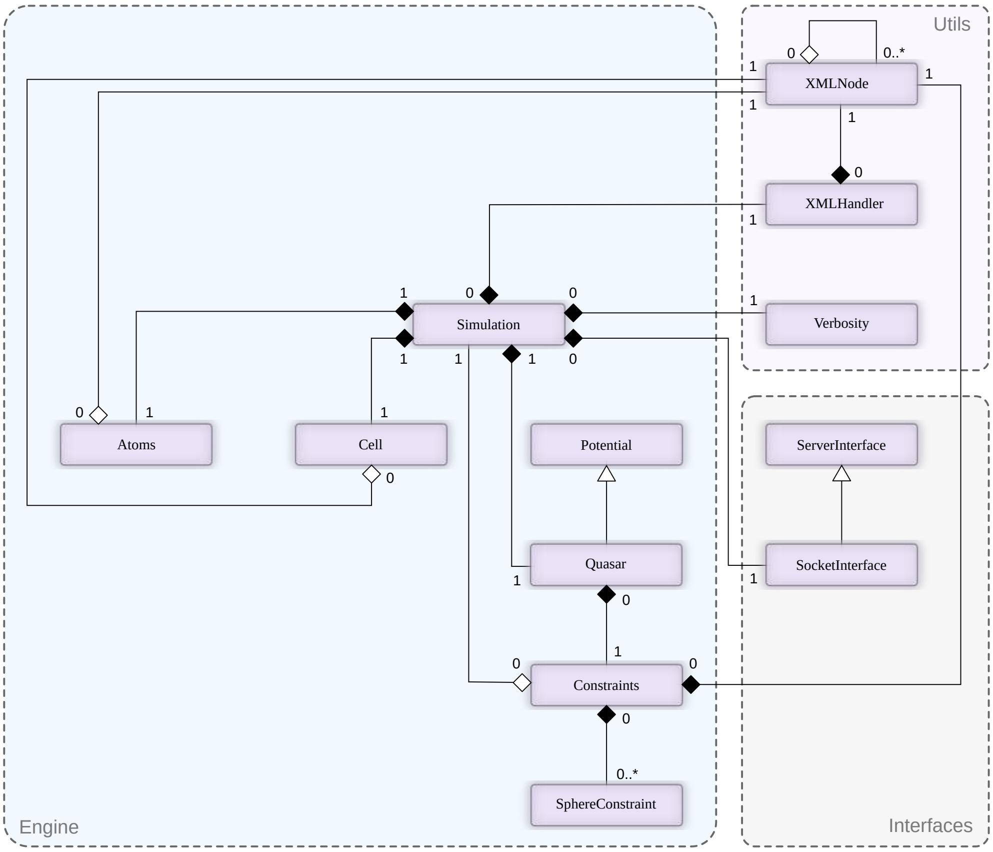


Figure 6.2: Class diagram showing the relationship between various classes/objects in UML format [213] (white diamonds denote aggregation, filled diamonds composition, numbers multiplicity, and white triangles inheritance). The *Simulation* object is the central (root) object, and the input data is stored in a tree of XML nodes.

required per sphere is the atom index at which the sphere is to be centered. An example is shown in Listing 6.2.

Listing 6.2: Specification of the spherical constraints within the i-QI constraints subconfiguration file (example).

```
<spheres>
  <sphere central_atom_id="3"/>
  <sphere central_atom_id="5"/>
  <sphere central_atom_id="9"/>
</spheres>
```

Generally also other basic shapes might be useful in practice, such cuboids or dividing planes, which might be implemented in the future.

6.3.2 PDBX File Format

Within the context of the QUASAR method atom types can be classified as either

(1) inside the adaptive spherical assembly constraint (QM level),

(2) outside the adaptive spherical assembly constraint (MM level).

Moreover, not all atoms have to be in the diffusive phase, which allows to classify them as being

(1) in the diffusive phase (constrained),

(2) in the non-diffusive phase (uncontrained).

Since these two classification schemes are independent there is a total of four atom types when using QUASAR. And i-QI needs to know for every atom of which type it is since it needs to compute the restraining forces on the restricted atoms only. Since our major target application are biomolecular simulation, and since i-PI can read in Protein Data Bank (PDB) files [22], we have introduced an extension of the PDB data format, referred to as the Protein Data Bank Extended (PDBX) file format. This format is exactly identical in its specifications as the PDB format, but extends it with additional fields to the ATOM and HETATM records. The original ATOM/HETATM records have the following format:

Listing 6.3: ATOM record specification of the original PDB file format. ATOM record specification of the original PDB file format. The record specification for the HETATM record is analogous. Source: [339]

```
COLUMNS        DATA  TYPE    FIELD        DEFINITION
-------------------------------------------------------------------------------------
 1 -  6        Record name   "ATOM  "
 7 - 11        Integer       serial       Atom  serial number.
13 - 16        Atom          name         Atom name.
17             Character     altLoc       Alternate location indicator.
18 - 20        Residue name  resName      Residue name.
22             Character     chainID      Chain identifier.
23 - 26        Integer       resSeq       Residue sequence number.
27             AChar         iCode        Code for insertion of residues.
31 - 38        Real(8.3)     x            Orthogonal coordinates for X in Angstroms.
39 - 46        Real(8.3)     y            Orthogonal coordinates for Y in Angstroms.
47 - 54        Real(8.3)     z            Orthogonal coordinates for Z in Angstroms.
55 - 60        Real(6.2)     occupancy    Occupancy.
61 - 66        Real(6.2)     tempFactor   Temperature  factor.
77 - 78        LString(2)    element      Element symbol, right-justified.
79 - 80        LString(2)    charge       Charge  on the atom.
```

The extension added by the PDBX file format are the specified in Listing 6.4.

Listing 6.4: PDBX ATOM (and HETATM) specification on top of the PDB file format. Two additional columns are added which are used by i-QI to classify the atoms in the context of the QUASAR method.

```
COLUMNS        DATA  TYPE    FIELD        DEFINITION
-------------------------------------------------------------------------------------
81             Atom          type         QUASAR phase identifier
82             Atom          type         QUASAR region identifier
```

The *QUASAR phase identifier* specifies whether the atom is the diffusive phase, i.e. in the constraint phase, which is specified by the letter 'C' in the PDBX file. If it is in the unconstrained phase the character 'U' is used. Normally biological macromolecules such as proteins are in the unconstrained phase, while the solvent, and possibly other small molecule, are in the diffusive phase. The *QUASAR region identifier* specifies whether the atom is outside the spherical assembly constraint (normally on the MM level), which is denoted by 'M'. If the atom is inside the constraint (normally on the QM level the character 'Q' is used). A short snipped of an example PDBX is shown below.

Listing 6.5: An example excerpt of a PDBX file, the file format which is used by i-QI to classify the atoms by their QUASAR types. In the QM phase a part of a protein is located, as well as a ligand binding to it and surrounding water molecules. While the protein is completely in the unrestricted phase, the solvent molecules and the ligand are in the constrained phase.

```
...
ATOM   2868  O    GLY R 208      21.829   3.554   0.121  1.00  0.00      RCP  O  MU
ATOM   2869  C    GLY R 208      22.119   3.899   1.299  1.00  0.00      RCP  C  MU
ATOM   2872  CA   GLY R 208      21.951   2.979   2.471  1.00  0.00      RCP  C  QU
ATOM   2873  HA1 GLY R 208      20.908   2.701   2.532  1.00  0.00      RCP  H  QU
ATOM   2874  HA2 GLY R 208      22.626   2.144   2.339  1.00  0.00      RCP  H  QU
ATOM   2870  N    GLY R 208      22.310   3.681   3.679  1.00  0.00      RCP  N  QU
ATOM   2871  HN   GLY R 208      21.572   4.061   4.227  1.00  0.00      RCP  H  QU
...
ATOM   3027  C1Q LIG L   1      -1.254   0.099   1.926  1.00  0.00      LIG  C  QC
ATOM   3028  C2Q LIG L   1      -0.999   0.679   3.185  1.00  0.00      LIG  C  QC
ATOM   3029  C3Q LIG L   1      -0.866   2.075   3.353  1.00  0.00      LIG  C  QC
ATOM   3030  C4Q LIG L   1      -0.964   2.944   2.261  1.00  0.00      LIG  C  QC
ATOM   3031  C5Q LIG L   1      -1.150   2.368   0.997  1.00  0.00      LIG  C  QC
ATOM   3032  C6Q LIG L   1      -1.296   0.981   0.834  1.00  0.00      LIG  C  QC
ATOM   3033  C7Q LIG L   1      -2.040  -4.148   1.611  1.00  0.00      LIG  C  QC
...
ATOM   3071  OH2 TIP3W   4      19.718 -23.086 -17.685  1.00  0.00      WT1  O  MC
ATOM   3072  H1  TIP3W   4      19.925 -22.909 -18.648  1.00  0.00      WT1  H  MC
ATOM   3073  H2  TIP3W   4      18.787 -22.802 -17.488  1.00  0.00      WT1  H  MC
...
ATOM   3080  OH2 TIP3W   7     -25.651 -25.332 -10.186  1.00  0.00      WT1  O  QC
ATOM   3081  H1  TIP3W   7     -25.606 -24.984  -9.279  1.00  0.00      WT1  H  QC
ATOM   3082  H2  TIP3W   7     -25.092 -26.120 -10.116  1.00  0.00      WT1  H  QC
...
```

The major advantage of the PDBX format is that PDB files are usually needed and available in any case in the context of biomolecular simulations, and therefore the PDBX files can be easily prepared by augmenting the PDB versions.

6.3.3 Uniform Force Distribution Mode

In Subsection 3.4.2 in which the QUASAR method was introduced the uniform force distribution mode was described. In this mode the forces due to the restraining potentials are not atom-wise, but molecule wise. In the configuration file this option is specified within the potential declaration, as is illustrated exemplary in Listing 6.6.

Listing 6.6: Excerpt of the main i-QI configuration file showing the specification of a the QUASAR potential with the uniform force distribution mode (with respect to the molecules).

```
<potential type="QUASAR">
  <constraints>
    <file type="xml">constraints.xml</file>
  </constraints>
  <forceconstant>0.001</forceconstant>
  <force_distribution>molecule</force_distribution>
</potential>
```

In order to be able to operate in the uniform-force-distribution mode i-QI it needs to know which atoms belong to which molecule, and it is able to determine this by using the chain and residue information of the PDBX input file.

6.3.4 Interfaces

i-QI can communicate with i-PI either via POSIX local inter-process communication sockets (IPC sockets) or alternatively via internet sockets. These two ways are currently the only possibilities which i-PI currently provides. One advantage of IPC sockets is that they are faster than internet sockets, but the latter have the advantage that they allow communication across different computers in local or remote networks. For example when using high performance computer clusters the possibility to run different clients on different nodes which are all connected to the same i-PI server can be very valuable.

The socket interface between i-PI and its clients is of a minimalistic nature in the sense that only the absolutely essential information is exchanged. i-PI sends to its clients the following data (per MD step):

(1) The current size.
(2) The positions of the atoms.

And clients, after having completed the energy and force evaluations, send back to i-PI

(1) the total energy,
(2) the forces on each atom,
(3) the pressure virial tensor,
(4) optional extra information.

The extra information is only needed in special cases, and i-QI does not require this feature.

An example of how the interface can be set up in the main configuration file of i-QI is shown in Listing 6.7.

Listing 6.7: Example interface configuration within the i-QI main configuration file.

```
<interface type="socket">
  <socket type="inet">
    <address>localhost</address>
    <port>31415</port>
  </socket>
</interface>
```

6.4 Test Simulations

To see if the QUASAR method is working as predicted, and to verify its implementation in i-QI, test simulations were carried out.

The Test System

For the test system a drug-like molecule was chosen to be solvated in a water box. Such a molecule is suitable for our verification purposes in particular because its elongated shape allows the placement of multiple restraining spheres along its scaffold. Moreover it has polar atoms which allows the formation of more complex water structures around the solute, allowing for a more refined verification. The (isomeric) Simplified Molecular Input Line Entry Specification (SMILES) of the compound is given by

```
c1ccc(-c2ccc(-c3nc([C@H]4CCOC4)no3)cc2)cc1
```

,

which corresponds to the structure shown in Figure 6.3.

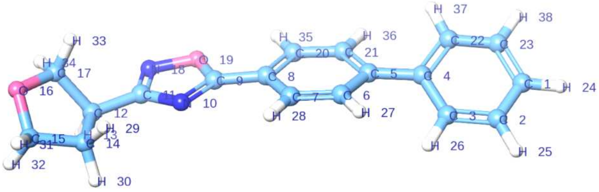


Figure 6.3: The organic compounds which serves as the solute in the test simulations involving QUASAR and i-QI. The (non-isomeric) IUPAC name of the compound is *5-(4-Biphenylyl)-3-(tetrahydro-3-furanyl)-1,2,4-oxadiazole*. The compound has a stereocenter at atom C12 belonging to the tetrahydrofuran ring on the left side. The numbers indicate the atom indices.

The waterbox in which the compound was dissolved was created such that it is cuboid and has a padding size of 15 Å from each boundary wall to the solute. This has resulted in 1594 water molecules, giving rise to a total number of 4820 atoms in the entire system including the solute.

Simulation Details

The system is modeled uniformly on the MM level since this the purpose of the simulations are to compare the ensemble under unrestricted conditions and under the presence of the QUASAR restraining potentials. The CHARMM c36 force field was used for the entire system. The solute was parametrized with CGenFF force field [304, 348] by the corresponding program of its developers [306, 307], and the water model is TIP3P.

At first an energy minimization was carried out with CP2K consisting of 500 optimization steps. Afterwards an NPT simulation was carried out with CP2K to allow the volume to equilibrate. And finally the production NVT simulations were run with i-PI/CP2K, and in the QUASAR simulations i-QI was used in addition. The temperature of 298 K was maintained by a Langevin thermostat with a friction coefficient of $\gamma = 5\ \mathrm{ps}^{-1}$, and the integration time step was 1 fs.

Three restraining spheres were employed for creating the QUASAR restraining potential. They were placed at the atoms with the indices 1 (C), 8 (C) and 15 (C), defining spheres 1, 2, and 3, respectively. The atom which were selected to be included in the adaptive spherical assembly restraint were selected by the criteria

(i) all atoms within 7 Å of the atoms with serial numbers 1, 8, 15 of the solute,
(ii) all atoms which belong to the same molecules as the atoms of condition (i).

The adaptive sphere assembly was rendered in Figure 6.4 for a specific timepoint of the dynamics. Condition (ii) above is important since otherwise the atom would be trapped between the inner and the outer region and experience forces into opposing directions, and it can also increase the initial maximum distance of 7 Å to up to 9 or 10 Å. The above criteria resulted in 92 water molecules to be added to the QM region $\mathcal{Q}$ (see section 3.3 on page 46 for the notation, in this case it is modeled on the MM level).

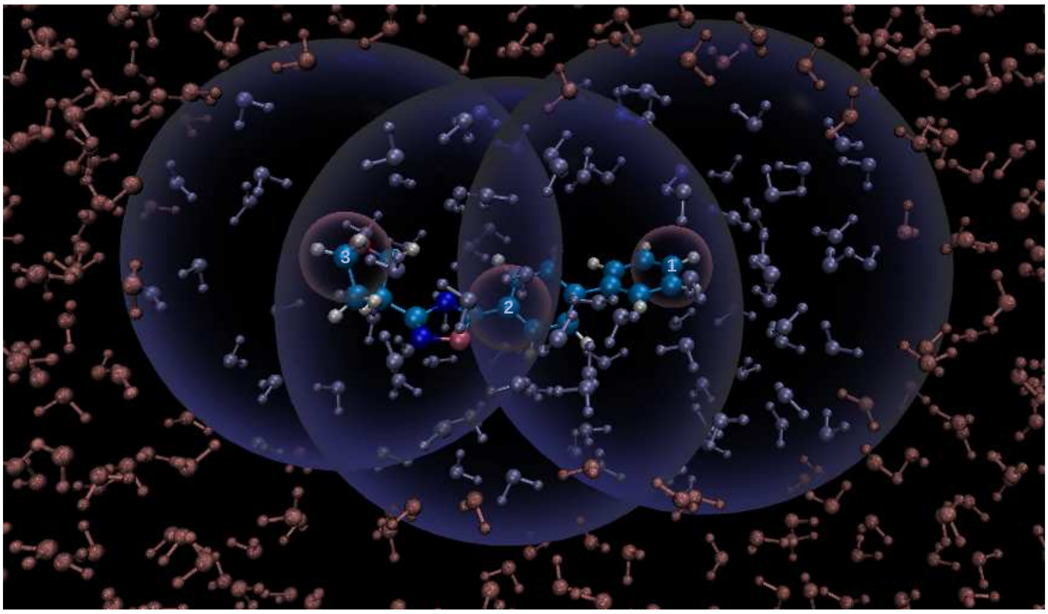


Figure 6.4: The solute (atoms colored by element, carbons here in light blue) of the test system is shown together with the QUASAR constraint, consisting of three spheres (transparent dark blue). The central anchoring atoms of the spheres can be recognized by their mauve-colored halos as well as the sphere indices in their center. The solvent molecules which are inside the QUASAR constraint are all visible (violet), while of the outer solvent molecules (mauve colored) only those are shown which lie near the frontal plane (i.e. the plane facing the viewing point) which runs through the solute. The spherical constraints have non-uniform sizes, are adapting to the position of the particles, and particles can freely pass from sphere to sphere.

Results and Analysis

Regarding the QUASAR method the central property to verify is whether or not it changes the thermodynamic ensemble. And in relation to I-QI it needs to be checked if the atoms which are supposed to be separated by the adaptive sphere assembly are indeed separated, i.e the atoms in $\mathcal{Q}$ remain within the inner region of the constrained region and the atoms in $\mathcal{M}$ outside during the time evolution. Both can be verified by investigating certain radial distribution functions (RDFs) of the system. Since the QUASAR boundary is in the solvent center around the solute, of particular interest are the RDFs which are centered at the atoms of the solutes.

It is relatively simple to demonstrate that the structure of the solvent does not change significantly when using the QUASAR method when choosing appropriate circumstances. As an example, one can include a large number of atoms within the QM region so that the solute is far away from the boundary and the influence of the boundary becomes negligible on small organic solutes. Instead we will look here at a system with a setup in which unnatural influences due to the QUASAR boundary are more likely to happen, in order to point out a potential sources of errors when using this method. In the current test system the QUASAR spheres were chosen to be very small, containing on average only around two solvent layers between the solute and the QUASAR boundary, implying that often during the dynamics only a single solvent shell will exist between them. With such small spheres it was possible to see changes in the solvent structure caused by the QUASAR method. This was for example the case with the O16(solute)-O(solvent) RDFs, which are shown in Figure 6.5 for both the reference system (no restraints) and the QUASAR system (the latter was scaled to the former in order to make them more easily directly comparable). While the match between the two RDFs is still relatively good, the first water shell around the central atom (O16) of the solute of the RDF has an increased peak height. This seems to imply that the QUASAR method stabilizes hydrogen bonds between the solute and the solvent. this seems plausible since the close restraining potential restricts the movement capabilities of the nearby solvent molecules. Therefore it might be more difficult to break already formed hydrogen bonds. But also if they break indeed and the solvent molecule moves away from the solute it could be reflected back from the boundary and caused to directly form again a new hydrogen bond with the solute. It follows that if the water structure is important close to the solute, the QUASAR spheres should contain a sufficient amount of many solvent molecules so that multiple solvent layers are on average between the solute and the restraining potential. In practice three or four layers should be sufficient.

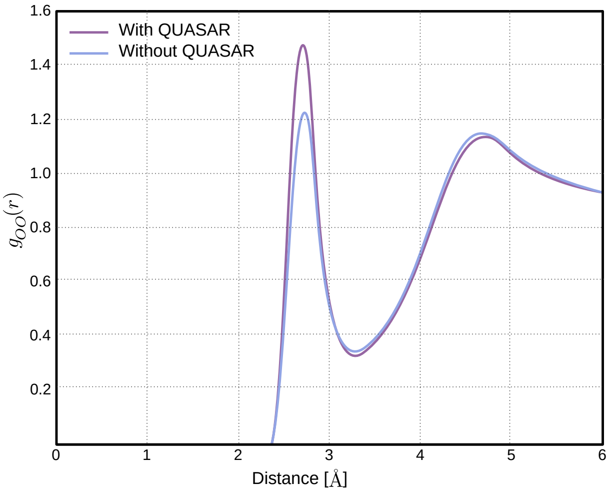


Figure 6.5: Radial distribution function between the oxygen atom (O16) of the solute compound and the oxygen atoms of the solvent molecules.

The distribution of the radii of the three spheres which were used in the above test system is shown in Figure 6.6. As can be seen in the figure the radii are approximately

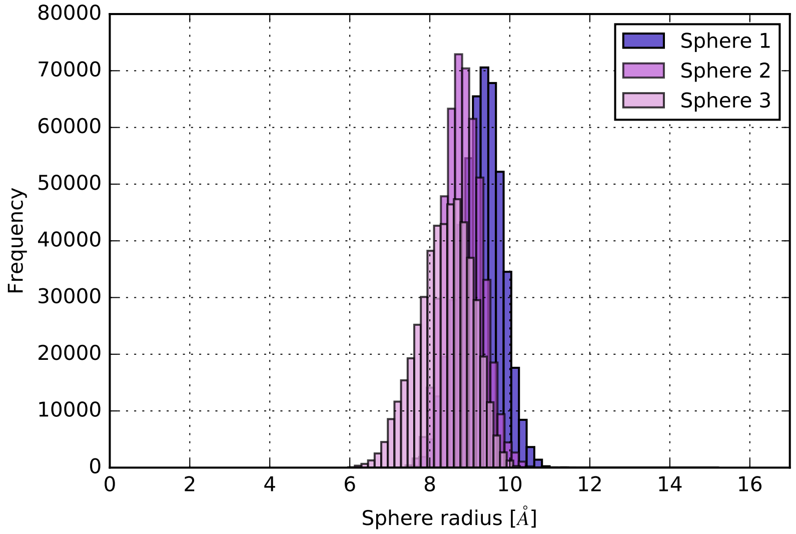


Figure 6.6: Distribution of the radii of the three QUASAR spheres of the simulations of the test compound (*5-(4-Biphenylyl)-3-(tetrahydro-3-furanyl)-1,2,4-oxadiazole*) within a water box.

Gaussian distributed. They are also moderately different from each other, which is not unexpected since the spheres are located at different regions of the molecule which have a different atomic structure.

Chapter 7

HyperQ

There is no scientific discoverer, no poet, no painter, no musician, who will not tell you that he found ready made his discovery or poem or picture – that it came to him from outside, and that he did not consciously create it from within.

William K. Clifford

Letter to the Royal Institution

Contents

7.1 Introduction

In section 4.3 on page 69 the QSTAR method has been presented for computing alchemical free energy differences. A new FES suite was developed, called HyperQ, with the primary purpose of implementing the QSTAR method. HyperQ was designed with the purpose to make it as simple, convenient, and practically useful as possible for scientists working in CADD to use and apply it. This purpose has lead to the following principle design features:

(1) **Binding Free Energy support.** Special support for small organic molecules (ligands) binding to a receptor structure (target), in particular proteins.

(2) **Highly Automatized.** Providing possibilities to automatically carry out the following tasks:

  (a) Preparation of required input files.

  (b) Running the FESs.

  (c) Postprocessing of FES output files.

(3) **Flexibility** Being highly customizable in an optional fashion.

(4) **Scalability.** The workflow suite should have maximum support for paralellization.

The automatization aspect is very important because the complexity of the alchemical QSTAR simulations involving path integrals can be daunting and is hardly manageable manually. In particular when one needs to run simulations of different systems or ligands the time and work which automatic procedures can save are dramatic. Automatization also reduces possible human errors which can occur when preparing and running the simulations manually. In addition, the simulations become more easily reproducable. And also many scientists working in the field of CADD who would like to use the new QSTAR method might lack the knowledge on how to carry out the required preparation steps manually.

While being specialized on binding free energies, HyperQ can principally be used for any type of alchemical transformation, such as solvation free energies or distribution coefficients (logD) as long as the required input files are provided (mainly the MM force field files).

7.2 Overview

The HyperQ[1] pipeline consists of six major procedural components:

(1) Structure preparation

(2) Geometry optimization

(3) Equilibration simulation

(4) Production (PIMD) simulation

(5) Energy cross evaluation

(6) Free energy computation

These steps are carried out consecutively, with the equilibration step being optional. The starting point of the pipeline are the initial input files. The general workflow and the flow of information is illustrated in Figure 7.1 on the next page.

Details on the individual components are described in Section 7.4.

7.2.1 Features

- **Extended QSTAR Support**. QSTAR is supported not only in its basic form, but in certain variants of its extended form. Most of the schemes which have been presented in section 4.4 on page 76 are supported, including the ι-schemes and ν-schemes which are the two primary types of η-schemes. Regarding the σ-schemes the full PEARL approach has been implemented, including PEARL-P,PEARL-N, PEARL-N, and PEARL-XH.

[1] HyperQ is written in Bash and Python, released under the GNU GPL v3, and available on GitHub via `https://github.com/cgorgulla/HyperQ`.

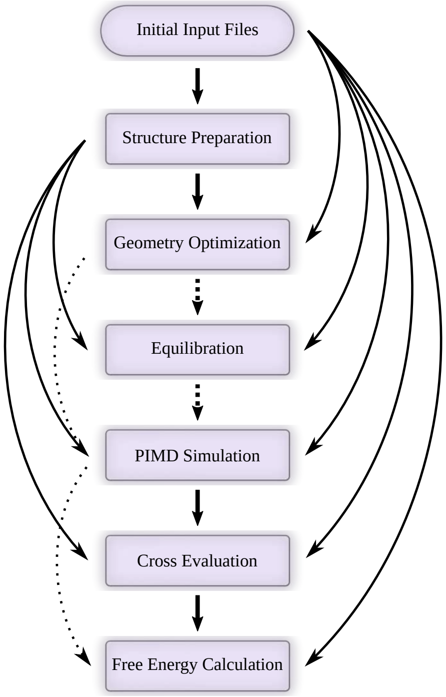


Figure 7.1: General HyperQ workflow. Arrows indicate the flow of information, dotted arrows imply setup-dependent information flows. The equilibration step is optional, if omitted the PIMD step will directly depend on the geometry optimization. Every component depends on the input files, and all but one (free energy computation) depend on the structure files.

- **i-PI Support.** HyperQ used i-PI as the code which drives the PIMD simulations. This makes all the advanced PIMD techniques which i-PI provides also available in HyperQ (many of them have been mentioned in subsection 6.2.1 on page 106).
- **CP2K Support.** HyperQ supports CP2K and uses it as the major code for providing the force field. Since CP2K is extremely versatile and supports all MM, QM and QM/MM simulations, all of these potentials can be readily used with HyperQ
- **i-QI Support.** HyperQ can prepare and run the simulations such that i-QI is an integral part of the simulations (i.e. a client of i-PI), which is mainly relevant when the QUASAR potential should be used during QM/MM simulations. It can work hand in hand with CP2K during such simulations.
- **Batchsystem Support.** In order to meet the scalability goal, and because of the potentially very high computational costs of the QSTAR method, a distributed resource management systems (DRMSs) module has been developed for HyperQ to allow to run on computer clusters managed by batchsystems. Currently supported are SLURM, PBS, LFS and SGE, and extensions to other batchsystems are straightforward.
- **Quality and Robustness.** HyperQ has been extensively tested and many weaknesses and causes of problems identified and removed, and therefore should be able to run stably on a wide range of systems. Any of the simulations can also be continued or extended, either after crashes or intentional breaks, and HyperQ can be set up such that it continues close to the point where the last run has

stopped without losing the already available simulation data.

- **Pure MD Simulations.** HyperQ can also be used to run vanilla PIMD or even standard MD simulations, without any alchemical transformations. Such simulations can be useful for other purposes than computing free energies, and HyperQ can be convenient for such simulations as it still can prepare the required input files, run the simulations efficiently and conveniently, making all the features of i-PI, CP2K and i-QI available.
- **Disk Usage Reduction.** PIMD-based FES can be extremely demanding regarding the available storage, in particular because one snapshot does not only contain one set of coordinates but as many as there are beads (path integral replicas), i.e. $P \in \mathbb{N}$, which can be further substantially increased by the energy cross evaluations which are required by the alchemical FEMs. HyperQ can be set up such that only the essential files are stored permanently, that they are compressed on the fly, and simulation output files which are not needed anymore deleted during the runtime.
- **Dynamic Workflow Control.** When using the batchsystem module, jobs can be monitored, terminated and continued at any time, and the job configuration adjusted easily from job to job. Batchsystem jobs are able to automatically start successive jobs autonomously and intelligently, being able to respond to different types of errors which can occur.
- **Multiple Types of Free Energies.** HyperQ supports direct alchemical pathways and various types of transfer free energies (see Subsection 7.2.2 below).

7.2.2 Supported Thermodynamic Pathways

HyperQ can carry out almost any direct alchemical transformation between different molecules (as described in Chapter 4). In the context of binding free energies of small molecules to receptors (as common in CADD), which is currently the primary target application of HyperQ, the molecules which are alchemically transformed into each other would be the ligands. Therefore in HyperQ the transforming species are always called ligands as it provides an uniform naming scheme throughout all types of simulations, even if these molecules should in fact not be ligands (e.g when computing solvation free energies).

In HyperQ molecules (ligands) can currently be transformed within three different molecular environments, viz.

(1) in vacuo,

(2) in solution (S)

(3) in solution with a receptor (R).

The resulting molecular systems which are supported by HyperQ are

(1) ligand-only (L),

(2) ligand and solvent (LS),

(3) receptor, ligand and solvent (RLS),

Moreover, in RLS systems the solvent can also consist of zero atoms, effectively allowing the simulation of RL-systems.

The three different environments allow the simulation of a variety of different types of free energies which are of interest in applications, all of them can be considered *transfer free energies*:

(1) *Solvation free energies*, i.e. the free energy difference between molecules in vacuum and a solvent,

(2) *Distribution free energies* (corresponding to distribution/partition coefficients), i.e. free energy differences of molecules in different types of solvent,

(3) *Binding free energies*, i.e. ligands binding to a receptor structure.

Transfer free energies can be computed either in an absolute way involving only a single molecule of interest, or in a relative mode in which the transfer free energy difference is computed between two different molecules. The latter are double free energy differences, i.e. $\Delta\Delta A$. The standard thermodynamic cycle for relative transfer free energies is illustrated in Figure 7.2. The relative transfer free difference is now defined to

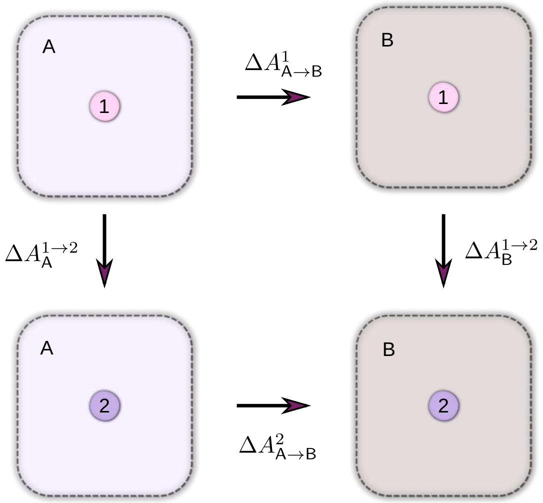

A
1
$\Delta A^1_{\mathrm{A}\to\mathrm{B}}$
B
1
$\Delta A^{1\to 2}_{\mathrm{A}}$
$\Delta A^{1\to 2}_{\mathrm{B}}$
A
2
B
2
$\Delta A^2_{\mathrm{A}\to\mathrm{B}}$

Figure 7.2: Illustration of the direct relative transfer free energy scheme. Two molecules (ligands (L), denoted by 1 and 2) are transferred from one environment A to another environment B.

be the difference between the (absolute) transfer free energies of the two molecules, i.e.

$$\Delta\Delta A^{1\to 2} := \Delta A^2_{\mathrm{A}\to\mathrm{B}} - \Delta A^1_{\mathrm{A}\to\mathrm{B}}. \tag{7.1}$$

Since in the total free energies of a complete thermodynamic cycle has to be zero it follows that we have

$$\begin{aligned} & 0 = \Delta A^2_{\mathrm{A}\to\mathrm{B}} - \Delta A^1_{\mathrm{A}\to\mathrm{B}} + \Delta A^{1\to 2}_{\mathrm{A}} - \Delta A^{1\to 2}_{\mathrm{B}} && (7.2)\\ \Leftrightarrow \quad & \Delta A^2_{\mathrm{A}\to\mathrm{B}} - \Delta A^1_{\mathrm{A}\to\mathrm{B}} = \Delta A^{1\to 2}_{\mathrm{B}} - \Delta A^{1\to 2}_{\mathrm{A}} && (7.3)\\ \Leftrightarrow \quad & \Delta\Delta A^{1\to 2} = \Delta\Delta A_{\mathrm{A}\to\mathrm{B}}, && (7.4) \end{aligned}$$

where

$$\Delta\Delta A_{\mathrm{A}\to\mathrm{B}} := \Delta A^{1\to 2}_{\mathrm{B}} - \Delta A^{1\to 2}_{\mathrm{A}}. \tag{7.5}$$

This means that we can compute $\Delta\Delta A^{1\to2}$ by computing $\Delta A_{\mathrm{B}}^{1\to2}$ and $\Delta A_{\mathrm{A}}^{1\to2}$, which is normally dramatically more efficient than computing $\Delta A_{\mathrm{A}\to\mathrm{B}}^{1}$ and $\Delta A_{\mathrm{A}\to\mathrm{B}}^{2}$.

When the special type of transfer free energy under consideration is the solvation free energy, then the environment A is vacuum and B consists of the condensed phase of interest, usually water. In the case of binding free energies the above more general scheme for transfer free energies takes the form illustrated in Figure 7.3. The ligand and the receptor are assumed to be so far separated in the unbound state that their interaction can be neglected, and thus they can be treated as individual systems. By the more general equation (7.4) it follows that

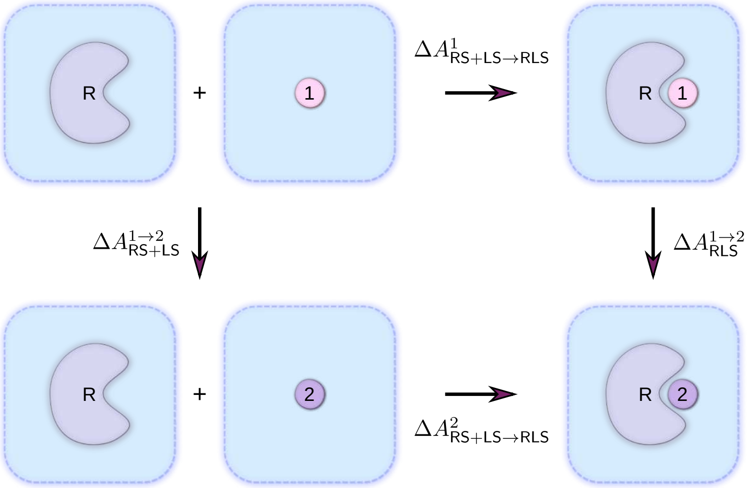

R
+
1
$\Delta A^{1}_{\mathsf{RS+LS\to RLS}}$
R 1
$\Delta A^{1\to2}_{\mathsf{RS+LS}}$
$\Delta A^{1\to2}_{\mathsf{RLS}}$
R
+
2
$\Delta A^{2}_{\mathsf{RS+LS\to RLS}}$
R 2

Figure 7.3: Illustration of the direct relative binding free energy scheme. Two ligands (L), denoted by 1 and 2, are dissolved in water in the initial unbound state, and in the final state bound to their target.

$$\Delta\Delta A^{1\to2} = \Delta\Delta A_{\mathrm{A}\to\mathrm{B}} \tag{7.6}$$
$$= \Delta A_{\mathrm{RLS}}^{1\to2} - \Delta A_{\mathrm{RS+LS}}^{1\to2} \tag{7.7}$$
$$= \Delta A_{\mathrm{RLS}}^{1\to2} - \Delta A_{\mathrm{LS}}^{1\to2} \tag{7.8}$$

since

$$\Delta A_{\mathrm{RS+LS}}^{1\to2} = \Delta A_{\mathrm{RS}}^{1\to2} + \Delta A_{\mathrm{LS}}^{1\to2} \tag{7.9}$$

and $\Delta A_{\mathrm{RS}}^{1\to2} = 0$.

While HyperQ can in theory compute both absolute and relative transfer free energies, it is currently tailored and efficient in particular to the computation for computing relative free energies. For absolute free energies the code should be extended with a few relatively simple additional features. To compute absolute binding free energies the double decoupling scheme is one of the most commons ones. In this scheme the ligand is decoupled both from the LS as well as the RLS systems. If no restraint is present the method is called *double annihilation method*, and one of the first applications of the scheme can be found in [135] by Jorgensen in the year 1998. The problem when no constraints are present is that the decoupled ligand can float around freely in the solvent, making the phase space overlap with the bound state of the ligand extremely small. Therefore restraints are normally used in addition to keep the ligand which was decoupled from the receptor at the binding site, while the ligand which is decoupled from the solvent only remains unrestrained. The free energy difference between the unrestrained and the restrained decoupled ligand can then be computed analytically. This approach in which restraints are present is called the *double decoupling method*, and first applications of it can be found as early as 1986 by Hermans [118]. These approaches have also been analyzed and extended in [241, 98, 34].

7.3 File and Directory Structure

The directory structure of HyperQ is designed in a way which closely resembles the workflow itself. Therefore it will be described in more detail in this section. Every HyperQ workflow has a root folder, and this folder is primarily structured by the pipeline components listed above in section 7.2 on page 118. For each of the pipeline components (2)-(6) a separate folders exists. In addition there are folders for the input-files, log-files, runtime-dependent files used by the HyperQ, and a folder for the batchsystem module.

HyperQ uses a specific nomenclature to be able to handle and clearly distinguish the various types of (molecular) systems occurring in the workflow:

(1) **Systems.** The term *system* is used to refer to the initial molecular systems between which the free energies should be computed. These are the endpoints (or end states) in alchemical pathways.

(2) **Subsystems.** Free energy simulations involve simulations with different environments, often only a part of the entire system (for instance simulations of a solvated ligand without receptor). These types of systems are called *subsystems*. HyperQ currently supports three types of systems, molecules/ligands without solvent (L), with solvent (LS), and with receptor structure (RLS).

(3) **Molecular System Pairs (MSPs).** Since only free energy differences between two systems are computed, the original systems need to be paired in specific ways *a priori* the simulations, one of them being the initial system and the other one being the final/target system. One such pair is called a molecular system pair (MSP).

(4) **Thermodynamic States (TDSs).** When the initial system of an MSP is transformed into the target system, this might not happen directly but via any number of intermediate thermodynamic states. These states can be of alchemical nature but don't have to be, therefore these states are referred to simply as thermodynamic state (TDS).

The three types of currently supported subsystems (L, LS, RLS) allow to compute direct alchemical transformations (in vacuum and in solution), solvation free energies, and free energies of binding.

The four types of systems introduced above give rise to multiple different hierarchies of states. In HyperQ the following organizational hierarchy is used:

$$\text{Systems} \rightarrow \text{MSPs} \rightarrow \text{Subsystems} \rightarrow \text{TDSs}$$

The directory structure of the root and `input-files` folder is shown in Figure 7.4.

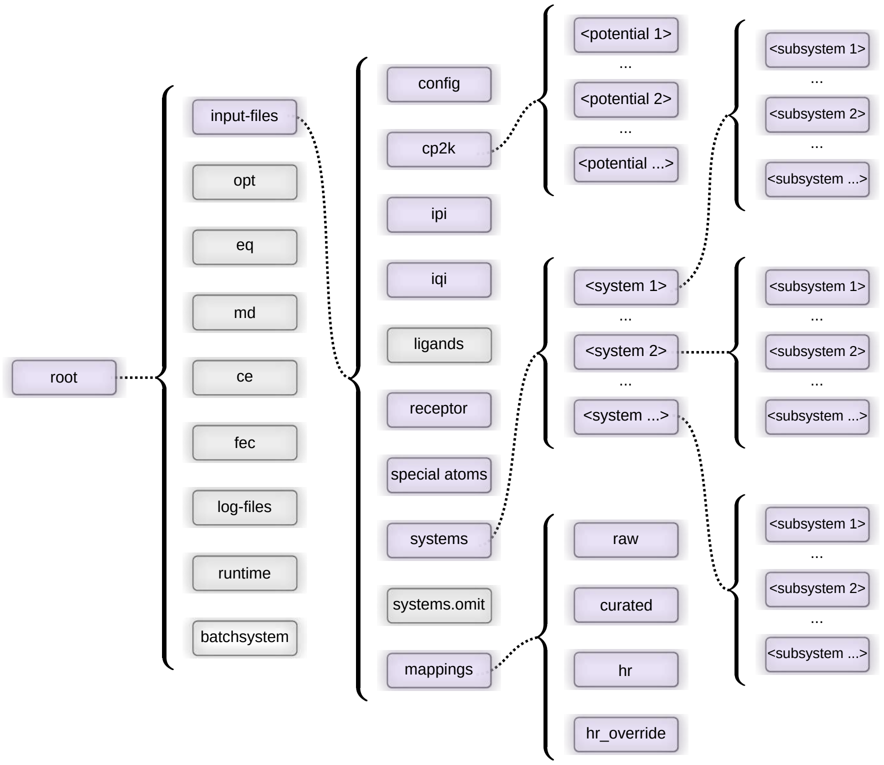


Figure 7.4: Overview the HyperQ root and `input-files` folders. For the directories colored in pale violet the subfolders are shown, while the folders in grey are not expanded (i.e. they have at least one subdirectory). The folders in the root directory contain the following abbreviations: `opt`: geometry optimization, `eq`: equilibration, `md`: production (PI)MD simulations, `ce`: energy cross evaluations, and `fec`: free energy computations.

Input Files

Component (1) of the HyperQ pipeline, i.e. the structure preparation shares a folder with the initial input files, which have both subfolders in the general `input-files` folder. Within this directory the subfolders `config`, `ipi`, `iqi`, `cp2k`, `ligands`, `receptor`, and `special-atoms` comprise the initial input files. For the folders `config`, `ipi`, `iqi`, `cp2k` and `special-atoms` default input files are provided by i-QI, but they can and have to be adjusted during the initial setup to fit ones needs. The ligand and receptor files always need to be provided. A default configuration file can be found in appendix B on page 243. The remaining folders, `systems`, `systems.omit` and `mappings` can be created by HyperQ automatically during the structure preparation step (component (1) of the HyperQ pipeline).

7.4 Pipeline Components

The six components of the HyperQ pipeline have been introduced in 7.2, which comprise structure preparation (1), geometry optimizations (2), equilibrations (3), production (PIMD) simulations (4), energy cross evaluations (5), free energy computations (6).

7.4.1 Structure Preparation

During the structure preparation step the required system-dependent input-files for the simulations are generated. As outlined in subsection 7.2.2 on page 120, each molecule in HyperQ belongs to one of three classes, L, S or R, which are handled differently.

Ligands

HyperQ can start from different ligand formats, either SMILES, PDB, MOL2, or the SDF format. Each ligand gives rise to a system, and HyperQ uses the name of the ligand as the system name.

For each ligand MM parameters are required for two potential purposes: For carrying out MM simulations and for modeling the dummy atoms which are used during the alchemical transformations. HyperQ can automatically parametrize the ligands using the program MATCH [344] based on the CGenFF force field [304, 348]. Alternatively, MM force field (topology and parameter) files can also be provided by the user which will be used automatically if present.

Receptors

HyperQ allows that different ligands use the same receptor, which is useful when one wants to test a number of different compounds for their binding affinity to the same receptor, e.g. during drug discovery. However, it is also possible to provide an individual receptor file for each ligand. Similarly as with ligands, also for the receptor structures MM force field files are required. These can be provided manually, or generated by HyperQ. Currently two types of receptors are supported during the automatic preparation step, proteins (P) and smaller host molecules (H) such as Cucubit[8]uril or octacids. Host molecules are low-molecular weight compounds which form cavities allowing small molecules/ligands to bind to them. An example of such a host is shown further below in section 7.6. For proteins HyperQ uses CHARMM36 force field by default, and for hosts the CGenFF force field. Host molecules are popular for instance testing FES methods for computing binding free energies as they can emulate the binding pockets of proteins, but are computationally less demanding due to their smaller size and to the lack of conformational changes on large time scales. For these reasons also the SAMPL host-guest challenges use these types of hosts as their test systems [206, 345].

Systems and Special Atoms

During the preparation of the systems based on the ligand and receptor files, dependent on the subsystem (L, LS, RLS) these molecules are assembled and solvated into one system. All the systems which are prepared are synchronized regarding the solvent phase so that each one has the same number of water molecules as the systems need to have the same number of degrees of freedom during the alchemical transformations.

In the `special-atoms` folder the atoms which unique roles can be specified. Atoms can have special roles in QM/MM approaches, and further types arise when the QUASAR method is used which was introduced in section 3.4 on page 50. The relevant atoms types have been summarized already in subsection 6.3.2 on page 108 in which the PDBX file format was introduced which is used by i-QI, namely

- restricted/unrestricted (R/U),
- quantum/classical (Q/M).

These types are also used by HyperQ, and only play a role during QM/MM simulations and/or simulations in which i-QI is used with the QUASAR scheme. However, one only needs to specify the unrestricted atoms and the quantum atoms as the other two types are their (set-theoretic) complements. HyperQ automatically creates the QM/MM input files for CP2K based on the Q-atoms, and PDBX files for i-QI based on the Q-atoms and U-atoms. The special types of atoms can be most conveniently specified in HyperQ as it makes available the powerful atom selection language of VMD.

Alchemical Mappings

Currently HyperQ supports primarily relative transfer free energies, meaning that it always computes the transfer free energies differences between pairs of systems (i.e. MSPs). [2]

For free energy computations HyperQ requires two types of mappings/associations:

(1) A list of MSPs, i.e. a list which states which systems form pairs as well which of them is the initial and the final state.

(2) An atom mapping for each MSP which specifies which atoms are transformed into which (only for the ligands since this is the species which is transformed in HyperQ). The mapping also implicitly implies how many dummy atoms are required and at which location.

One of the advantages of relative transfer free energies is that if one wants to compute the free energy difference between each pair of M molecules, one does need to compute the free energy difference directly between all $M(M-1)/2$ possible pairs of systems corresponding to a fully connected graph, but for a edges of a graph which connects all the systems by at least one path. This means that only $M-1$ MSPs are required. But it can be useful to have some redundancies in the graph for verification purposes and for the case that one or more of the FES should fail or diverge for some reason. An example of a slightly redundant graph is shown in Figure 7.5 on the facing page.

Any number, direction, and combination of individual systems to form MSPs can be specified manually within HyperQ. The atom mappings can be provided manually as well. However, HyperQ can carry out both the identification of system pairs which

[2] To compute absolute free energies of binding one can in theory use for the second system a ligand which has no atoms at all, meaning that all atoms of the first system are transformed into dummy atoms.

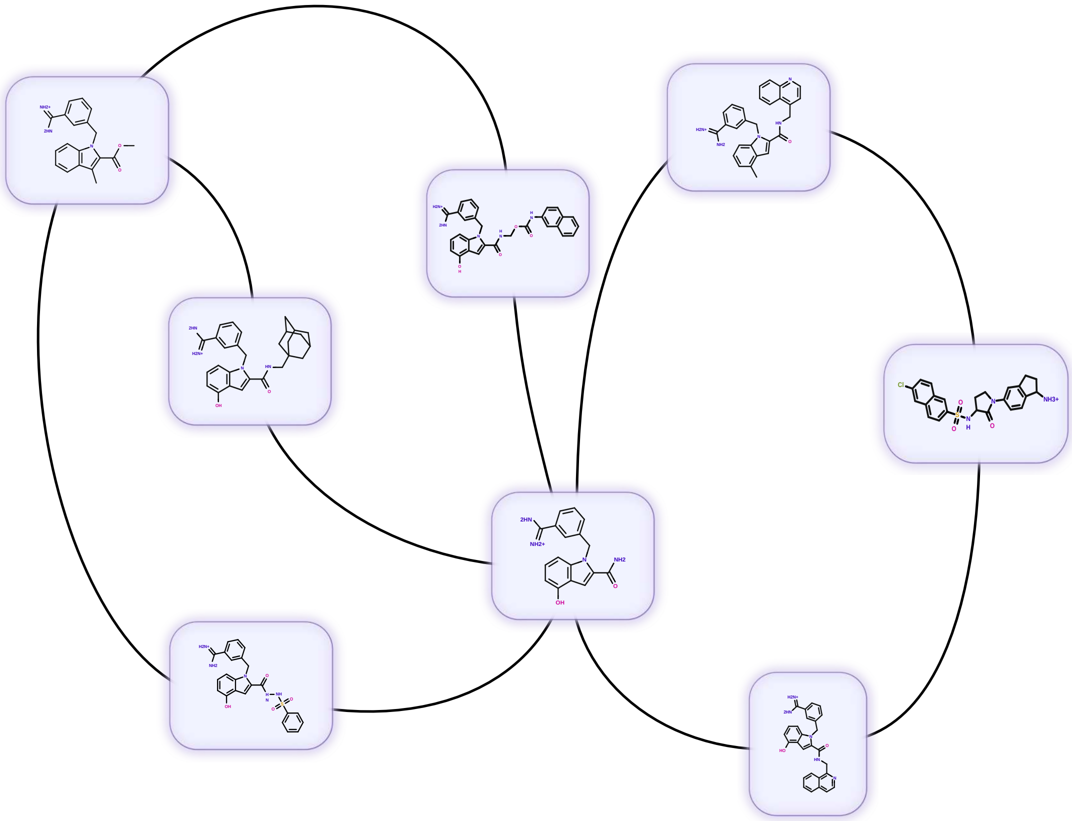


Figure 7.5: Example of a relatively small graph of eight molecules showing which systems have been paired to form MSPs.

suit to each other as well as the associated atom mapping. It does so by delegating the task to a tool called Lead Optimization Mapper (LOMAP) [181], which can provide both system and atom mappings. LOMAP itself depends on RDKit [286] to compute the maximum common substructure between pairs of molecules. The system pairings play a very important role in alchemical FES, and can affect the quality of the result dramatically. Therefore using a tool which is specialized on this task can be of significant benefit. Moreover the automatic mapping can be extremely useful when one has a large number of ligands for the same receptor and if one needs to find those pairs which are most easily transformed into each other. The number of molecular pairs rises exponentially with the number of compounds and becomes soon unfeasible for manual treatment.

For each MSP an atom mapping is required which specifies which atoms should transformed into which ones, which implies indirectly also which atoms have to be transformed into dummy atoms. An illustration of a single-topology atom mapping of an MSP via their common substructure shown in Figure 7.6.

HyperQ curates the output files of LOMAP and provides them in a human readable format which can be most easily modified by the user. If one wants to use a custom mapping file for one or more of the ligands, one can place them in the `input-files/mappings/hr_override` folder and HyperQ will automatically use these files instead. An example mapping file which can be adjusted manually is shown in Listing 7.1 on the next page.

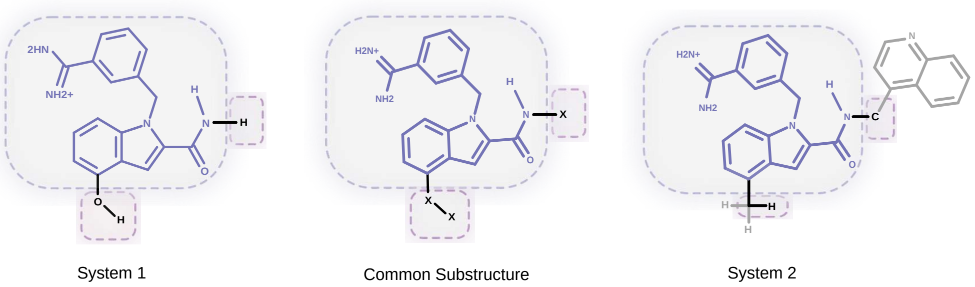


Figure 7.6: Simple example of a mapping of two molecules via a common substructure. Bonds in pale violet connect atoms which are identical in both substructures, bonds and atoms in black are present in the common substructure but are of a different type, and bonds in grey are present in only one substructure (which will be transformed into dummy atoms).

Listing 7.1: Example atom mapping file in HyperQ. The first two columns are relevant and processed, the remaining four columns are present as an aid to allow the atoms to be more easily identified.

```
# All indices are based on the order of the atoms in the pdb/psf files)
#
# Column 1: System 1 reduced indices (without receptor if present)
# Column 2: System 2 reduced indices (without receptor if present)
# Column 3: System 1 total indices
# Column 4: System 2 total indices
# Column 5: System 1 atom names
# Column 6: System 2 atom names

    1      9    404    412 C1Q    C7Q
    2      8    405    411 C2Q    C6Q
    3      7    406    410 C3Q    C5Q
    4      6    407    409 C4Q    C4Q
    5     11    408    414 C5Q    C9Q
    6     10    409    413 C6Q    C8Q
    7     15    410    418 C7Q    C13Q
    8     14    411    417 C8Q    C12Q
    9     13    412    416 C9Q    C11Q
   10     12    413    415 C10Q   C10Q
   11     17    414    420 C11Q   C15Q
   12     16    415    419 C12Q   C14Q
   13      5    416    418 C13Q   C3Q
   23     40    426    434 H1Q    H7Q
   24     39    427    442 H2Q    H6Q
   25     42    428    445 H3Q    H9Q
   26     41    429    444 H4Q    H8Q
   27     45    430    448 H5Q    H12Q
   28     44    431    447 H6Q    H11Q
   29     43    432    446 H7Q    H10Q
   30     47    433    450 H8Q    H14Q
   31     46    434    449 H9Q    H13Q
```

7.4.2 Simulations

In the context of HyperQ the term *simulation* refers to all components of the pipeline in which force/energy evaluations are carried out, implying that the following four parts fall into this class:

(i) **Geometry optimization.** To remove sterical clashes which could lead to crashes of the simulations.

(ii) **Equilibration.** To bring the system into equilibrium, allowing for instance the volume to adjust in NPT simulations when the production runs are NVT.

(iii) **Production (PIMD) simulation.** These simulations provide the sampling and the native energies which are required by the alchemical free energy methods.

(iv) **Energy cross evaluation.** Most alchemical free energy methods required the evaluation of conformations at other potentials than the sampling potential.

While all these types of simulations are different they share several commonalities and similarities in HyperQ, for instance regarding the hierarchical folder structure, the configuration/setup, and parts of the automatic preparation procedures. Each of the simulation steps consist of the following subcomponents:

(1) **Preparation.** Preparation of the all the files and folders which are required to run the simulations.

(2) **Execution.** Running of the simulations for single force evaluations in the case of the cross evaluations.

(3) **Postprocessing.** Only relevant for optimization and equilibration.

The PIMD simulation as well as the cross evaluation step don't possess postprocessing steps because these are integrated in the preparation steps of the subsequent components of the workflow, e.g. the preparation subcomponent of the cross evaluation step postprocesses the data the from the PIMD simulation step. This design is by purpose and makes the handling of those steps more simple.

Folder Structures

The hierarchical folder structure is basically the same for all types of simulations up to a certain level, and it exhibits the relation of the various simulations which are required. Moreover, it reveals to some extend the complexity and the large number of simulations which are required when computing RBFEs between multiple systems/molecules. The expanded folder structure for the `md` folder is illustrated in Figure 7.7. Geometry optimization and equilibration have a relative folder depth (with respect to their individual root folders) of four, while the PIMD simulations require an additional level due to the number of beads. Energy cross evaluations require even one more level because for each snapshot/cross evaluation a separate one-time point simulation is required.

Configuration Files

There is a considerable number of different types of configuration files required to run simulations with HyperQ.

(1) At first the primary HyperQ configuration files themselves in the folder `input-files/config`, which contains at least one file which is called *general.txt*. This file will be used by HyperQ for MSP independent settings and for all MSPs, except for those for which a file with the name of the MSP itself is present which allows to override the general configuration for this MSP. A default primary configuration file of HyperQ can be found in the appendix in Listing B.1 on page 243.

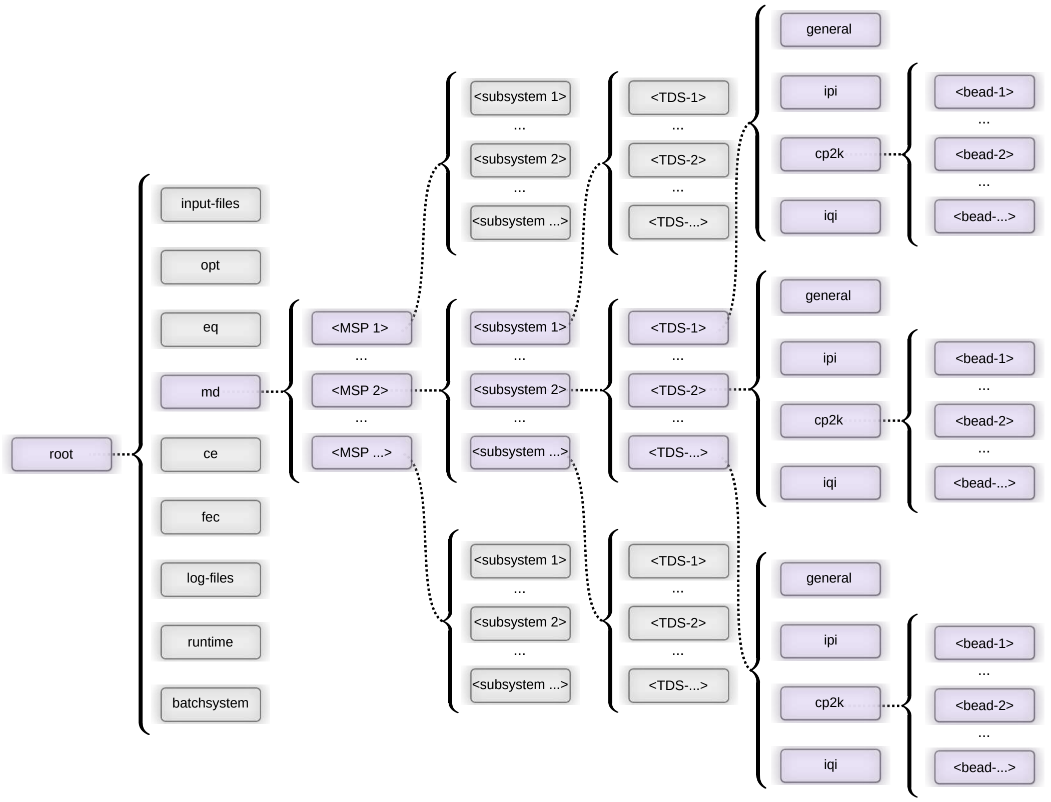


Figure 7.7: Overview of the HyperQ MD folder, which reveals the hierarchical structure of the states which need to be simulated. For directories colored in pale violet the subfolders are shown, while the folders in grey are not expanded (i.e. they have at least one subdirectory). The folders in the root directory contain the following abbreviations: opt: geometry optimization, eq: equilibration, md: production (PIMD) simulations, ce: energy cross evaluations, and fec: free energy computations.

(2) The HyperQ configuration files for the optional batchsystem module (more details can be found in section 7.5.2 on page 138). A default batchsystem configuration file of HyperQ can be found in the appendix in Listing B.2 on page 253.

(3) Input files for i-PI, the server which drives the PIMD simulations and cross-evaluations.

(4) Input files for CP2K, which is employed during all types of simulations.

(5) Input files for i-QI (optional) in case that QM/MM simulations are carried out with the QUASAR scheme.

HyperQ provides default configuration files for all the above mentioned types. The number of required configuration files is multiplied by

(1) the number of different types of simulations (optimization, equilibration, ...), since each of them requires specific settings,

(2) the subsystems, since different types of subsystem (L, LS, RLS) can require different settings,

(3) the type of potential, such as MM, QM/MM, or any of the available QM potentials.

(4) the two systems which are contained in each MSP (and thus require two different force evaluations, one for the initial state and one for the endstate)

If all the resulting combinations of possible input-files would be provided naively by HyperQ, it would result in at least two problems:

(1) **Modification Complexity.** If one would like to modify a single setting throughout the simulations, one would need to edit multiple files, and possibly also multiple values in the same configuration file since it might contain settings which need to be set multiple times (e.g. in QM/MM simulations this can be the case in CP2K). Or if one wants to change a global parameter in all CP2K files this would mean that a large number of files would have to be edited. If one needs to find the proper parameters the required editing process can become extremely cumbersome.

(2) **Additional Error Source.** If one changes a value of certain parameters it can happen that they have to be in sync with other parameters, either in the same configuration file or in configuration files of one of the other programs. This can easily lead to errors by the user.

To address these problems, several techniques have been implemented in HyperQ which provide a clean, intelligent and efficient way to deal with all the configuration files described below.

Bundling via HyperQ. There are a few settings which need to be changed and adapted frequently, such as the simulation length (number of steps), the stride at which conformations or properties like energies are written to output files, or the simulation temperature. These settings have been made native HyperQ settings, and HyperQ in turn adjusts all the required input files automatically when preparing the simulations (see also Figure 7.8). In addition a new set of *external* HyperQ *parameters* has been introduced for the external simulation codes which can be used within the configuration files of these programs. This allows to easily set up new template-configuration files for these programs in a manner which allows to customize the way in which HyperQ processes these file during the preparation of the simulations.

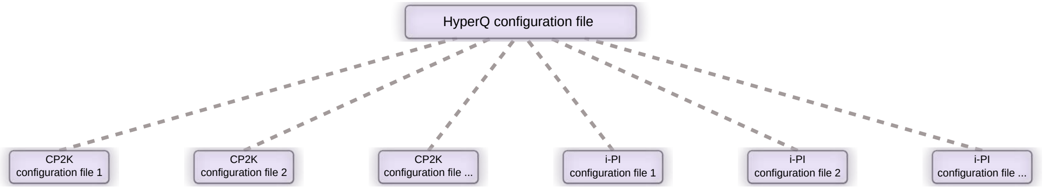

HyperQ configuration file
CP2K configuration file 1
CP2K configuration file 2
CP2K configuration file ...
i-PI configuration file 1
i-PI configuration file 2
i-PI configuration file ...

Figure 7.8: Key settings and shared parameters which are required to be in sync are delegated to and bundled within the HyperQ main configuration file. These settings involve parameters like the total simulation length/number of steps, the temperature or scaling factors which can be used for the EWALD parameters via the external HyperQ variables in CP2K.

Modular Template File Structure. The lion's share of the template configuration files are for CP2K, since CP2K provides either the full or the major part of the force field in all types of simulations, and because there are many different potentials from MM over QM/MM to full QM methods. To reduce the number of configuration files arising from the possible combinations, the CP2K template files are splitted into fragmental files (subconfiguration files) which are organized in a modular manner. Each elementary potential type (e.g. PM6 or DFTB3) is stored in its own subconfiguration file, which is true for the dummy atom force evaluations as well. Moreover, for each type of simulation (e.g. equilibration) primary CP2K input files exist, and they are set up such that they source the subconfiguration files. Because the subconfiguration files depend on variables which differ from simulation to simulation, the external HyperQ parameter infrastructure (which have been introduced in the previous paragraph about the variable bundling) is used to handle this issue as well.

In addition, the CP2K template configuration files are organized in folders of the types *general* and *specific*. In the HyperQ configuration file it is possible to specify two input folders for CP2K which should be used, and files in the specific CP2K input folder will always override files in the general folder if they should be present there as well. This mechanism allows a convenient and flexible management of all the required files and settings.

Simulation Preparations

During the structure preparation step all the structure files have been prepared which belong to the individual systems. For the actual simulations of a certain MSP which have to be carried out a joint system has to be created. This joint system takes into account the atoms of the individual systems which are mapped onto each other as well as the atoms which are not. Since during the initial structure preparation step it is not known yet which pairs of systems will be simulated, the preparation of the joint system is only carried out directly before the simulation together with the preparation of the required folders and other simulation files.

The joint system consists of all the solvent (S) and receptor (R) atoms if present, which are always the same in both systems. The ligand (L) in the joint system consists of the merged ligand of both individual systems. If an atom of the ligand 1 is not mapped onto an atom of ligand 2 and vice versa (i.e. not belonging to the common substructure), dummy atoms are added automatically. In addition dummy atoms might arise when the PEARL scheme is applied, as then the ligands of both systems are simultaneously transformed in dummy atoms in counter-rotating directions. These types of atoms give rise to different versions of the system (X stands for 1 or 2, and Y for the other of the numbers, i.e. 2 or 1, respectively):

(1) **SX.** In the case of X=1 the initial system 1, but not necessarily the original system but rather the initial state of current alchemical transformation. In the case that the PEARL scheme was applied some atoms might be present only in the form of internal dummy atoms (see XS.SXD).

(2) **SX.SXD.** The dummy atoms (D) of system X within the original system X which might have arisen due to the PEARL scheme. These are internal dummy atoms of the system, i.e. the dummy atoms of system X for system X (therefore they are denoted by SX.SXD).

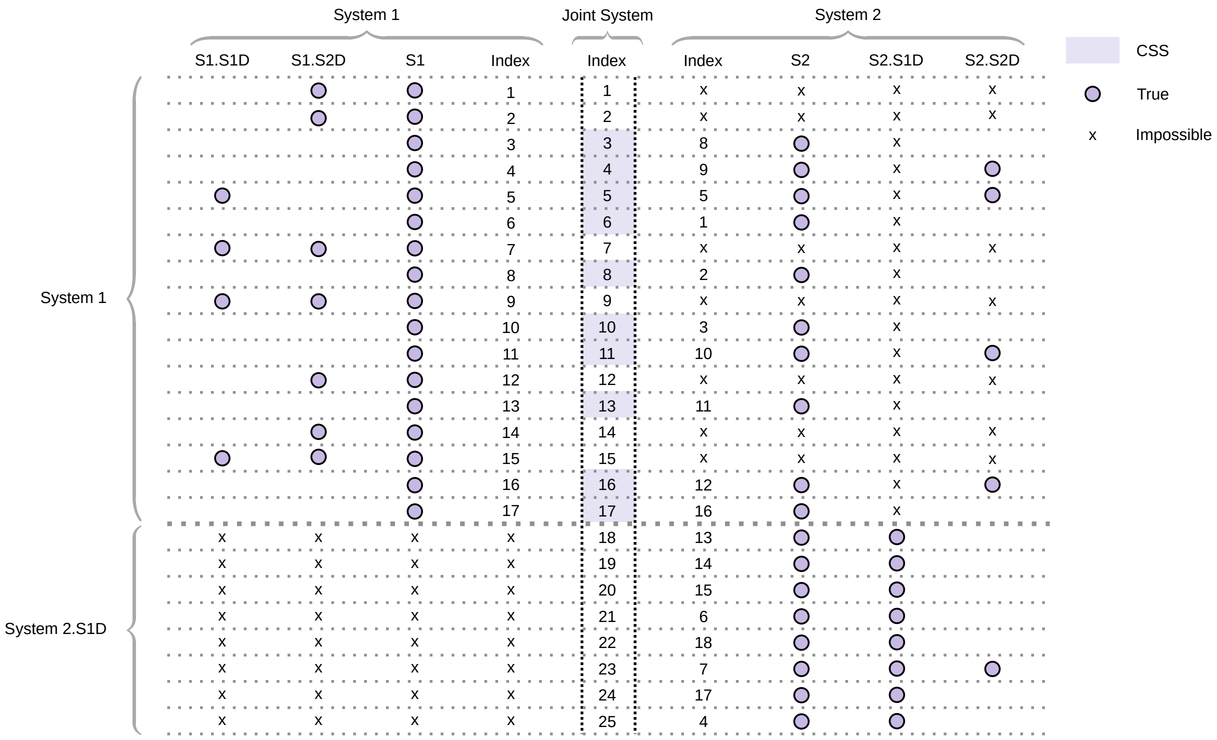
(3) **SX.SYD.** The dummy atoms of system X which are present in the joint system, and are evaluated with system Y to hold these atoms (or some of their beads)

in place.

(4) **SX.red.** The reduced system X, meaning SX without its internal dummy atoms SXD.

(5) **SX.aug.** System X together with the dummy atoms of the other system, SX.SYD, is called the augmented system X.

An overview of the atom types present in the joint system is illustrated in Figure 7.9. In HyperQ the joint system is based upon system 1 regarding the order of atoms, adding at first the atoms of system 1 (S1) and then the remaining atoms of system 2 (S2), which become automatically dummy atoms in the augmented system 1, denoted by S2.S1D. The indices of the atoms of system 2 however become mixed up as they are mapped on the joint system.



	System 1				Joint System	System 2			
	S1.S1D	S1.S2D	S1	Index	Index	Index	S2	S2.S1D	S2.S2D
System 1		○	○	1	1	x	x	x	x
		○	○	2	2	x	x	x	x
			○	3	▨ 3	8	○	x	
			○	4	▨ 4	9	○	x	○
	○		○	5	▨ 5	5	○	x	○
			○	6	▨ 6	1	○	x	
	○	○	○	7	7	x	x	x	x
			○	8	▨ 8	2	○	x	
	○	○	○	9	9	x	x	x	x
			○	10	▨ 10	3	○	x	
			○	11	▨ 11	10	○	x	○
		○	○	12	12	x	x	x	x
			○	13	▨ 13	11	○	x	
		○	○	14	14	x	x	x	x
	○	○	○	15	15	x	x	x	x
			○	16	▨ 16	12	○	x	○
			○	17	▨ 17	16	○	x	
System 2.S1D	x	x	x	x	18	13	○	○	
	x	x	x	x	19	14	○	○	
	x	x	x	x	20	15	○	○	
	x	x	x	x	21	6	○	○	
	x	x	x	x	22	18	○	○	
	x	x	x	x	23	7	○	○	○
	x	x	x	x	24	17	○	○	
	x	x	x	x	25	4	○	○	

▨ CSS
○ True
x Impossible

Figure 7.9: Example of the ligand within a joint system of an MSP, showing the classes to which each atom belongs to. Atoms belonging to the CSS, i.e. atoms which are mapped onto each other, have a pale violet background color in the central column. Atoms which are not part of the CSS have to be present either in system 1 or 2, but not both. These atoms automatically become dummy atoms within the other (augmented) system.

Since the dummy atoms are always treated on the MM level, the potential of the augmented systems SX.aug can be represented as a sum.

$$V_{\mathrm{SX.aug}} = V_{\mathrm{SX.red}} + V_{\mathrm{SX.SXD}} + V_{\mathrm{SX.SYD}}. \tag{7.10}$$

In CP2K the potential evaluations are carried out by evaluating these three potentials separately due to implementation aspects and flexibility, and moreover it has the advantage that it is also applicable when QM/MM potentials are used. Since during alchemical transformations both systems are present at the same time (via a mixed potential), this means that up to six for potential evaluations are carried out with CP2K per time step. The energy evaluation of the joint system (with the PEARL scheme) via its components in CP2K is illustrated in figure 7.10.

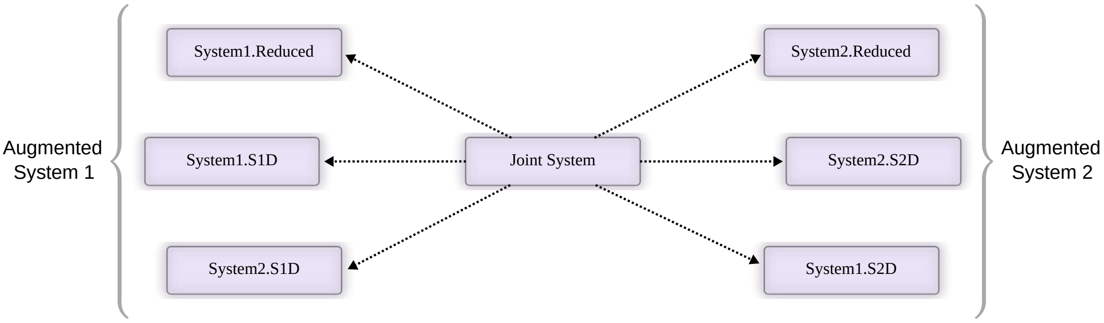


Figure 7.10: Energy evaluation of the joint system of an MSP with CP2K via its components when the PEARL scheme is active (which gives rise to internal dummy atoms).

When the reduced system is modeled by the QM/MM method, then $V_{\mathrm{SX.red}}$ might be further splitted up into subpotentials (e.g. MM, QM, and QM/MM-interaction parts). When the PEARL scheme is active the internal dummy atoms SX.SXD are not present and thus no potential evaluation for them is required. The potentials of the augmented system 1 and system 2 are only evaluated for the same particles in λ-schemes where Hamiltonians are directly mixed since in the κ-schemes of the QSTAR method each bead is modeled only by one of the two augmented systems (more details on these schemes have been described in Subsection 4.4).

7.4.3 Free Energy Computations

Regarding the free energy evaluations, at the moment only the BAR method is implemented in two versions, a primitive approach based on the minimum distance of the two curves, and alternatively a variant which uses a root finding algorithm. The latter is more efficient, but the former allows to create the characteristic plots innate to the BAR method. The computation of the free energies (as well as the cross evaluations) can be carried out on the fly while the PIMD simulations are still running. This allows one to observe the convergence of the results without affecting the simulations. Summary files for each free energy evaluation are created, plots related to the BAR method and the energy distribution generated, and the results of repeated free energy computations are automatically saved in a history folder. An example summary output file can be seen in listing 7.2.

Listing 7.2: Example free energy summary file in HyperQ where the BAR method has been used. The input parameters are required for the BAR method, and either obtained from the primary HyperQ configuration file or automatically computed/extracted from the simulations by HyperQ. The parameter C is described in the original publication [20], n_i are the numbers of snapshots considered, and F stands for the free energy.

```
 ***   Input data   ***
delta_F_min: -1000.0 kcal/mol
delta_F_max: 1000.0 kcal/mol
Temperature: 298.0 K
n_1: 394
```

```
n_2: 406
C_min: -1688.62628524
C_max: 1688.68628974
Absolute C-tolerance: 0.1
Reweighting: False

 ***   Results   ***
Min (Delta F_eqn1 - F_eqn2) found at C-value: 84.3328122913
Delta_F equation 1: 49.92290 kcal/mol
Delta_F equation 2: 49.92301 kcal/mol
```

Based on these individual free energy computations the resulting free energy differences for each thermodynamic window of the alchemical pathway is summed by HyperQ to provide the total free energy difference between the two end states.

Additional features, such as non-Boltzmann sampling (i.e. umbrella sampling or reweighting) is rudimentarily implemented but not fully functional yet. Alternative free energy methods such as MBAR are planned to be implemented as well.

7.5 Serialization and Parallelization

The HyperQ workflow consists of six primary components and several subcomponents as outlined already earlier in the chapter. All components and subcomponents of HyperQ are summarized below.

(1) Structure preparation (SP)

(2) Geometry optimization (OPT)

  (a) Preparation (PRO)
  (b) Execution/Running (ROP)
  (c) Postprocessing (PPO)

(3) Equilibration (EQ)

  (a) Preparation (PRE)
  (b) Execution/Running (REQ)
  (c) Postprocessing (PPE)

(4) Production MD/PIMD simulation (MD)

  (a) Preparation (PRM)
  (b) Execution/Running (RMD)

(5) Cross evaluation (CE)

  (a) Preparation (PRC)
  (b) Execution/Running (RCE)

(6) Free energy computation (FEC)

  (a) Preparation (PRF)
  (b) Execution/Running (RFE)
  (c) Postprocessing (PPF)

A graphical overview of all the primary components and subcomponents is shown in Figure 7.11 on the following page. The structure preparation (SP) step is

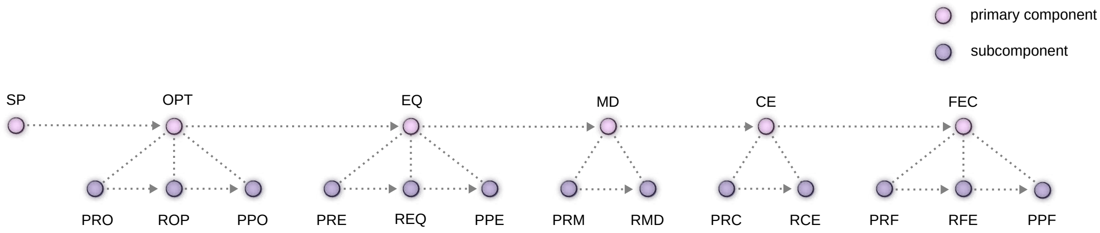


Figure 7.11: Overview of the various components and subcomponents in the HyperQ workflow.

carried out for all systems at once rather than for individual systems (which is not possible) due to two reasons.

(1) The number of solvent molecules is synchronized between all of the individual systems.

(2) The MSPs pairings are created by LOMAP, which takes into account all systems at once.

All the other steps of the workflow can be carried out for individual MSPs and specific subsystems.

The number of simulations which are required to compute the transfer free energies of a certain type (e.g. binding free energies) can become very large. As an example, if we have 30 different ligands forming 30 MSPs for which the binding free energy should be computed to a common receptor structure, and each alchemical pathway (defined by the QSTAR method) consists of 9 intermediate steps (i.e. 10 TDS), then a total of 2 400 simulations are required since

$$30 \text{ MSPs} \times 10 \text{ TDSs} \times 2 \text{ subsystems} \times 4 \text{ simulation-types} = 2\,400. \qquad (7.11)$$

If the simulations of each MSP are carried out in parallel (each type of simulation in serial), and if each atom in the PIMD simulations is represented by 12 beads while for each bead an individual CP2K instance is running, then 7 200 CP2K instances have to run in parallel. From these numbers it becomes clear that running all these simulations manually is unfeasible, but that automatic tools and mechanisms are essential. In order to be able to carry out all the steps of the workflow efficiently and conveniently, HyperQ provides two types of tools corresponding to one of two running modes, a *standalone mode* and a *batchsystem mode* for running on computer clusters. These two modes are described briefly in the next two subsections.

7.5.1 Standalone Mode

The basic standalone mode of HyperQ allows it to run on individual Linux machines.

Structure Preparation. For carrying out the initial structure preparation step of the individual systems, the commandline tool

```
hqf_sp_prepare_all.sh <subsystem> [lomap]
```

is provided. As mentioned above in HyperQ this happens for all the individual systems at the same time. The subsystem (L, LS, and RLS) can be freely chosen, as well as if LOMAP should be used for the preparation of the system pairings and the atom mappings or not (in which case they need to be prepared manually).

Managing Single MSPs. Besides the structure preparation step, all the remaining components of the HyperQ worfklow can be carried out for individual MSPs independently from each other. Even more, all TDSs belonging to each MSP and subsystem can be prepared, run and postprocessed independently.

HyperQ provides a single command which allows to run all of these workflow components and subcomponents conveniently, namely

```
hqf_gen_run_one_pipe.sh <MSP> <subsystem> <pipeline type> [<TDS range>].
```

The value of the argument `<pipeline type>` can be a concatenated expression of *elementary components* (subcomponents) and *macro components*, the available options of which are shown in Listing 7.3.

Listing 7.3: Available components for the option `<pipeline type>` to the command `hqf_gen_run_one_pipe.sh`.

```
Elementary components (subcomponents):
 _pro_: preparing the optimizations
 _rop_: running the optimizations
 _ppo_: postprocessing the optimizations
 _pre_: preparing the equilibrations
 _req_: running the equilibrations
 _ppe_: postprocessing the equilibrations
 _prm_: preparing MD simulation
 _rmd_: postprocessing MD simulation
 _prc_: preparing the crossevaluations
 _rce_: postprocessing the crossevaluations
 _prf_: preparing the free energy calculation
 _rfe_: postprocessing the free energy calculation
 _ppf_: postprocessing the free energy calculation

Macro components:
 _allopt_: equals _pro_rop_ppo_
 _alleq_ : equals _pre_req_ppe_
 _allmd_ : equals _prd_rmd_
 _allce_ : equals _prc_rce_
 _allfec_: equals _prf_rfe_ppf_
 _all_   : equals _allopt_alleq_allmd_allce_allfec_
```

As an example, if the optimization should be postprocessed and the equilibration be prepared, the pipeline specification would be `_ppo_pre_`. The available pipeline options are visualized in Figure 7.12 on the next page. The pipeline is designed in a way that the simulations can be continued after crashes or intended breaks, and that the next component can start while the previous one is still running. This is for instance useful if one wants wants to compute the free energy while the simulation is still running, in which case one can start the cross evaluations

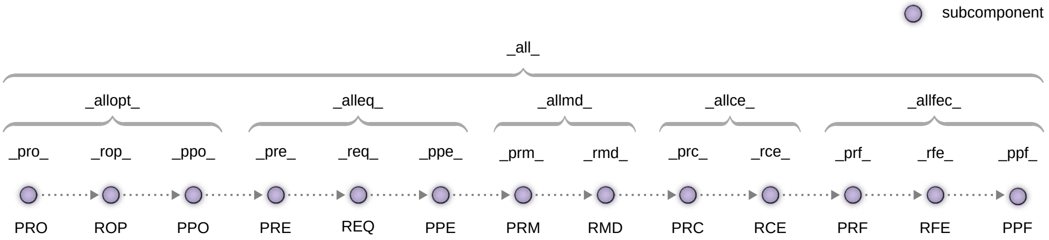


Figure 7.12: Overview of the scope of the available pipeline components for a single MSP in the commandline tool `hqf_gen_run_one_pipe.sh`.

followed by the free energy computation.

The HyperQ pipeline tool `hqf_gen_run_one_pipe.sh` executes the various components for a single MSP in serial. However, each MSP contains a number of TDSs which form the alchemichal pathway, and the corresponding simulations can be run independently from each other. Therefore the pipeline tool `hqf_gen_run_one_pipe.sh` allows to specify which of the TDSs of the MSP should be run, and moreover how many of them in parallel. If there are more TDSs than parallel slots the remaining ones will automatically run in serial, i.e. will not start until another simulation has finished. In addition it is possible to specify how many CPUs are used by each simulation (in particular CP2K). The corresponding settings are all located in the main HyperQ configuration files located in `input-files/config`. All these mechanisms together allow to run precisely as many simulations in parallel and serial as is suitable for the machine it is running on. Moreover, it allows to run different parts of the same MSP on different computers.

Managing Multiple MSPs. For the case that one has many MSPs and wants to manage the execution of an arbitrary subset of them at the same time, HyperQ provides the commandline tool

`hq_gen_run_all_pipe.sh`,

which itself makes use of the command `hq_gen_run_all_pipe.sh`. This above command allows one to run the same pipelines specification on all MSPs which have been provided (via a text file). This command provides an additional level of parallelization by allowing to specify the of MSPs which are allowed to run in parallel, which causes the remaining MSPs to be queued automatically. This can for example be useful when one wants to lead all the MSPs in sync through the individual components of the pipeline. An example is illustrated in Figure 7.13 on the facing page.

7.5.2 Batchsystem Mode

Due to the high computational costs HyperQ is normally only run on a single general-purpose computer for testing purposes or for very small systems. For

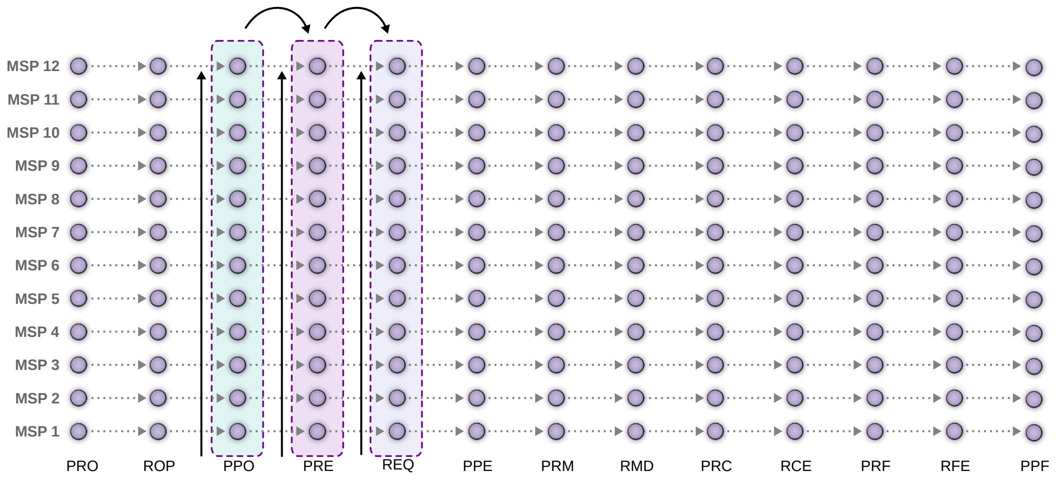


Figure 7.13: Example of how the command `hqf_gen_run_all_pipe.sh` can be used to run multiple MSPs in synchronization through the individual pipeline components. The number of MSPs which are supposed to run in parallel can be specified, as can the range of the TDSs which should be affected.

most production runs it needs to run on a high performance computer such as computer clusters managed by DRMS. Regarding such clusters, ideally it should be possible to run HyperQ on clusters

(i) with any batchsystem type and configuration,
(ii) in a way which allows high scalability,
(iii) as flexible and convenient as possible.

For achieving these goals and additional features a DRMS module for HyperQ was developed.

Serialization and Parallelization

In order to be able to run a large numbers of simulations as efficiently and quickly as possible on the available cluster resources, the batchsystem module has been designed such that it allows to run the MSP-dependent components of the HyperQ workflow (i.e. components (2) to (6)) in a parallel as well as serial fashion in a complementary way. For this purpose a multilevel hierarchical architecture was designed which consists of the following components:

Multiple Workflows per Cluster. A HyperQ workflow can be given a unique workflow identifier (WFID), which allows to run multiple workflows at the same time on the same cluster without causing interferences.

Multiple Jobs per Workflow. In each HyperQ in theory any number of independent batchsystem jobs can be run in parallel. Each job has an internal job identifier (SJID) in HyperQ.

Multiple Subjobs per Job. Each batchsystem job can have multiple subjobs, where each subjob represents a job step in batchsystem such as SLURM. Each subjob has its unique subjob identifier (JID). Normally subjobs are required within HyperQ to distribute the resources of large multinode jobs among the allocated resources (in most cases one subjob is required per compute node). Subjobs are for example required when jobs need to have a certain amount of minimum nodes, which is the case in particular for some supercomputers (such as the HLRN).

Multiple Tasks per Subjob. Tasks are used within subjobs to run specific pipelines on single nodes (either all CPUs of the node or some subset of them). Each task has a task identifier (TID).

Multiple Job Serial Numbers Per Job. Each job (and with that its subjobs and tasks) has a job serial number (JSN). The serial number allows the same job to run multiple times while keeping track of the iteration. Running the same job multiple times is for example useful when the finished job has not yet completed its task, or if the simulations should be extended. For this purpose the continuation mode can be used of the standalone part of HyperQ, which will automatically continue the simulations where they have been stopped. Therefore exactly the same job/task can be run without modifications to extend previous simulations.

Multiple Jobtypes per Job. The HyperQ pipeline consists of a variety of different components, from geometry optimization to the free energy computation. These components have different demands on the computational resources. By far the most resources are normally required by the PIMD simulations as the number of degrees of freedom is multiplied by the number of path integral beads. Due to these differences on the computational resources, different classes of jobs are required for different pipeline components. Therefore each job is assigned a to a certain job class, and each class is identified by a jobtype letter (JTL) (a lower case Latin letter).

The above design allows to have a unique JID for each MSP (or a specific range of its TDSs), independent of the job iteration/serial number or the pipeline component. This makes the handling of the various batchsystem jobs more intuitive and simple.

The above hierarchy consists of six levels. If we only consider the batchsystem jobs, not the subjobs or tasks, then four dimensions remain. These are visualized in Figure 7.14 on the next page. Therefore the batchsystem jobs can be interpreted as being organized in a four dimensional space. If in addition the subjobs and tasks which each job has are considered, these can be seen as being hidden inside each of the small cubes which represent the individual jobs, implying that the tasks represent the sixth dimensions of the hierarchy. Subjobs and tasks can both be set up to run in serial or in parallel. These options already give considerable flexibility. Since the batchsystem module is based on the standalone mode of HyperQ, all its serialization and parallelization layers become available in addition to the layers which the batchsystem module itself provides. Altogether they comprise eleven, which are summarized below.

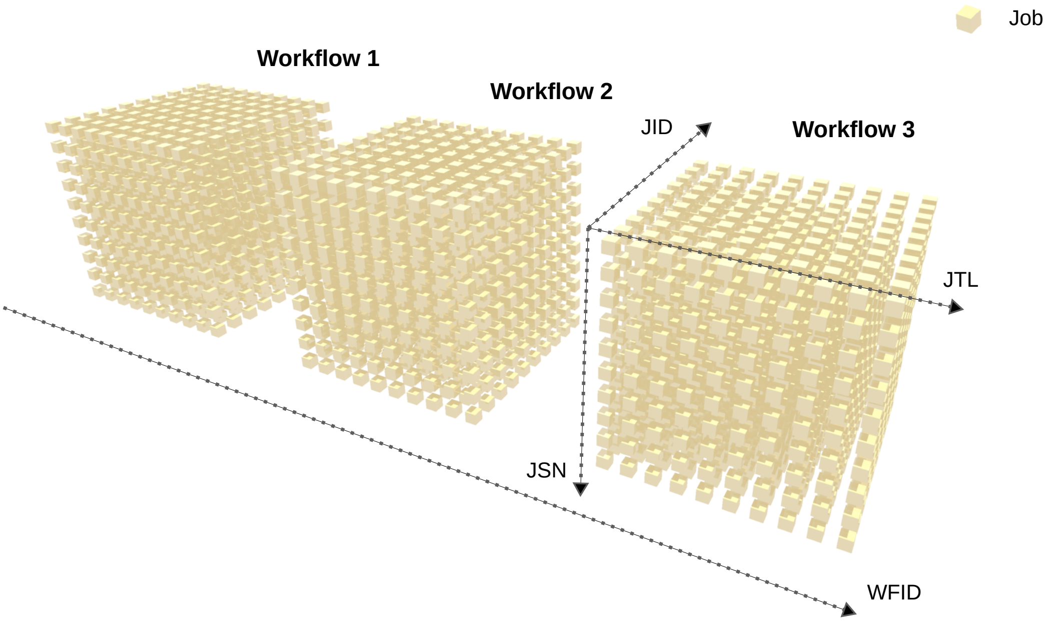


Figure 7.14: Top four levels of the hierarchical batchsystem job architecture. Multiple workflows can coexist on the same cluster, each having a unique workflow ID (WFID), corresponding to three macro cubes in the figure. Each workflow consists of jobs which have a job ID (JID), job serial number (JSN), as well as job type letter (JTL).

(1) Multiple workflows per cluster (WFIDs).
(2) Multiple jobs per workflow (JIDs).
(3) Multiple subjobs per job (SJIDs).
(4) Multiple taks per subjob (TIDs).
(5) Multiple jobtypes per job (JTLs).
(6) Multiple serial numbers per job (JSNs).
(7) Pipeline serialization/parallelization over MSPs and TDSs ranges (`hqf_gen_run_all_pipe.sh`).
(8) Pipeline serialization over multiple workflow components (`hqf_gen_run_one_pipe.sh`).
(9) Pipeline serialization/parallelization over TDS of the same MSP (`hqf_<sim-type>_run_one_msp.sh`).
(10) Path integral replica parallelization (one CP2K instance per bead).
(11) External code parallelization (multihreading, OpenMP, MPI, ...).

The tenth level, i.e. the path integral replica parallelization, can currently not be deactivated within HyperQ. However, i-PI is generally able to run with any number of CP2K instances connected to it, providing the in theory the possibility to use less CP2K instances as there are beads.

Additional Features

Automatic Preparation. The number of files required for running the HyperQ workflow on cluster via batchsystems is relatively large. At first there are the primary job files for each single job. Secondly, each job has subjobs which are listed in the subjob list files (one per job). Furthermore, each subjob consists of a number of tasks, which are defined in individual subjob files (one per subjob). All these files need to be prepared for each application according to the specific needs. The batchsystem module of HyperQ provides several commandline tools which can prepare and set up all these files semi-automatically.

Autonomous Job Management at Runtime. When a job of a certain jobtype is running for a certain MSP (or a range of its TDS), then a number of things can happen:

(1) The job runs without errors and completes its task before reaching the walltime.

(2) The job runs without errors but does not complete its task before reaching the walltime.

(3) The job experiences errors during the runtime. Errors can happen due to problems of the cluster/batchsystem, within the primary jobfile, within the subjobs, or within the HyperQ pipeline.

Depending on which of the above cases is occurring, the desired action can vary. It can depend for example on the job type or in the case of an error on the specific type. When running possibly thousands of jobs in parallel, the task of simply responding to the above job terminations in the appropriate way can become unfeasible. Therefore HyperQ was equipped with the feature to be able to respond to job terminations autonomously. The actions which it should take can be specified, depending on various criteria such as the job type, the job ID, or the nature of the job termination. The actions which can be taken are to do nothing, to prepare the next job but not submit it, to submit the next job, and moreover it is possible to specify of which jobtype the next job should be. In this way one can elegantly set up the entire batchsystem workflow, allowing HyperQ to move along specific paths within the three dimensional cubes shown in Figure 7.14 on the preceding page.

As an example, the batchsystem workflow can be set up as follows. Jobs are at first started on the geometry optimization level.The simulations can then progress from the optimizations to the PIMD simulations. The PIMD simulations can be set up such they are very long, possibly longer than the walltime of the jobs, in which case jobs will start continuation PIMD jobs. While the PIMD simulations are still running, cross evaluation jobs together with free energy computations (`_allce_allfec_`) can be started together in single pipelines and can run continuously in parallel to the PIMD simulations, steadily carrying out cross evaluations of new snapshots and recomputing the free energies. This has the advantages that the convergence of the simulations can be monitored in real time and that time is saved since energy cross evaluations can be computationally very expensive when a more accurate method is used within reweighting (umbrella sampling) approaches. This example scheme illustrated in Figure 7.15 on the next page.

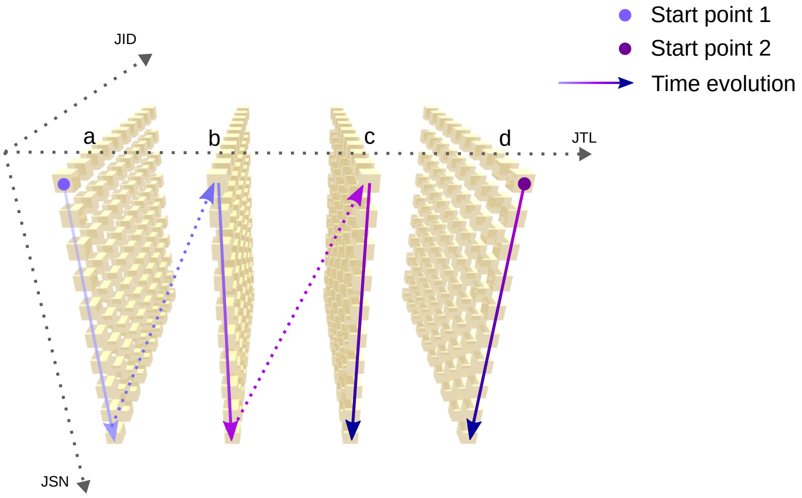


Figure 7.15: Example setup of the batchsystem workflow within HyperQ using four types of jobs (JTLs a to d). Jobtype a comprises the optimizations, jobtype the b equilibrations, jobtype c the PIMD simulations, and type d the cross evaluations followed by the free energy computations. Jobs are started at the geometry optimization level as well as the cross evaluation level, and are running in parallel with the PIMD simulations to continuously carry out the cross evaluations. The colors of the arrows indicate the time point.

When the free energies seem to be converged sufficiently the entire workflow can be halted. The more detailed HyperQ job management policy might be designed as follows:

(i) If the optimization of an MSP (or TDS range) is not completed when the walltime is reached, prepare the job for the next iteration and start it again.

(ii) If the optimization of an MSP (or TDS range) is completed before the walltime is reached, start the equilibration.

(iii) If the equilibration of an MSP (or TDS range) is not completed when the walltime is reached, prepare the job for the next iteration and start it again.

(iv) If the equilibration of an MSP (or TDS range) is completed before the walltime is reached, start the PIMD simulation.

(v) If the PIMD simulation of an MSP (or TDS range) is completed before the walltime is reached, do nothing.

(vi) If the PIMD simulation of an MSP (or TDS range) is not completed when the walltime is reached, prepare the job for the next iteration and start it again.

(vii) If the cross evaluation and free energy computation of an MSP is completed before the walltime is reached, prepare the next job and start it again.

(viii) If the cross evaluation and free energy computation of an MSP is not completed when the walltime is reached, prepare the next job and start it again.

(ix) If an error occurs of any type during the simulations except that the job is aborted due to the requeuing mechanism of the batchsystem, do nothing. (Then we can investigate the error before manually continue the job.)

(x) If an error occurs because the job has been terminated by requeuing mechanism of the batchsystem, prepare the new job but do not submit it. (It will be submitted automatically by the batchsystem.)

Flexible Workflow Control. The HyperQ batchsystem settings/control files are stored in the folder `batchsystem/control` (a default configuration file can be found in B.2 on page 253).

The batchsystem workflow can be controlled both dynamically during the runtime. Some HyperQ batchsystem options allow to control the jobs nearly in real time, for instance they can be stopped in a clean way at any time (without using the batchsystem itself to cancel the jobs, which may result in abrupt job terminations, moreover HyperQ might restart the next jobs by itself in this case if it was setup to do so during certain types of errors). Additional batchsystem parameters are synced from the control files each time before a job is started (or restarted), as for example the number of CPUs, the memory or the walltime.

The batchsystem workflow in HyperQ might involve a large number of jobs and TDSs to simulate. As outlined before, there are different types of jobs requiring different computational resources, and therefore also multiple control files are required. In addition, HyperQ enables us to have individual control files for jobs with specific job identifiers to manage the computations of specific MSPs or TDSs.

Parallel Robustness. Different jobs in HyperQ are generally independent from each other. However, there is the possibility of clashes on the file system (as it has turned out in test simulations). Cluster file systems are often slow and exhibit delay times of several seconds. Thus when one job creates or deletes a file, another job might not see this change yet, respond to it, while in the meantime the situation has changed. This can lead to errors and job abortions. Clashes can occur in HyperQ because each subsystem folder of the simulation components (i.e. `<simulationtype>/<MSP>/<subsystem>/`) contains files which are shared among the various TDSs of each MSP. Therefore if the simulations of different TDSs of the same MSP are running in parallel with different jobs, their workflow might arrive at the next type of simulation at the same time, causing both of them to prepare the common files of the joint system. To prevent this two mechanisms were implemented, most importantly a delay time and signposting mechanism, which prevents virtually all clashes (tested in practice).

Supported Batchsystems. Currently supported are the DRMSs SLURM, Moab/TORQUE/PBS, LSF, and SGE. HyperQ can be easily extended to other batchsystems due to its modular and generic jobfile structure. The major file which needs to be provided for new batchsystems is the top level job template.

7.6 Test Simulations

In order to test whether the implementation of HyperQ works as intended it was applied to a number of test systems for which experimental free energies are available as reference values. Here we will only describe two of them in more detail, the relative solvation free energy between methanol to ethane, as well as the relative binding free energy of two small organic molecules involving a low molecular weight host system.

Methanol-Ethane Relative Solvation Free Energy

To compute the relative solvation free energy between methanol and ethane the thermodynamic cycle described in Figure 7.2 on page 121 has been applied where one of the solvents is vacuum and the other one water. Two dummy atoms are needed during the transformation to compensate for the different numbers of atoms (see also Figure 7.16).

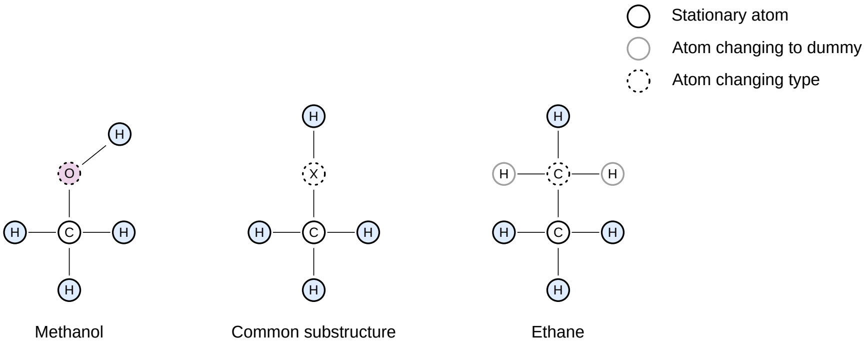


Figure 7.16: Alchemical transformation between methanol and ethane. The single topology approach is applied in which the oxygen atom of methanol is transformed into a carbon atom of ethane. The two excess dummy atoms of ethane are transformed from/into dummy atoms within the augmented methanol molecule.

During the creation of the solvated systems the molecules were placed in a water box such that the distance from the solutes to each boundary wall is at least 10 Å, resulting in 293 water molecules. Periodic boundary conditions were applied for both the L and the LS system, and both of them modeled on the MM level with the CHARMM36 force field. Five independent runs of the simulations have been carried out for each system, involving 500 steps of geometry optimization, 100 ps of equlibration, and an MD simulation (NVT) of more than 1 ns in length. The temperature was 300 K, the integration time step 1 fs, and the friction coefficient γ of the Langevin thermostat 5 ps^{-1}. A single path integral bead was used, and the alchemical transformation was carried out in a single step. The free energies were computed with the BAR method where the energies of every 500th time step were used. The results are summarized in table 7.1 on the next page. The results are in excellent agreement with experimental measurement, and the error is of the same order of magnitude as others have obtained when using single-topology approaches [33]. Notably we have obtained our results with a single alchemical step only. The characteristic plot associated to the two BAR equations (as introduced already by

Quantity	*BAR (*HyperQ*)*	*Experiment*
$\Delta\Delta A_{\text{solv}}$	6.77 ± 0.91	6.93

Table 7.1: Relative solvation free energy between methanol and ethane in kcal/mol. Predicted values are based on five independent runs. Experimental values retrieved from [19].

Bennett in his original publication [20]) are shown in Figure 7.17 for one of the LS systems.

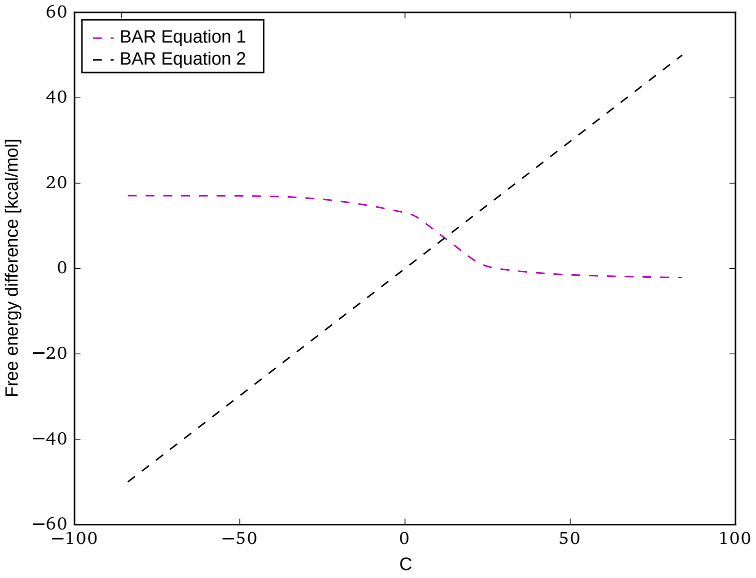


Figure 7.17: Plot of the two characteristic BAR equations of one of the LS simulations in which methanol was transformed to ethane. The intersection of the two lines yields the free energy difference predicted by the BAR method.

Cucurbit[7]uril-Guest Relative Binding Free Energy

The system involving the transformation of methanol into ethane as described above has tested simulations of the subsystems L and LS, as well as the major components of the workflow, from structure preparation to the free energy computations. In order to test HyperQ also for cases involving subsystems of the type RLS and in situations where the QSTAR method is applied, the binding free energies of a host-guest system were computed. In addition the PEARL scheme was applied.

The system which was chosen is a host-guest system in which the host is cucurbit[7]uril. The advantages of a host system instead of full biomolecular protein is that it is relatively rigid and small, thus minimizing the computational efforts by allowing for shorter sampling times as well as smaller system sizes. Still it allows to emulate the pocket of a biological macromolecule such as a protein and the binding of small organic molecules. The host is shown in Figure 7.18 on the next page.

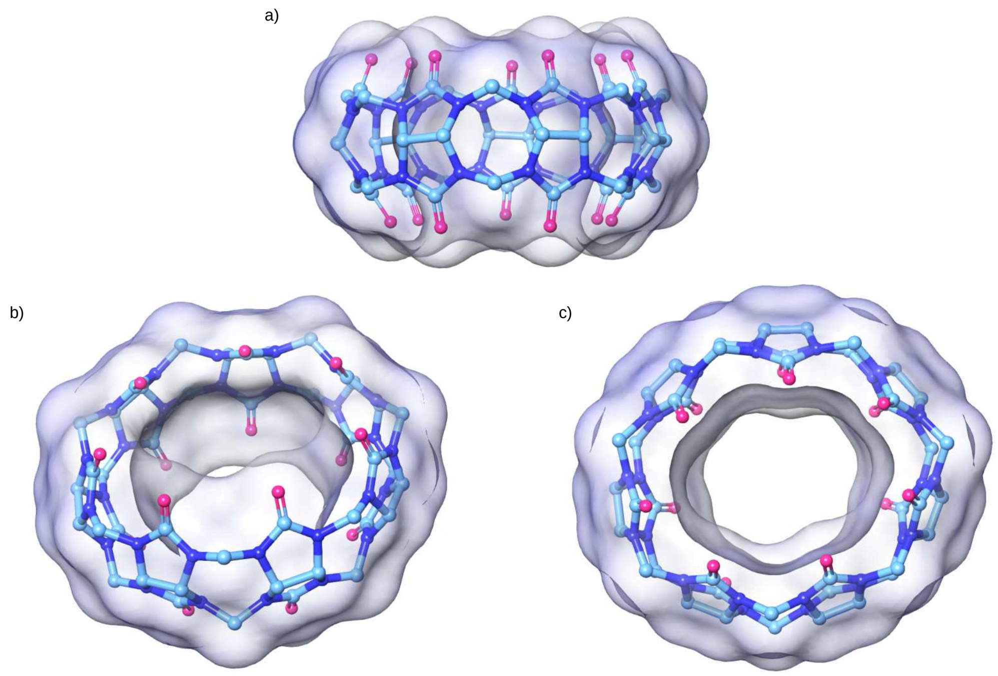


Figure 7.18: The host molecule cucurbit[7]uril shown from different perspectives. The tunnel has the appropriate size for many small organic compounds to fit in, rendering the host suitable for binding organic compounds of low molecular weight.

The ligands (guests) which are used for computing the RBFE are *adamantane-1,3-diamine* and *1-adamantanecarboxylic acid*. These compounds have a common three-dimensional cage structure which is the major part of the CSS (illustrated in Figure 7.19).

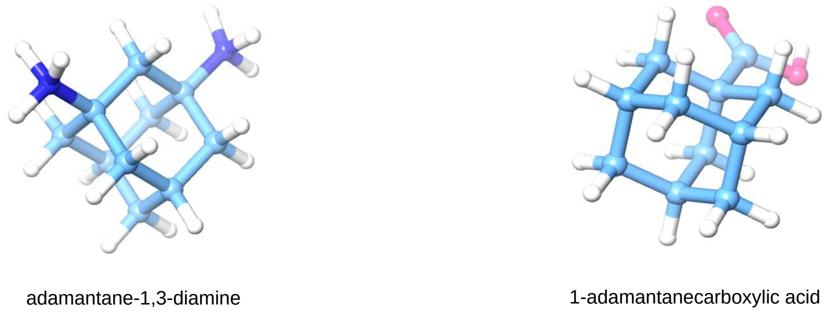


Figure 7.19: The molecules *adamantane-1,3-diamine* and *1-adamantanecarboxylic acid* are the two guest molecules for the RBFE will be computed.

In order to test the unique nature of the QSTAR method more than a single path integral bead has to be used. We tested several different scenarios involving up to 8 beads, the results are shown in table 7.2 on the next page.

With one or very few thermodynamic windows it can be expected that the results deviate strongly from the experimental values due to the virtually non-existent phase space overlap between the adjacent states. The more intermediate states the higher the overlap and thus the better the convergence behavior.

Beads	*States*$^{\mathbf{a}}$	*Alchemical Scheme*	*Macroscheme*	$\Delta\Delta A_{\text{bind}}(HyperQ)$
1	2	ι-scheme	no	-42.5 ± 8.2
2	3	ι-scheme	no	-28.1 ± 5.1
4	5	ι-scheme	no	-11.8 ± 3.9
8	9	ι-scheme	no	-15.2 ± 4.4
4	5	PEARL-B	no	-2.8 ± 1.9
4	7	PEARL-B	yes$^{\mathbf{b}}$	-0.1 ± 1.0

Table 7.2: Relative binding free energies between *adamantane-1,3-diamine* and *1-adamantanecarboxylic acid* in kcal/mol. $^{\mathbf{a}}$Number of states during the thermodynamic pathway. Experimental binding free energy: $-0,26 \pm 0.11$ kcal/mol [202]. $^{\mathbf{b}}$The charges of the atoms which are transformed in dummy atoms as well as of their direct neighbors were decoupled/recoupled in a single step prior/after the PEARL and QSTAR alchemical transformations, adding two addition thermodynamic windows.

However, in our test case more than 4 beads and 5 thermodynamic states did not further improve the results, possibly indicating that the phase space overlap does not further improve by simply increasing the number of states and/or beads. The cause of the considerable deviations from the experimental values, as well as the high standard deviations, are probably related to endpoint problems due to the appearance of new atoms which can clash with already existing atoms. For such problems the PEARL scheme was designed, which indeed improved the predicted relative binding free energies considerably. It was possible to improve the results further by decoupling/coupling the charges of the atoms which were transformed into dummy atoms in separate step. The resulting macroscheme is of the type $\lambda - \sigma - \lambda$ (see subsection 4.4.3 on page 87 for more details). The total number of TDSs which were used is relatively small, commonly more than are used, but sometimes even several dozens as for example in [7].

Simulation details: Solvent box padding size (to the solute) 10.0 Å. Geometry optimization: 500 steps. NPT equilibration: 100 ps. NVT PIMD simulations: 100 ps to 1000 ps. The thermostat used was a Langevin thermostat with a friction coefficient of $\gamma = 5\text{ps}^{-1}$. The cutoff value for nonbonded interactions was set to 12 Å. The tests shown here were carried out for qualitative purposes to verify the software works as intended and that the implementation is correct. However, more extensive testing should be carried out in the future (see also the discussion in the last chapter 12 on page 199), which is a project of its own.

Chapter 8

Virtual Flow

An expert is a person who has made all the mistakes that can be made in a very narrow field.

Niels Bohr

Contents

8.1 Introduction

In chapter 5 on page 91 about virtual screenings it was shown the scale of the virtual screening does play a role regarding both the binding affinity of the final hit compounds as well as the true hit rate. Moreover multistage virtual screening procedures were outlined and its benefits described.

Because of the vastness of the chemical space of small organic compounds there exists a virtually infinite number of compounds which could in theory be screened, even when the chemical space considered is constrained by additional criteria. Moreover, the computing power which is available to researchers has increased substantially over the past decades, partly due the circumstance that HPCs of various types and sizes became more and more available.

In order to carry out multistage structure-based virtual screening procedures on extreme scales special software is required. Regarding computer clusters, one of the primary classes of HPCs (into which most supercomputers fall as well), no suitable code was available which was able to carry out the envisioned virtual screenings. Therefore we have developed a new virtual screening suite named Virtual Flow which provides all the desired features, in particular perfect scalability and high automatization. Many software tools exists already within the field of CADD, including innumerable docking programs [71, 262, 173, 62, 256, 183, 54, 78]. However, most of these programs run only on a single computer. Therefore it seems that one type of virtual screening suite which would be particularly useful would be workflow software which can use existing external programs designed for single computers to carry out the individual docking procedures while using many instance of them in parallel on the cluster.

The number and sizes of the databases which can be used as ligand-sources have grown significantly. One of the most attractive databases for virtual screenings is the ZINC database [130, 268], containing at the time of writing already more than 600 million compounds. Other examples are the members of GDB family of databases [243] containing multiple billions of molecules created specifically for virtual screenings. However most of the molecules are available only as SMILES, the most common chemical line notation format. This is mostly true for the GDB databases, but also to some extend for the ZINC databases. As an example, in early versions of the ZINC 15 database (when this project was in the beginning) only approximately 10 % of the compounds were available in 3D format.[1] Since most docking programs require ligands in some three dimensional format, sometimes even in a specific molecular file format not provided by the ligand database, it can be possible in many cases that the ligand database needs to be prepared/curated by the user before carrying out the virtual screening procedure.

The preparation of ligand databases and the carrying out of the virtual screening procedure itself are two tasks of a very similar nature. Both of them require the processing of very large numbers of ligands (or elementary tasks) which are independent from each other, and require for each specific task an input file and generate some output files. Therefore the same underlying technique can be used to carry out both types of tasks, and for this reason two versions of Virtual Flow were developed which share the same core architecture (outlined in section 8.2). The first version which is dedicated to the preparation of ligand databases is Virtual Flow for Ligand Preparation (VFLP), subject of section 8.3 on page 153, which currently support the AutoDock-based family of docking programs. The second version, Virtual Flow for Virtual Screening (VFVS), is tailored for the virtual screening procedure itself (described in section 8.4 on page 155).

8.2 Common Core Architecture

In this section the general architecture and features of Virtual Flow are described.[2]

Features

- **Scalability.** Parallelization and scaling without bounds regarding the CPUs.
- **Any Cluster Configuration.** The workflow tool can use any conceivable hardware configuration regarding the number of cores/CPUs, the number of sockets, the number of nodes, etc. The software runs on any Linux

[1] At the time of writing it has risen to around 40 %.

[2] While they are all implemented in the latest version of VFVS, a few of them are not yet implemented in the latest version of VFLP. These updates are relatively trivial and are planned for the near future.

distribution

- **Automatized Workflow.** Virtual Flow can run fully automatically for any duration until it has processed all ligands specified in the input files by autonomously submitting new jobs into the batch system.
- **Monitoring.** Realtime monitoring of the workflow.
- **Realtime Control.** The workflow can be controlled and modified during the runtime (e.g. adjusting the hardware configuration, pausing and continuing the workflow).
- **Automatic Archiving and Compression.** Automatic decompression and unpacking of input files, compression and archiving of the output files.
- **Robustness.** Virtual Flow is robust regarding unexpected errors and interruptions. It was tested extensively on various clusters of different types. Virtual Flow can respond to signals sent to it by the (batch or operating) system in the case of cluster problems (which are not uncommon), but even after termination without warning it can simply be restarted.
- **Multiple Batchsystem Types.** Virtual Flow currently supports the following DRMS: SLURM[3], Moab/TORQUE/PBS[4], LSF[5] (VFVS only) and SGE[6] (VFVS only). It can be easily extended to other job schedulers by creating additional job-templates without the need to change any of the core or front end components.
- **Availability.** VFLP and VFVS are free and open source software.

Implementation

Parallelization. In order to allow for optimal scalability on computer clusters of any configuration and batchsystem type Virtual Flow uses four hierarchical levels of parallelization. One Virtual Flow instance can employ multiple jobs, one job is able to use several job steps, one job step may run any number of queues, and one queue employs the external docking programs which can have their internal parallelization (e.g. multithreading). The multilevel architecture is illustrated in Figure 8.1.

Collections. Due to the potentially vast number of ligands to be screened, the total number of input and output files could become extremely high in large-scale production runs. A virtual screening of 100 million ligands will roughly involve 1 billion files. Such large numbers would lead normally to at least two major problems if directly stored on the disks. At first, transferring/copying millions of smalls files is often extremely slow in comparison to copying the same amount of data stored in a few files. And secondly, the (file) systems themselves frequently run into serious trouble if the number of files is too high (in particular if they are located in the same folder). Therefore on many clusters the number of files which are allowed per user is often limited (less than one million is not uncommon). To

[3] `http://slurm.schedmd.com`

[4] `http://www.adaptivecomputing.com`

[5] `https://www.ibm.com/us-en/marketplace/hpc-workload-management`

[6] `http://gridscheduler.sourceforge.net/`

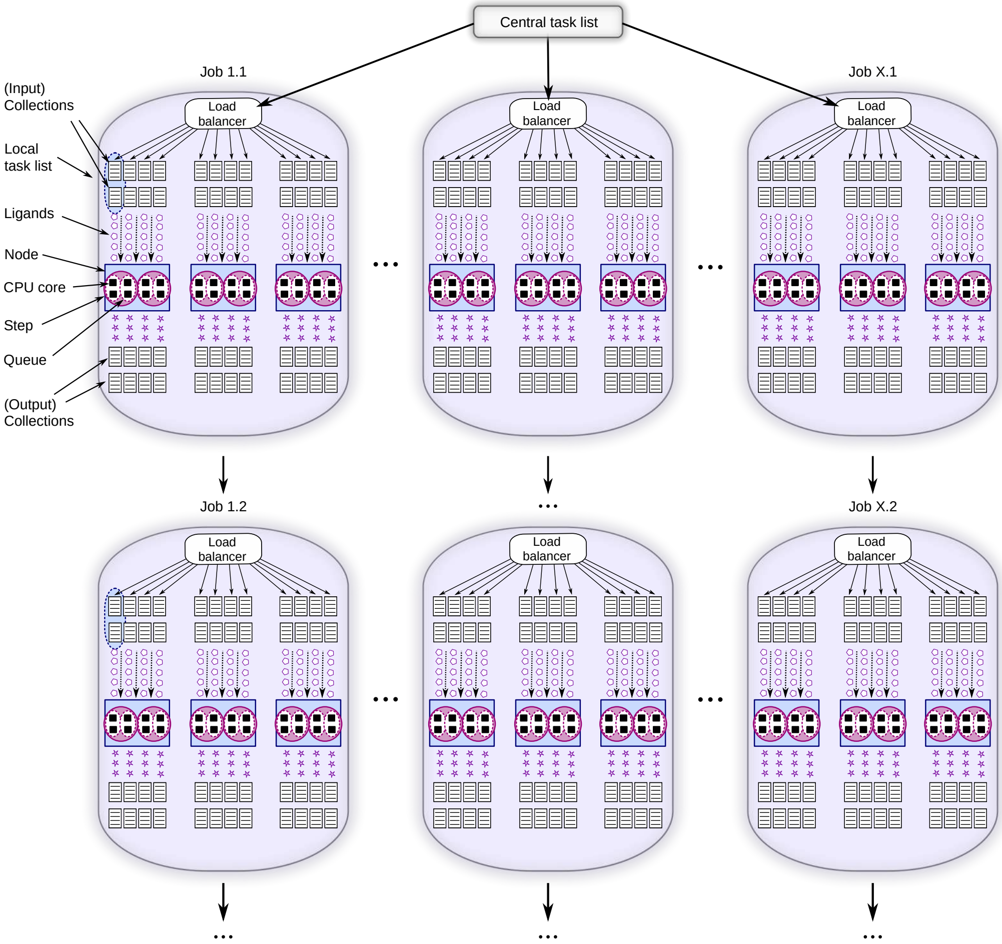


Figure 8.1: Overview of the workflow within Virtual Flow in terms of the hardware, hierarchical levels of organization, task lists and collections. Any number of batchsystem jobs can be employed in parallel, and after individual jobs have ended successive jobs are started automatically. One item of the task lists corresponds to one ligand collection. There is one central task list as well as local task lists associated to each queue.

address these issues regarding virtual screenings, in Virtual Flow the input and output files are organized in multidimensional archives (i.e. archives of archives of ...) called collections. Input and output collections (databases) consist of three layers consisting of tranches, (elementary) collections and the ligands themselves. In addition the files are compressed, which usually decreases the size of the files by more than one order of magnitude.

Workload Balancing. Because the ligands are processed in parallel a mechanism is needed which makes sure that every ligand is processed exactly one time in the workflow. This task, while sounding theoretically trivial, was one of the major challenges during the development of Virtual Flow. To simply work off a common list of ligands does not work well on most clusters, in particular not if thousands of queues are working in parallel. The major reason lies in the fact that cluster file systems are often very slow (delays of several seconds are not uncommon), rendering also locking mechanisms useless. These delays are normally not present within single jobs due to the caches the cluster file systems employ, but they are often present between independent jobs. More advanced approaches such as using message-passing systems (e.g. MPI or OpenMP) are also no option because they allow only intra-job communication, but not inter-job communication.

We solved this issue by the introduction of the before mentioned collections in concert with the implementation of a workload balancer which distributes collections at the very beginning of each job. This reduces access frequency on the central task to an absolute minimum of only a single time per job, making the jobs completely independent during their runtime, which allows for a perfect linear scaling without virtually upper bounds. Furthermore two backup mechanisms were implemented just in case. One for the case that two jobs should try to access the central task list at the same time. This is done by dispersing arriving jobs in time. The second backup mechanism saves the last valid version of the central task list in a backup folder, and in the case that the central task list should despite the other measures get lost or wiped (which are the most common effects when two processes try to access the task list at the same time) automatically the last backup version is restored and used. Further backup versions can be restored manually if needed.

Reduction of I/O Load. When a large number (possibly tens of thousands in the case of Virtual Flow) of small programs are running in parallel, the generated I/O workload can easily bring the cluster file system down. Sometimes more sophisticated prevention mechanisms of clusters prevent I/O problems by limiting the maximum amount of I/O per user, in which case the running software exceeding this limit would be slowed down. In order to avoid such problems Virtual Flow carries out as many operations and processes automatically on the local memory-based tmpfs which is available on almost any Linux system by default, thus relieving the cluster file system.

BASH. Virtual Flow is implemented in BASH because shells are specialized on handling files, data flows and running external programs, which are all essential tasks when dealing with workflows. Moreover, BASH has essentially no overhead, is readily available on almost any Linux distribution, and its code can be easily modified by the user.

8.3 Virtual Flow for Ligand Preparation

VFLP is dedicated to the preparation of ligand databases, and provides them in a format which is directly usable (e.g. regarding the structure of the input collections). But VFLP can be used independently of VFVS and virtual screenings in general, in particular since it is able to provide the ligands in almost every possible chemical file format.

Implementation

The first step of the VFLP workflow consists of the computation of the major microspecies, i.e. the major protonation and tautomeric state of the ligand at physiological pH (7.4). This is done with the tool `cxcalc` which is part of the *Marvin* package of ChemAxon [7], after which the three-dimensional structure of the ligand is computed by ChemAxon's `molconvert` and stored in the PDB format. If this conversion fails (which may happen with certain ligands) another attempt is made using Open Babel [212]), which also carries out the conversion into the desired final output format. The workflow diagram of a single queue is shown in Figure 8.2.

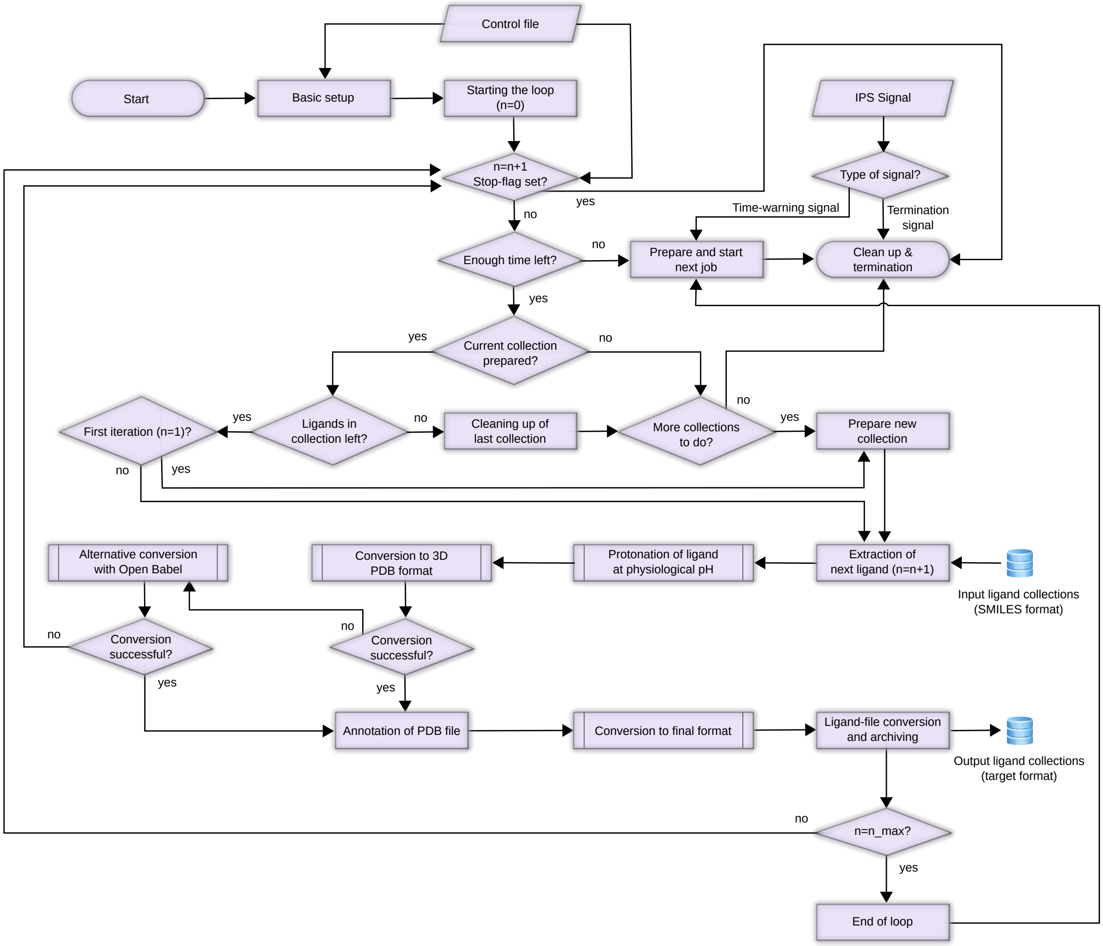

Control file
Start
Basic setup
Starting the loop (n=0)
IPS Signal
Type of signal?
n=n+1 Stop-flag set?
yes
no
Time-warning signal
Termination signal
Enough time left?
no
Prepare and start next job
Clean up & termination
yes
yes
Current collection prepared?
no
no
First iteration (n=1)?
yes
Ligands in collection left?
no
Cleaning up of last collection
More collections to do?
yes
Prepare new collection
no
yes
Alternative conversion with Open Babel
Conversion to 3D PDB format
Protonation of ligand at physiological pH
Extraction of next ligand (n=n+1)
Input ligand collections (SMILES format)
no
Conversion successful?
no
Conversion successful?
yes
yes
Annotation of PDB file
Conversion to final format
Ligand-file conversion and archiving
Output ligand collections (target format)
no
n=n_max?
yes
End of loop

Figure 8.2: Simplified workflow diagram of a single queue in the VFLP software. The workflow until preparation of the collections is relatively generic and similar in all Virtual Flow implementations. The VFLP specific tasks start after this step (i.e. the ligand extraction step).

[7] `www.chemaxon.com`

Results

For testing and simultaneously applying VFLP to a real problem, we have converted more than 130 million ligands from the ZINC 15 database (at the time of application only a small fraction of the database was available in a ready-to-dock format) on clusters of our institute where one node has two Intel Xeon X5570 processors. The conversion time of a single ligand is almost the same on a desktop computer with an i5-2320 processor when only one core is used in both types of systems. In average 105 nodes (i.e. 840 cores) were used in parallel during the conversion, which took approximately 17 days (∼ 350 000 core hours). VFLP was later applied a second time after the ZINC database was extended, and all 350 million ligands were converted for use with VFVS. In both cases the scaling was virtually linear with respect to the number of CPUs used. The scaling behavior will be illustrated in the next subsection about VFVS.

Discussion

With VFLP a workflow tool was created which allows the efficient use of high-performance computers for the preparation of ligand databases of any size and type, and thus it can be used within individual virtual screening procedures as well as for the creation of the publicly available ligand databases of the future. While the currently available ligand databases providing ready-to-dock ligands already contain millions of molecules such as ZINC, this number becomes vanishingly small when considering the chemical universe of small organic molecules (as pointed out in section 5.2 on page 94). Hence tools that facilitate ligand preparation as the one presented here will likely prove valuable in the future in general, which can be used to prepare databases like the GDB which contain billions of compounds [29, 243]. In addition VFLP can be used in seamlessly concert with VFVS. VFLP is available at `https://github.com/cgorgulla/VFLP`.

8.4 Virtual Flow for Virtual Screening

VFVS is dedicated for carrying out the virtual screening procedure itself, and can for example use the ligand collections prepared by VFLP.

Features

- **Multiple Docking Programs.** VFVS is able to use various AutoDock-based docking programs, at the moment AutoDock Vina [294], QuickVina 2 [112, 8], Smina (Vinardo) [150] and ADFR [235]. It can easily extended to accommodate other docking programs.
- **Multiple Docking Scenarios.** VFVS allows to carry out multiple docking scenarios per ligand. A docking scenario in VFVS is defined by the receptor structure, by the docking parameters (such as exhaustiveness), rigid or flexible receptor docking, the choice of flexible receptor side chains or the docking program. Since multiple receptor conformations can be used, VFVS

can carry out the so-called ensemble dockings. Therefore VFVS is able to carry out dockings involving both side chain flexibility as well as backbone flexibility. This feature can be very important since biological macromolecules are often relatively flexible, and the flexibility can have a significant influence on the binding process.

- **Multiple Replicas.** Each docking scenario can be carried out multiple times. This possibility can be very useful when one wants to increase the chance that the docking program finds the docking pose with the global minimum relative to the scoring function which it employs. Even though most docking programs allow to adjust the exhaustiveness of the search space, in our tests carried out with AutoDock-based docking programs it has turned out that it is considerably more likely that the global minimum is found when multiple runs of lower exhaustiveness are carried out than few with higher exhaustiveness.
- **Realtime Docking Results.** Even while the virtual screening procedure is still running, VFVS provides the possibility to see results in realtime for each docking scenario which was defined for the screening procedure. It can provide both statistical information regarding the docking scores of all docked ligands, as well as list the highest scoring compounds along with their highest score.
- **Tools.** A separate tools-package (VFTools) for Virtual Flow was created, which contains mostly tools for VFVS which can assist to create the ligand collections in the required layout (provided the ligands are already in the correct format), as well as to automatically postprocess and curate the output files.

The feature of VFVS to run multiple docking scenarios and replicas renders the software also useful for docking a few ligands very thoroughly rather than carrying out a virtual screening procedure.

Comparison to other Virtual Screening Software

There exist several other programs which allow to carry out virtual screening procedures with AutoDock-based docking programs. Among them are VcPpt [26], VSDK [12], AUDocker LE [244], DockingApp [69], DockoMatic [42, 43], Autodock4.lga.MPI [64], VinaMPI [82], VinaLC [352], VSDocker [229], MOLA [2], and DOVIS [351, 133]. But none of them is able to allow to carry out largest-scale virtual screening procedures on computer clusters as efficiently as VFVS.

Among these programs the following are not able to run on (multiple nodes of) Linux-based computer clusters. AUDocker LE, Vina interface of UCSF Chimera, DockingApp, VcPpt, VSDK, and AUDocker LE all run only on single machines running MS Windows. DockingApp is platform independent but also is limited to a single computer. VSDocker can run on computer clusters, but only on Windows-based clusters which are at first rarely found, and secondly not optimal for scientific research.

MOLA is able to run on Beowulf type clusters, which consist usually of a number of Desktop style computers connected via a local area network (LAN). MOLA works inside a customized Live CD GNU/Linux operating system which needs to be booted on every computer of the cluster.

Autodock4.lga.MPI, VinaMPI and VinaLC are Message Passing Interface (MPI)-based versions of AutoDock, which allows them to run on multiple nodes of computer clusters. However, they suffer from several significant disadvantages. Because they are based on MPI, the virtual screenings need to be run in a single huge batchsystem job (since MPI does not provide inter-job communication). While jobs of extreme size might be possible on some clusters, it is often only possible or efficient when one submits large numbers of smaller jobs, sometimes even single-core jobs which allow to fill in many small free gaps within the cluster (which sometimes have special mechanisms like backfill schedulers). Virtual Flow provides vastly more flexibility in this regard, since it can be set up to use any number and type of job. Moreover, MPI-based jobs scale sublinearly with the number of cores due to the required communication, while Virtual Flow has effectively eliminated the communication. Furthermore, the two MPI versions of AutoDock Vina are based on a single docking program, while Virtual Flow allows to use a variety of different docking programs. Also, the MPI versions lack the autonomous job management (e.g. for restarting their own successor jobs) or sophisticated techniques which would allow them to manage billions of ligands (e.g. VinaMPI stores all the output files in a single folder, which effectively limits the number of ligands they can handle dramatically[8]).

DockoMatic and DOVIS are virtual screening programs which is able to run on Linux-based computer clusters managed by a batchsystem. DockoMatic has many features such as a graphical user interface and options for homology modeling, but it is not specialized nor suitable for large scale virtual screening procedures alone because of the file and directory management. Also it has other disadvantages such that the only batchsystem which is supported is PBS.

DOVIS is probably the program most close to VFVS. While it has some features which VFVS does not posses such as a graphical user interface, it has a sigificant number of disadvantages, some of them critical. Most importantly, it is not useful for highly-parallelized workflows, as it scales linearly only up to 128 CPU cores (DOVIS 1.0) or 256 CPU cores (DOVIS 2.0). Large clusters and supercomputers however often provide more than 100 000 cores, and in order to use such facilities efficiently the scaling behavior should ideally be favorable for very large numbers of cores. Also, DOVIS supports only the old generation of AutoDock (DOVIS 1.0: AutoDock 3.05, DOVIS 2.0: AutoDock 4.0), but not the new versions such as AutoDock Vina [294] or QuickVina [112, 8] which are considerable more efficient and also more accurate than the old AutoDock versions. Indeed, the speed improvements can be up to two orders of magnitude [294, 8], and for virtual screenings on the largest scales the docking

[8] In our experience the access to folders with several hundred thousands of files can already be very slow, while several million files per folder can cause the entire cluster file system to crash. Often clusters and supercomputers also limit the numbers of files which a user can create, e.g. on the HLRN III in Berlin/Hanover the limit is 150 000 files.

programs with minimal costs are essential. Another problem is related again to the number of ligands which can be screened in a single virtual screening, which seems to be limited in DOVIS due to the file and directory structure. It also lacks features such as autonomous batchsystem management (e.g. submission of successive jobs) or ensemble docking. Another problem is that it uses the Python scripts of MGLTools[9] which have a very unfavorable I/O behavior due to the initialization of the Python environment for each single ligand, which can cause problems for the cluster file system. Virtual Flow, on the other hand, suffers from none of the above mentioned problems.

Implementation

The workflow diagram of a single queue of VFVS is shown in Figure 8.3 on the facing page. It is simplified, and leaves away details such as the possibility of new jobs to continue to work on ligands which have only partially be completed.

An example configuration for VFVS can be found in the appendix in Listing B.3 on page 256.

Results

Virtual Flow has virtually no restriction regarding the number of cores, as it scales linearly in theory without any practically relevant bounds. Speedup measurements with VFVS were done during production runs on Odyssey (Harvard's largest computer cluster). The first time up to 18 000 CPUs were used on a partition of mixed node types consisting mainly of AMD Abu Dhabi processors (32-64 cores per node) and some Intel Xeon processors. The second test was carried out using up to 10 000 cores on nodes with two Intel Xeon E5-2683V3, each having 16 physical cores. The docking program in use was QuickVina 2. The ideal scaling behavior shown in the plots of figure 8.4 on page 160 is an extrapolation based on the average processing time per ligand on a single CPU. In both cases the real speedup is nearly the same as the optimal speedup, which can be expected since there is no communication at all between the parallel processes. Minor fluctuation during the measurements of seemingly random nature where observed without significant deviations from the ideal scaling behavior. Minor random fluctuations can be expected due to various reasons, such as different ligands which were processed (which were on average of same computational expense), or fluctuations in the load of the used nodes due to other processes. The Intel Xeon cores used were on average nearly three times as fast as the Abu Dhabi cores of AMD.

Discussion

One of the primary goals of developing VFVS was scalability, which is provided already by the generic part of Virtual Flow. VFVS was in addition designed to be suitable for multistaging virtual screenings. This can be accomplished by using the large variety of docking programs and options. As an example, in the

[9] `http://mgltools.scripps.edu/`

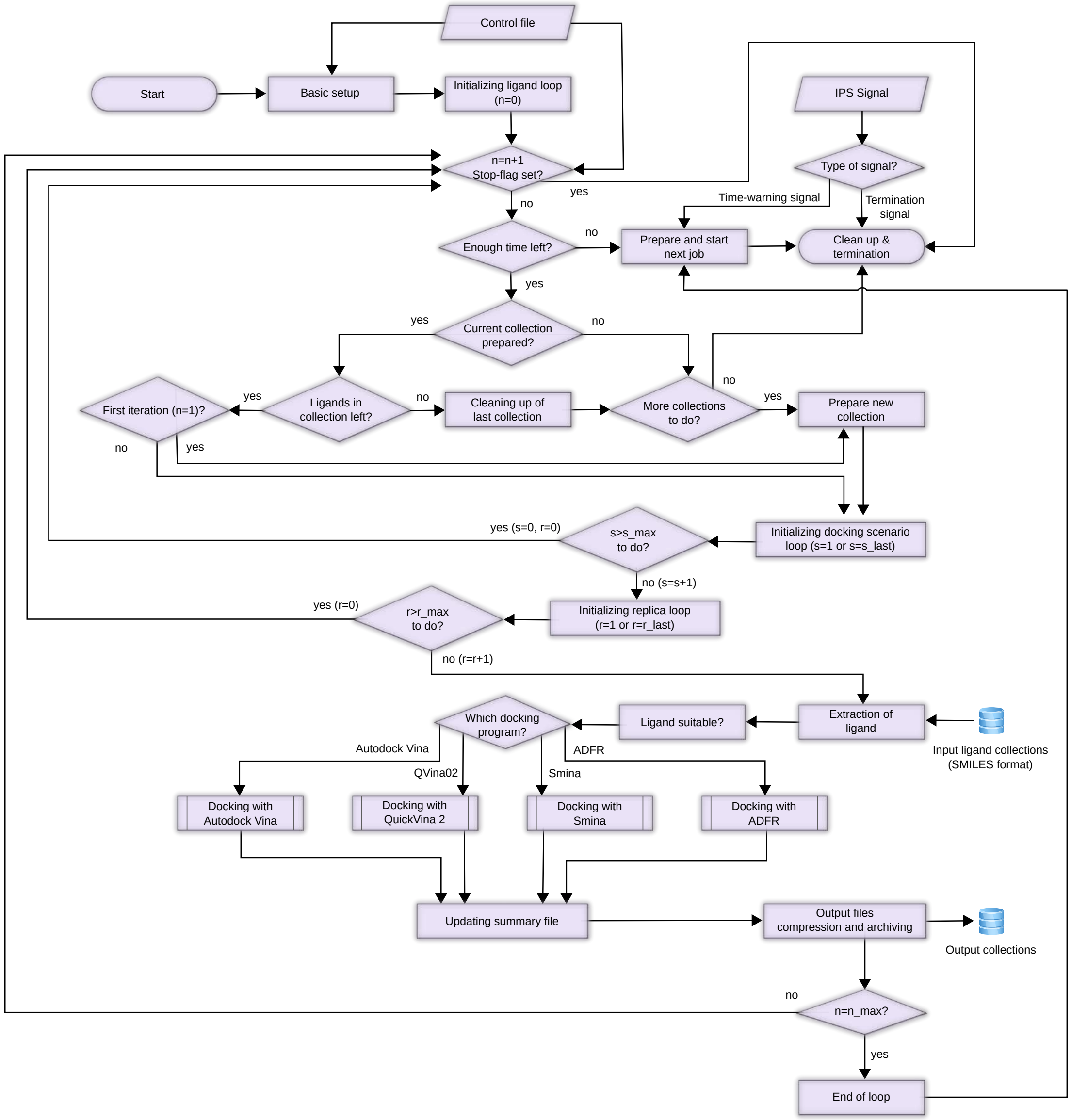


Figure 8.3: Simplified workflow diagram of a single queue in the VFVS software. The workflow until preparation of the collections is relatively generic and similar in all Virtual Flow implementations. The VFVS specific tasks start after this step (i.e. the initialization of the docking scenario loop). The single queue workflow contains three loops, the outermost one over the ligands, the central loop is over the various docking scenarios per ligand, and the innermost loop over the replicas per docking scenario.

primary virtual screening involving all the ligands rigid-receptor docking could be carried out. From these the 10 % highest scoring compounds could be used as the input ligand database for a higher accuracy rescoring procedure involving multiple docking programs, flexible residues, and several replicas. In addition further rescoring could be applied on top of that, for example using HyperQ on the third and possibly fourth levels.

Even though only up to 18 000 CPUs were used during the measurements, the linear scaling behavior is not expected to break down when using even higher number of CPUs due to the fact that there is no communication between different

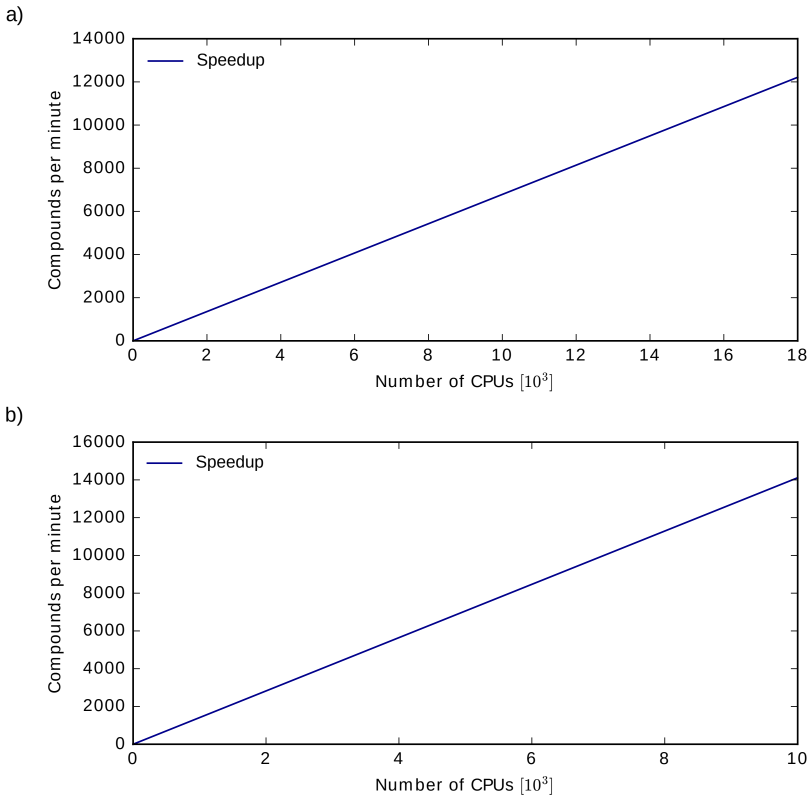


Figure 8.4: Illustration of the scaling behavior of VFVS. In plot a) nodes were used consisting mainly of AMD Abu Dhabi cores, while the second benchmark shown in plot b) was run on modern Intel Xeon processors. The real speedup is nearly indistinguishable to the shown ideal speedup without any indication of slow down even at the upper end of the range of employed CPUs, which can be expected since there is no communication at all between the parallel processes.

jobs. Possible bottlenecks which might reduce the efficiency of Virtual Flow are the number of jobs which can be submitted to the batchsystem, or the central task list which every job has to access at the beginning of each job. However, if these aspects should indeed become a bottleneck there is much room for optimization by using job layouts which minimize the number of jobs (e.g. by the length of the walltime or the number of CPUs used per batchsystem job).

VFVS is available under `https://github.com/cgorgulla/VFVS`, and the additional and optional VFTools package can be found at `https://github.com/cgorgulla/VFTools`.

Part III

Applications in Drug Discovery

Introduction

The software developed in this thesis was applied in a variety of drug discovery projects, in particular Virtual Flow. The purpose for carrying out these applications was on the one hand to have real-world test cases which allow to provide valuable feedback on how to improve the methods and software to be most useful and convenient to use. On the other hand, the work done within these applied projects might have a significant value by itself.

Application Overview

An overview of all the applications of VFVS can be seen in figure 8.5. VFLP was used for the preparation of the ligand databases (as described in 8.3 on page 153), which were used in all the applications of VFVS. HyperQ (introduced in chapter 7 on page 117) was only completed at the end of the PhD project but could still be applied within two of the real applications for initial testing purposes.

The initial plan was to work on a new drug discovery project, which is described in chapter 9 on page 165. It was intended as the only major application within this PhD project due to the time, effort, and resources required when starting a new drug discovery project in a serious way. However, other research groups became interested in our newly developed methods and software very soon, and therefore the new tools were applied by the author already within this PhD to several additional projects of collaborative groups (outlined briefly in chapter 11 on page 193). Also, one more joint drug discovery project was started by Magdalena Czuban and the author with the aim of finding candidates for a new class of

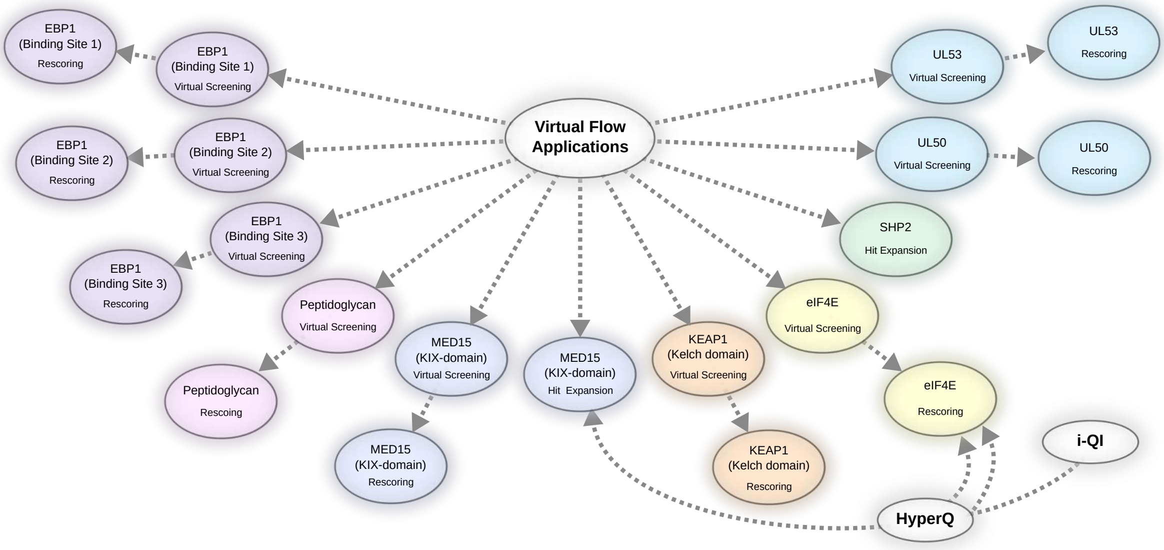


Figure 8.5: Overview of the applications of Virtual Flow and HyperQ during this PhD. Many of them are two-stage virtual screenings for hit discovery, while in two of them the purpose was to find more potent analogs of experimental hits (hit expansion).

antibiotics (summarized in chapter 10 on page 185). For testing and publication purposes an additional target was used, namely the Kelch domain of human KEAP1 (Kelch-like ECH-associated protein 1), which with 300 million screened ligands contained the largest of all virtual screenings carried out with VFVS in this PhD.

Chapter 9

EBP1

Wherever the art of medicine is loved, there also is love of humanity.

Hippocrates

Contents

9.1 Introduction

One of the most common types of disease is cancer. Approximately 40 % of the female and about 50 % of the male population in Germany is diagnosed with cancer during their life [238]. The mortality varies with the type of cancer, but on average this class of disease exhibits a survival rate of roughly 50 %, which renders it a major cause of death [238, 214]. As a consequence approximately 20 % to 25 % of all death are caused by cancer [265, 214]. Because other causes of death have decreased due to various reasons, the mortality rate due to cancer has risen over the past decades and the cancer incident rate is projected to continue to rise significantly [84]. Intense effort was made in the past decades to develop a magic bullet which is able to cure all types of cancer at once, but this endeavor has failed so far. An alternative approach to curing cancer is to develop new medications which are tailored to the specific type of cancer. All types of cancer depend on certain biomolecular pathways which they deregulate and exploit for their proliferation. Different types of cancer utilize different pathways, some are vital for their growth while others are not but enhance their progression [111]. As more and more is known about the specific biomolecular pathways which tumors exploit, this avenue has become considerably more promising over the past decades, and comprises one of the most promising approaches for healing cancer. This approach was followed in the project described in this chapter.

Choosing the Target

When looking for a suitable target to work on in this project, the following selection criteria were used.

(i) **High-Quality Structure.** The availability of a high resolution crystal structure was an important characteristic required, because it paves the way for carrying out meaningful structure-based computational drug design.

(ii) **Promising Target.** Inhibition of the target by small organic molecules should hold the promise of having a significant impact on the cure of the specific disease in question.

(iii) **New Drug Target.** The target should be one of which no other research group is or was reported on the development of drug candidates.

The last criteria rendered the finding of a target meeting all desired features a significant challenge. However, the aspect of working on a completely new target addresses an important issue. There exist a number of drug targets which are very promising and relatively secure in the sense that it is known with a high certainty that new ligands for these targets will have the desired effects, and that the target is not too challenging to drug. Therefore often numerous research groups, both academical and from industry, are working on these save havens. The more risky or difficult a potential target is, the less likely it is that drug design projects are already underway. A central question is now what has to our estimation the most benefits for pharmaceutical research and thus humanity? If a multitude of research groups is working on a small number of save targets, or if these groups would work all on different targets. Or alternatively, if a new drug design project is started, does it hold more promise to start working on a save target or start working on a completely new target even if moderately less promising? Questions like these are also the subject of economics, and in this discipline closely related formalized concepts are the *marginal utility*, *ordinal utility* or the *cardinal utility*. The law of *diminishing marginal utility* states that the more often an item is consumed or obtained the less utility or benefits the customer yields from it, in other words the marginal utility function MU is strictly monotonically decreasing:

$$\mathrm{MU}(1) > \mathrm{MU}(2) > \mathrm{MU}(3) > ... > \mathrm{MU}(n), \in \mathbb{N}. \tag{9.1}$$

These concepts can be generalized and applied to many situations, including the making of decisions [79] and health care [5]. Coming back to the question about the drug targets, the marginal utility of working on a promising novel target seems to be considerably higher rather than working on one of the save targets since

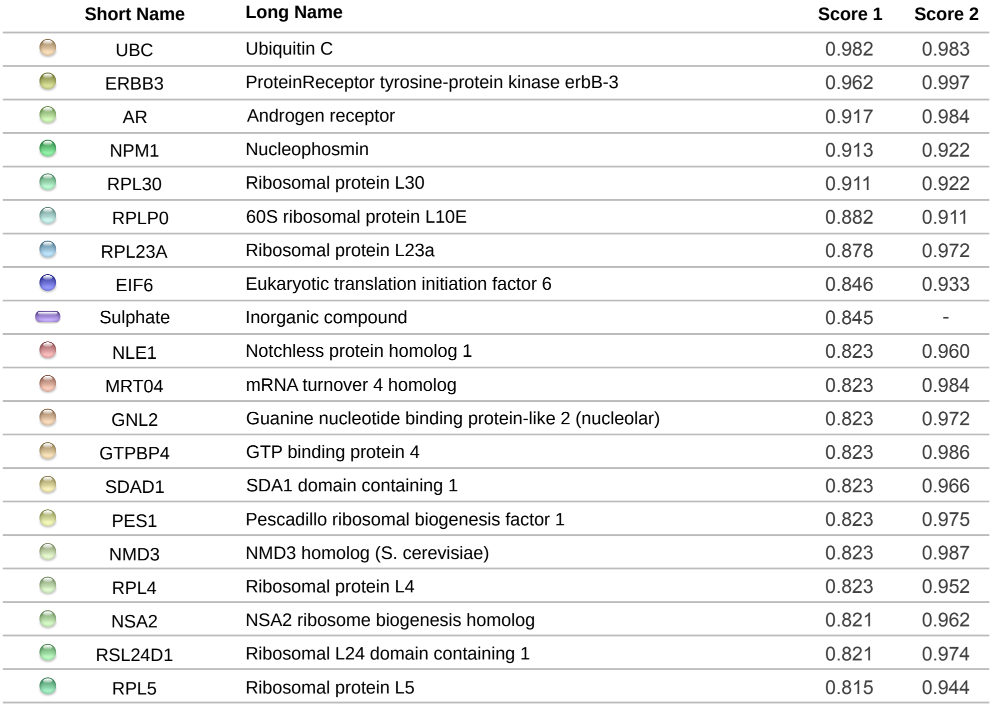
- more new knowledge is likely generated and duplicated knowledge avoided,
- it is hard to catch up with groups who are already working on a certain target,
- single new drugs for different targets are likely more beneficial than multiple new drugs for the same target.

For these reasons we have decided to work on a completely new target, even though the risk of failure is higher.

The target which was chosen after an extensive search is *ErbB3-binding protein 1* (EBP1), also known as *proliferation-associated protein 2G4* (PA2G4). It satisfies condition (i) from the list of criteria above as two high-resolution crystal structures are available, 2V6C (2.5 Å) and 2Q8K (1.6 Å). Also criterion (iii) seemed to be fulfilled to the best of our knowledge, no indications of ongoing drug discovery efforts were reported. And finally, also condition (ii) was met since EBP1 is a verified oncoprotein and finding inhibitors for certain binding regions on the protein is predicted to have anti-tumorigenetic effects regarding a variety of cancers.

Binding Partners

EBP1 binds to a large number of other biological macromolecules, including dozens of proteins and ribonucleoproteins, as well as RNA and DNA. The precise number and identities of all binding partners are not yet fully known (as for most proteins), but the likely interactions with the highest confidences are listed in Table 9.1.



	Short Name	Long Name	Score 1	Score 2
●	UBC	Ubiquitin C	0.982	0.983
●	ERBB3	ProteinReceptor tyrosine-protein kinase erbB-3	0.962	0.997
●	AR	Androgen receptor	0.917	0.984
●	NPM1	Nucleophosmin	0.913	0.922
●	RPL30	Ribosomal protein L30	0.911	0.922
●	RPLP0	60S ribosomal protein L10E	0.882	0.911
●	RPL23A	Ribosomal protein L23a	0.878	0.972
●	EIF6	Eukaryotic translation initiation factor 6	0.846	0.933
▬	Sulphate	Inorganic compound	0.845	-
●	NLE1	Notchless protein homolog 1	0.823	0.960
●	MRT04	mRNA turnover 4 homolog	0.823	0.984
●	GNL2	Guanine nucleotide binding protein-like 2 (nucleolar)	0.823	0.972
●	GTPBP4	GTP binding protein 4	0.823	0.986
●	SDAD1	SDA1 domain containing 1	0.823	0.966
●	PES1	Pescadillo ribosomal biogenesis factor 1	0.823	0.975
●	NMD3	NMD3 homolog (S. cerevisiae)	0.823	0.987
●	RPL4	Ribosomal protein L4	0.823	0.952
●	NSA2	NSA2 ribosome biogenesis homolog	0.821	0.962
●	RSL24D1	Ribosomal L24 domain containing 1	0.821	0.974
●	RPL5	Ribosomal protein L5	0.815	0.944

Table 9.1: Binding partners of EBP1 with the highest experimental confidence levels represented by score 1. Score 2 takes not only experimental data into account, but also textmining, databases, co-expression, neighborhood, gene-fusion, co-occurence and computational predictions. Data provided by the STITCH database [160, 161, 162, 163, 277].

Many of the binding partners interact also with themselves,partly due to the circumstance that EBP1 is part of multiprotein complexes such as the pre-ribosomal ribonucleoprotein complex. The physical interactions of EBP1 which are known or predicted with the highest confidence levels according to the STITCH database are shown in Figure 9.1 on the next page. The various interactions give rise to a relatively large number of functions of EBP1, which besides its role in

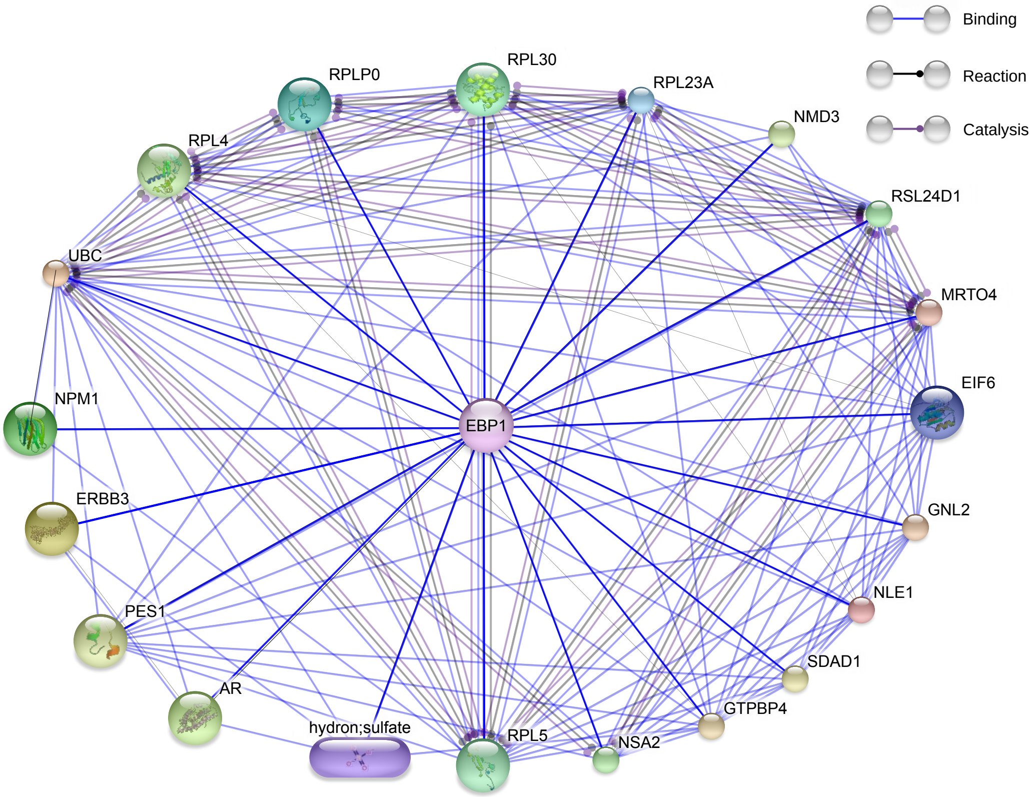


Figure 9.1: Network graph of EBP1 and its binding partners with the highest experimental confidence levels. In contrast to large nodes small nodes indicate that no 3D structure is known. Pale edges are connecting structures not involving EBP1 directly. Graphic adapted from the network graph provided by the STITCH database.

the assembly of ribosomes are large related to the regulation of cell growth and differentiation [263, 184].

Role in Cancer

Target Region 1 and 2.

Cancer cells typically deregulate certain signaling pathways for their proliferation. The precise pathways which are exploited differ between cancer types. One protein which is part of such oncogenic signaling pathways is podoplanin (PDPN), which is highly overexpressed in cells of certain types of cancer, such as squamous cell carcinoma (SCC) of the mouth, the lung, the cervix, the skin, the esophagus, in certain tumours of the central nervous system, in dysgerminomas, or in mesothelioma. Podoplanin also plays a role in other cancers such as breast and pancreatic tumors. [273, 325, 299] It is relevant in both tumorigensis and proliferation, and increases the risk of the formation of metastases. Its role in tumor invasion is also connected due to its amplifying effects on the migration of human fibroblasts [273].

It follows that if podoplanin is downregulated in some way in cells of certain types of cancer it might lead to tumor-suppressing effects. Targeting podoplanin directly by structure-based drug design is not directly possible because no sufficiently complete crystal of the protein is yet available. In addition, podoplanin is a vital protein for higher organisms such as mice, as knock-out experiments with PDPN-deficient mice have shown [299] which die immediately after birth. Therefore it might be more favorable to interrupt the signaling pathway which triggers the overexpression of podoplanin. A key protein which regulates podoplanin is EBP1, and indeed it seems to be responsible for the overexpression of podoplanin within the cancer cells since EBP1 is upregulated by itself as shown in [196]. This study also showed that downregulation of EBP1 in cancer cells via silencing RNA (siRNA) indeed reduces the malignant characteristics of EBP1. EBP1 is known to be a transcriptional activator of podoplanin, and since a full high-resolution crystal structure for it is available it seems to be particularly suitable for structure-based drug design.

Another important question is if EBP1 is vital to living organisms. If yes, this circumstance might render it unsuitable as a drug target due the resulting toxicity of potential drugs candidates. However, it was shown in [353] on EBP1-deficient mice that EBP1 is not vital, but when missing during the growth there are some alterations in the mice (such as reduced size). This seems to indicate that when EBP1 is inhibited by potential drugs during a certain period of time it might only have minimal effects on healthy tissues, while at the same time such inhibitors could prevent cancer cells from exploiting the EBP1 signaling pathway.

The question is now which site at EBP1 to target in drug discovery projects. The precise mechanism of how EBP1 is carrying out the transcriptional upregulation of podoplanin remain elusive. What makes the situation more complicated is that there are two isoforms of EBP1, p42 and p48. The latter is the full version of EBP1, while the former is missing the N-terminal 54 amino acids. p42 is located in the cytoplasm, while p48 has its major functions within in the nucleus.

One avenue worth pursuing might be to prevent EBP1.p48 to localize within the nucleus. It was shown in [263] that the first 48 amino acids of the N-terminal of EBP1 are required for its localization in the nucleus. It was also demonstrated that only the two amino acids K20 and K22 are essential for its nuclear localization, and when mutated to alanine caused EBP1 to be localized in the cytoplasm. Therefore it was suggested to be part of a nuclear/nucleoloar localization signals, both types of which are often occuring together and overlapping, which seems to be the case also here. Furthermore the residues K20 and K22 are highly conserved among homologs from different species, indicating that these two residues play indeed a critical role. Later when the crystal structure was solved it was revealed that K20 and K22 are closely located near the lysine rich loop ^{63}IFKKEKEMK^{71}, and lysine/arginine rich sequences are typical for nuclear/nucleolar localization signals. [203, 263, 154].

Another avenue which might be able to yield the desired cessation of the upregulation of podoplanin by EBP1 is to prevent EBP1 to bind to the podoplanin promoter. However, the precise binding mechanism in this regard is not yet fully understood.

The two regions surrounding the residues K20 and K22 seem to be among

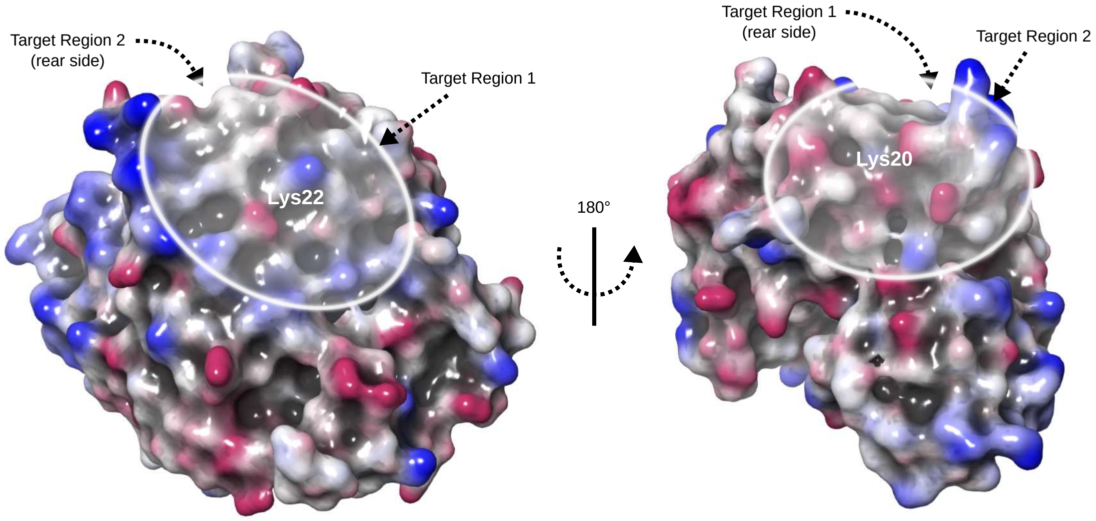


Figure 9.2: Target regions 1 (shown on the left) is located around the proximity of the amino acid K22. Target region 2 (on the right side of the figure) is on the other side over the crest of the protein centered around residue K20.

the most promising sites to target. On the one hand because it might prevent EBP.p48 to be transported into the nucleus. On the other hand the regions around these two residues might be critical for transcriptional activation of the podoplanin expression, and small molecular inhibitors might be able to prevent the binding of EBP to the podoplanin promoter or other components of a multiprotein complex which interact with the DNA. The sidechains of the two residues, even though nearly neighbors in the sequence, are spatially separated on the surface of EBP1.p48. Therefore they can be seen as two individual target regions, and in this thesis they are referred to as *target region 1* (K20) and *target region 2* (K22), both shown in Figure 9.2.

Target Region 3.

Interestingly, EBP1 also strongly promotes other types of cancer, such as glioblastoma, via a mechanism which involves its bindings to nuclear Akt (also known as Protein Kinase B (PKB), as well as the cyclin-dependent kinase 2 (CDK2). It was shown in [149] that the phosphorylation of the residue S34 of EBP1.p48 by CDK2 is vital for the acceleration of tumor cell growth. Therefore a small molecule inhibitor at the binding interface between CDK2 and EBP1 might be able to prevent the ability of the former to phosphorylate the latter, and in this case might lead to anti-tumorigenic effects as well as prevent its proliferation. The region around amino acid S34 is referred to as *target region 3* and shown in Figure 9.3 on the next page.

Besides being the phosphorylation site of CDK2 it is possible that this region also plays a role in the podoplanin-based tumor promoting functions of EBP1.

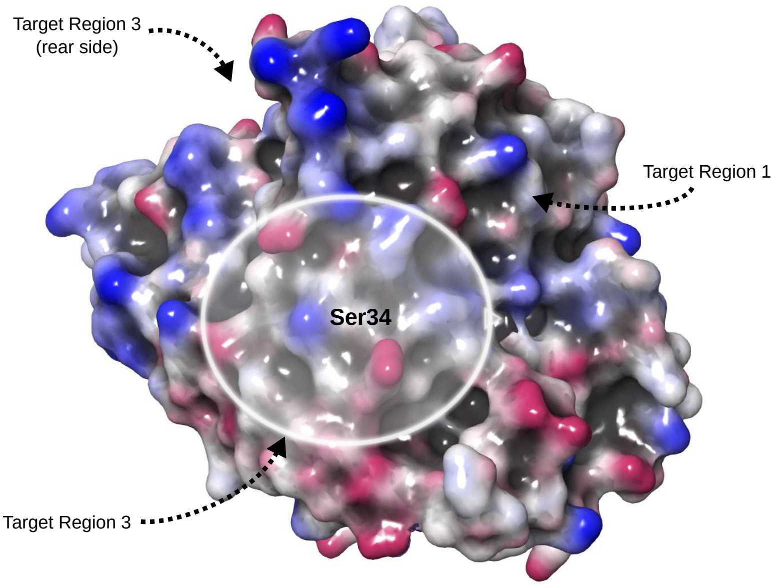


Figure 9.3: Target regions 3 is located in the vicinity of amino acid S34. Mutation studies have shown that phosphorylation of this residue is vital for oncogenic functions of EBP1.p48 within certain types of cancer. Inhibiting this site with a small molecule is expected to render EBP1 unable to perform these cancer supporting activities.

Role in Hepatitis C

Very recently it became known that EBP1 also plays several roles in relation to the Hepatitis C virus (HCV) [201].[1] Most interestingly, the situation is similar as within cancer in the sense that EBP1.p42 strongly suppresses HCV replication, while EBP1.p48 promotes the proliferation of the virus. The pathogenic activities of EBP1.p48 happen again most likely within the nucleus as is the case for cancer cells. This indicates that targeting the mechanism which is responsible for transporting EBP1 to the nucleus might be able to suppress the HCV-enhancing functionality of EBP1, and this mechanism is already aimed at by the target sites 1 and 2. Alternatively, the interaction of EBP1 with certain binding partners which are involved in enhancing the HCV replication can be targeted. The precise mechanism is again not yet known, but it is possible, if not likely, that the N-terminal region which is targeted by target regions 1-3 is involved in such interactions. Therefore it is conceivable that successful inhibitors of these regions could be dual-inhibitors for both HCV as well as certain types of cancer.

Protein-Protein Interactions

Target regions 1-3 on EBP1 are very challenging to inhibit under several aspects, which is probably one of the reasons why no drug design projects on this target

[1] This information, which makes EBP1 an even more interesting target, was only published two years after this PhD project was started, and therefore it was not taken into account when when EBP1 was chosen as the target.

were yet reported.[2] Among them are the following:

- The target regions are relatively flat surfaces taking part in PPIs, which are generally very difficult to drug in comparison to deep pockets as often found in enzymes. For instance due to the reduced interaction surface area between potential binders and the receptor.
- While the target regions 1-3 are in the proximity of specific residues which are known to play key roles, the precise locations which are relevant for disrupting the desired PPIs are not yet known. This leads to relatively large target regions, and possibly also to successful binders which are not able to disrupt the desired PPIs.
- The above two aspects are of a physical nature. And they lead to a significant challenge also for computational methods. For example, a large potential binding region means that during docking procedures there is a relatively large search space to explore. Moreover, when amino acid side chains are allowed to be flexible within the target regions the number of residues which are involved increases. This can become a serious issue because the computational costs in molecular docking often rise exponentially with the degrees of freedom. Due to the relatively large target regions it also becomes nearly impossible to define useful pharmacophores which would allow to narrow down the input ligand database.

Even though PPIs are in most cases extremely challenging to inhibit by small organic compounds they are one of the holy grails of drug discovery for a number of reasons. Among them are the following.

- Only about 10 % to 15 % of all protein-encoding genes encode enzymes according to the Human Protein Atlas [228, 300].
- An enzyme has normally only a single enzymatic site, even when it can process multiple types substrates. In contrast a protein often has multiple PPIs which multiplies the number of possible target sites for the development of medications. Indeed, the humane interactome was estimated to contain between 130 000 and 650 000 PPIs [257, 60]).
- Targeting PPIs allows for more selective drugs and more specific effects, because often enzymatic sites are similar to each other, causing the same inhibitor to bind to multiple related enzymes.
- Enzymes have already been extensively in the focus of drug development projects since decades, and the most promising of them have already been drugged multiple times. PPIs on the other hand were largely ignored as drug targets, and only in the last decade first endeavors where reported. According to [257] approximately 10-15 small molecule drugs (NMEs) inhibiting PPIs have already successfully been developed (i.e. obtained at least one jurisdiction), but this is still a tiny fraction when compared to the totality of 1 600 to 1 800 approved NMEs.

[2] Another reason might be that these regions are not well known target regions. We came to the conclusion that they are promising target regions by combining information from different articles. In addition, while it becomes more and more clear that EBP1 is a promising target, many functional details and mechanisms related to the protein are still unknown.

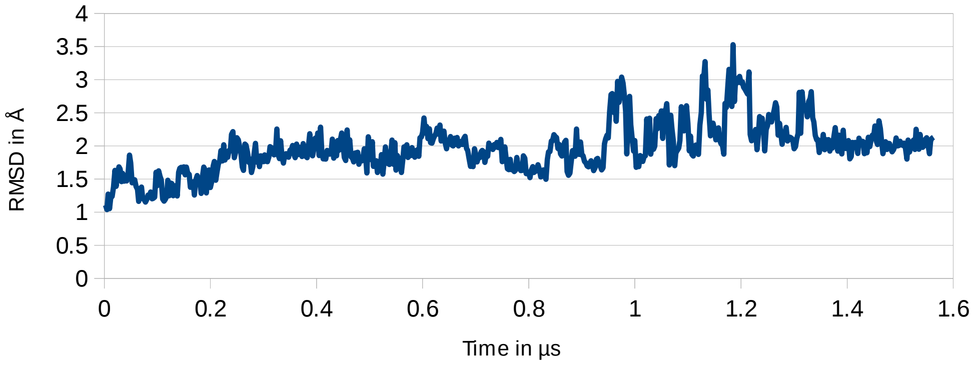


Figure 9.4: RMSD of one of the MD simulations of EBP1.

The above points indicate that there are not only fewer target sites when considering only enzymatic sites rather than the class of PPIs. It also means that the number of diseases which can be cured when targeting enzymes is expected to be considerably smaller than for PPIs since the latter most likely plays a role in significantly more diseases. PPIs play for example a central role in diseases like Alzheimer's disease and Parkinson's disease, where certain proteins in the central nervous system are misfolding and forming aggregates via PPIs, and these protein aggregates cause damage to the cells. Therefore it is conceivable that small molecules which bind to these misfolded proteins can prevent their aggregation.

Due to the challenge which PPIs pose as drug targets better tools and techniques become even more important and useful. The new method and software presented in this thesis go into this direction. Therefore challenging targets such as PPIs seemed to be the ultimate test scenarios for them, which was another reason why we decided to work on EBP1.

9.2 Structure Preparation

The starting conformation was the PDB crystal structure 2Q8K. Missing residues were modeled by Modeler. Geometry optimization was carried out (1 000 steps), followed by extensive conformational sampling, both with NAMD 2.10 on CPUs as well as on graphics-accelerated nodes running NVIDIA Tesla GPUs. The sampling was done in order to obtain biologically relevant conformations, but also to understand the dynamic behavior and the flexibility of the receptor. Three independent NPT-simulations were carried out, each with a total length of more 1 µs. The root-mean-square deviation (RMSD) of one of them is shown in Figure 9.4.

The force field used was CHARMM36, the solvent was modeled explicitly via the TIP3P water model and was ionized with NaCl (0.15 M). The total system contained 59207 atoms. Periodic boundary conditions were used for the rectangular shape of the system, the temperature during the production runs was 310 K, and the time step was set to 2 fs. For the virtual screening procedure a single receptor conformation was obtained from the MD trajectories via clustering of the conformations with the program GROMACS.

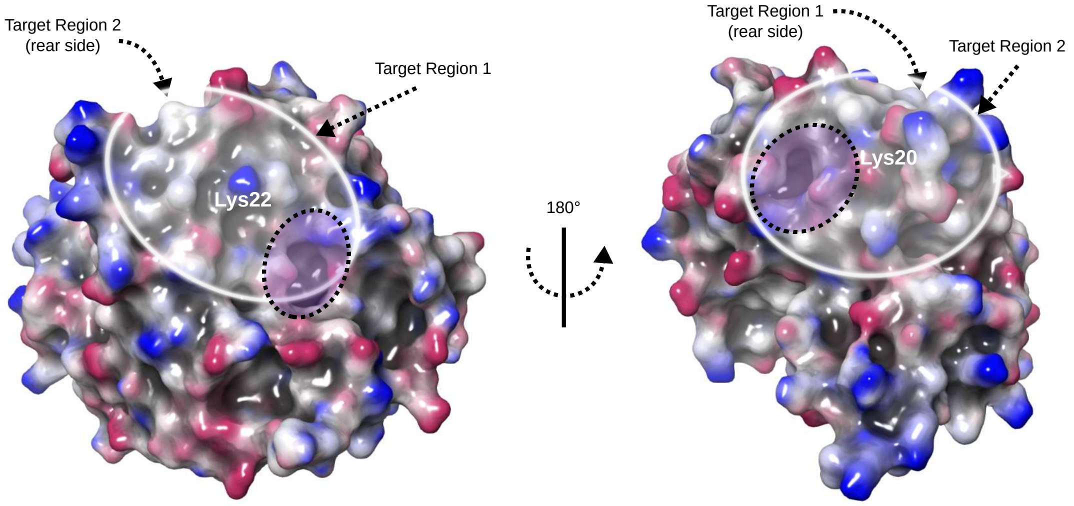


Figure 9.5: Left: Target region 1 of EBP1 after extensive conformational sampling and clustering. The tiny pocket which was visible in the crystal structure (2Q8K) has widened (marked with the smaller circle in the figure), hypothetically providing the possibility for smaller functional groups of potential drug candidates to bind into it and with that enhance their overall binding affinity to the protein, which can be essential for inhibiting difficult PPIs as this one. Right: A new relatively deep pocket has formed at the left of the target residue K20.

The conformation obtained by the clustering exhibits several differences to the original crystal structure, two of which might be of critical importance when developing new drugs. The first difference is present at the boundary of target region 1. In the crystal structure a tiny pocket was visible with an even smaller entrance (as can be seen in Figure 9.2 on page 170), apparently not large enough for small molecules to enter it. However, in the cluster-structure the pocket and also the entrance to it has considerably widened. Even more, it has extended in space into the direction of residue K22 and is therefore located to a large extent in target region 1 (see Figure 9.5 on the left). This pocket might increase the potential binding affinity of small ligands to the increased binding surface. Excitingly, a similar phenomenon has occurred in relation to target region 2. Also here a cavity has formed, but in contrast to target region 1 originates very close to the key residue of this target region (K20), and from there extends towards the boundary of the target region (shown in Figure 9.5 on the right).

Regarding target region 3, a formation of a new semi-cavity could not be observed. However, protein-protein docking between EBP1 and the CDK2/Cyclin A complex with GRAMM-X [292] was carried out to see if the likely binding interaction between these two proteins could be predicted. This turned out to be possible with high certainty alone due to the circumstance that the predicted binding pose shows an interaction mode between the two proteins which matches exceptionally well regarding the geometry (illustrated in Figure 9.6 on the facing page). As can be seen be seen in the figure, the shape of CDK2/Cyclin A matches the shape of EBP1 nearly perfectly in the vicinity of target region 3. The residue S34 which is phosphorylated by CDK2 and located at the center of target region 3 is not directly touched by CDK2/Cylcin A complex (visible in the top view of the figure), which is as expected. Such structural information is often extremely

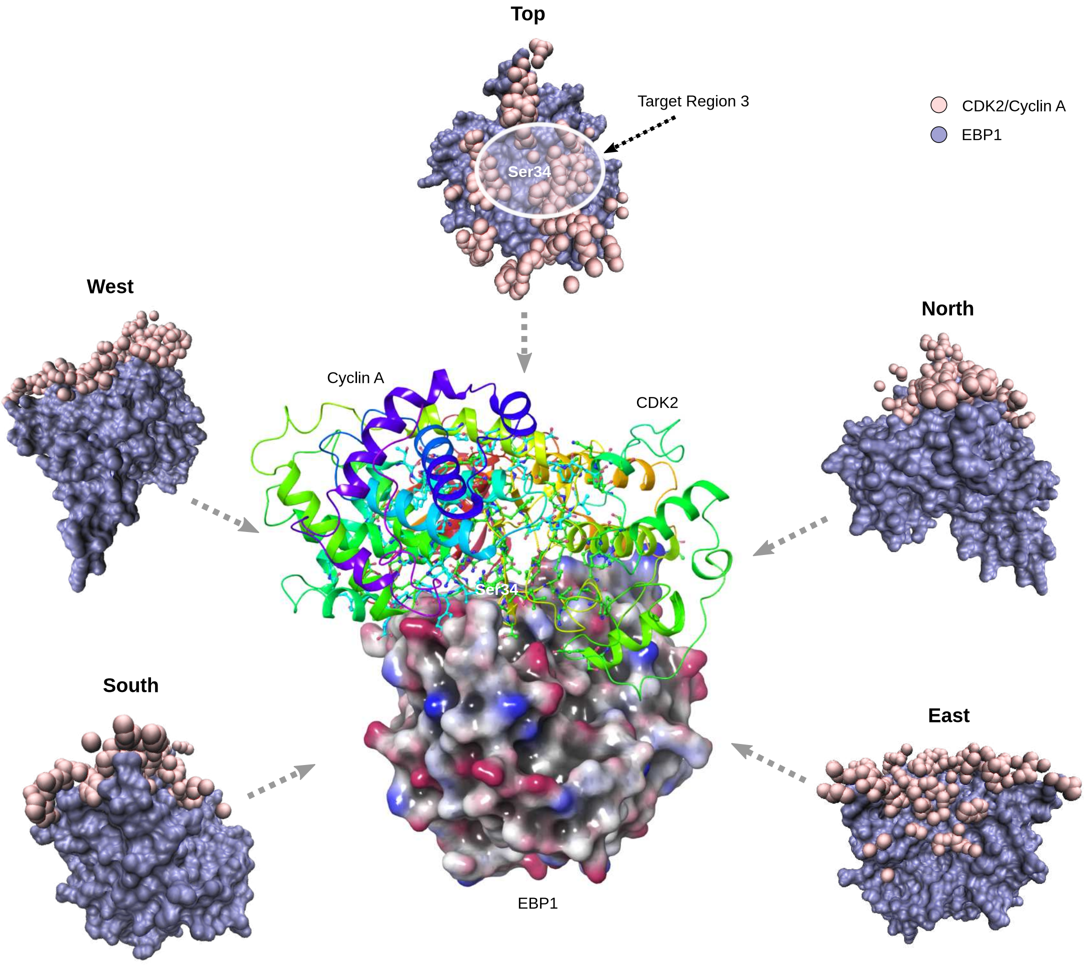


Figure 9.6: Protein-protein docking between EBP1 and the CDK2/Cyclin A complex. For EBP1 the same structure was uses as for the virtual screenings (i.e. the one obtained from the MD simulations/clustering). Center: EBP1 (solid surface) and the CDK2/Cyclin A complex (rainbow colored ribbons) binding to target region 3. Surrounding images: EBP1 (grey violet) and the atoms of the CDK2/Cyclin A complex (peach) which have a distance to EBP1 of less than 5 Å from different viewpoint. The grey arrows indicate the directions of the views onto EBP1.

valuable for structure-based drug design, since one knows exactly which residues take part in the PPI. This information can for example be used in docking-based virtual screenings to narrow down the search space.

9.3 Hit Identification

Virtual Screenings and Rescorings

Primary Virtual Screenings. To identify new potential hit compounds for each of the three target regions a two-stage virtual screening procedure was carried out with VFVS (described in section 8.4 on page 155). In the first stage (primary virtual screening) a primary virtual screening was carried out in which

130 million compounds of the ZINC 15 database were screened. A major advantage of ZINC is that contains most of the commercially available compounds which makes experimental verifications often considerably cheaper and faster (as the compounds can be ordered). The database was prepared before with VFLP (see section 8.3 on page 153), where the compounds were protonated at physiological pH and converted into the PDBQT format which is used by AutoDock-based docking programs. During the primary virtual screening QuickVina 2 was used with an exhaustiveness value of 8 while the receptor was held rigid. The receptor structure used was the one obtained from the procedure described in the previous section 9.2 on page 173.

Rescoring Procedures. After the primary virtual screening the compounds were ranked by their binding affinity, and approximately the top 100 000 highest scoring compounds were used for the rescoring procedures for each of the three target regions. Each ligand was docked three times in each rescoring procedure, carried out again with VFVS. One time with Smina Vinardo with the receptor held rigid, and two times (one time with AutoDock Vina and one time with Smina Vinardo) with the amino acid sidechains of EBP1 allowed to be flexible at the targeted regions (14 amino acids of target site 1, and 15 amino acids of target site 2 and 3). [3] In our experience this is on average the maximum number of residues which can be allowed to be flexible such that AutoDock Vina-based docking programs still perform well. With more residues the results become useless as the search algorithm seems to break down. Since more than 15 residues are part of the targeted regions, a subset needed to be selected. The more flexible a residue is within the structure, the more important it also becomes during molecular dockings that the residue is flexible. As a measure for the flexibility of the residues the B-factors of the crystal structures were used, as well as the dynamical data obtained by the MD simulation during the structure preparation. The results of the rescoring procedure of target region 1 in terms of their scores is illustrated in Figure 9.7 on the facing page. Binding affinities below -12 kcal/mol are already relatively high, even for inhibitors for deep enzyme pockets, as at around -12.3 kcal/mol corresponds to dissociation constants of 1 nM at standard conditions. It is therefore remarkable that it was possible to identify several hundred compounds in this region despite the fact that target region 1 is a relatively flat PPI. However, it can be expected that many of the top scoring compounds have overestimated binding affinities due to the approximate nature of fast docking programs. Since the threshold for classifying compounds as hits can be set considerable lower, for example to 10 kcal/mol, implying that up to several thousands of compounds can be seen as (computational) hits.

The virtual screenings and rescorings of the other two target sites (2 and 3) were similarly successful, having also led to thousands of computational hit compounds with predicted binding affinities.

[3] Vina was not used in a rigid mode because QuickVina 2 which was used in a rigid mode already during the primary virtual screening uses the same scoring function as Vina.

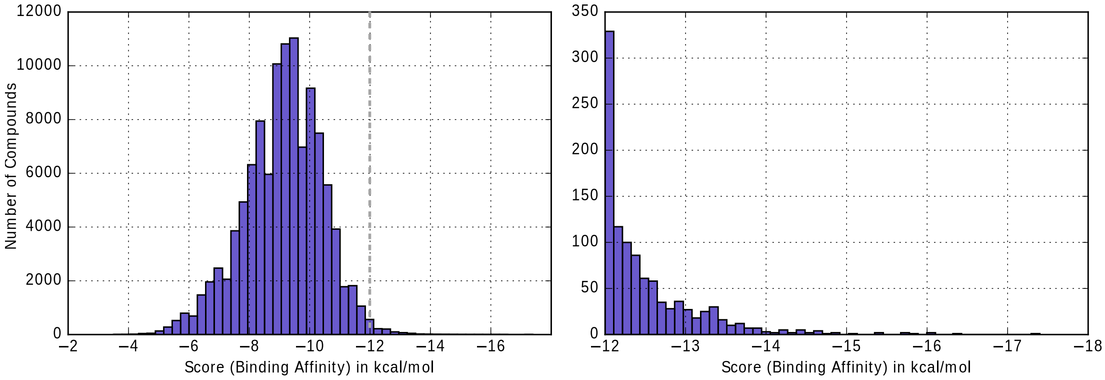


Figure 9.7: Histograms of the predicted binding affinities by Smina Vinardo of the compounds which went from the primary virtual screening into the rescoring procedure. The left side shows the entire histogram while the right side shows only the part of the diagram where the predicted binding affinities are stronger than -12 kcal/mol (the lower/more negative the score/predicted free energy of binding, the stronger the predicted interaction.)

Postprocessing and Filtering

Even though Vina and Smina Vinardo are both predicting the binding affinity, the scores of the latter were on average around 15 % higher than the scores of Vina for the targeted regions when the receptor was allowed to be flexible. Also rigid dockings generally result in lower final scores than flexible dockings since there is less room for optimization. Therefore, in order to be able to directly compare the scores of the different docking programs/scenarios, the docking scores were scaled relative to the docking scenario involving Smina Vindardo and flexible residues. This was done by multiplying the Vina (flexible) score of each ligand by 1.15, the Smina (rigid receptor) score 1.25, and the QuickVina 2 score by 1.4 in order to match the scores of the 100th hit compound in each of the hit lists.

Of all the rescored compounds a variety of physico-chemical properties was computed with DataWarrior [245] in order to have additional selection criteria for the compounds which should be selected as the most promising hit candidates. Among the criteria which were taken into account were the following:

- Score of QuickVina 2 (rigid)
- Score of AutoDock Vina (flexible)
- Score of Smina Vinardo (rigid)
- Score of Smina Vinardo (flexible)
- clogP value (DataWarrior)
- logP value (ZINC database)
- Molecular weight (DataWarrior)
- Ligand efficiency (score/number of heavy atoms)
- Topological polar surface area (Data Warrior)
- Druglikeness-measure (DataWarrior)
- Tumogenic risk assessment (DataWarrior)
- Mutagenic risk assessment (DataWarrior)

- Reproductive effect risk assessment (DataWarrior)
- Number of rotational bonds (DataWarrior)
- Number hydrogen donor/acceptors for hydrogen bonding (DataWarrior)
- Visual inspection of the docking poses of the highest scoring compounds

An example hit compound, named A1, for which the various criteria were applied is shown in Figure 9.8. The compound has very favorable pharmacological and physico-chemical properties. It satisfies all of Lipinski's rule of 5, for instance it has a predicted logP value of between 2.5 and 3, which renders it unlikely to be an aggregator and at the same time more easy to handle in experiments than very hydrophobic compounds. With a topological polar surface area (TPSA) of only approximately 55 Å^2 compound A1 should not only be easily able to penetrate cell walls, but might also be able to penetrate the blood brain barrier (which could be important when treating certain types of brain cancer which exploit EBP1 for their proliferation). Also no toxicity warnings (mutagenic effects, reproductive risks, tumorigenic effects) were present for this compound. The predicted binding pose of this compound (shown in Figure 9.9 on the facing page) makes use of the new pocket which has widened during the MD simulations. As can be seen in the figure, one end of the compound suits perfectly into this cavity, while the other end binds directly in the vicinity of residue K22 which is the central amino acid of target region 1. Compound A1 is predicted to have a variety of specific interactions with EBP1, including three to four hydrogen bonds and even a halogen bond (besides multiple hydrophobic interactions). Among the other hit compounds is great chemical variety, and the same is true for the site at which they are binding. While a certain fraction binds in a similar mode to EBP1 as the above example compound, other compounds bind on other locations of target region 1.

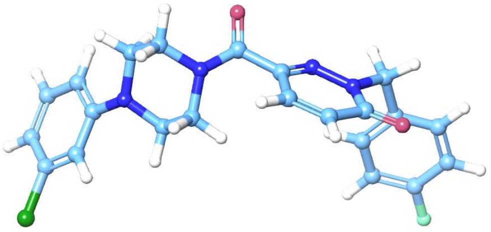


Figure 9.8: Example hit compound A1 for target region 1 of EBP1. The heterocyclic molecule contains two oxygen atoms, four nitrogens, as well as one chlorine atom (green) and one fluorine atom (cyan).

9.4 Experimental Verifications

The next step within this drug development procedure would at this point normally consist of experimental binding assays to determine which of the predicted compounds are indeed real binders. This information would allow one to concentrate on the most promising confirmed hits and their subsequent optimization.

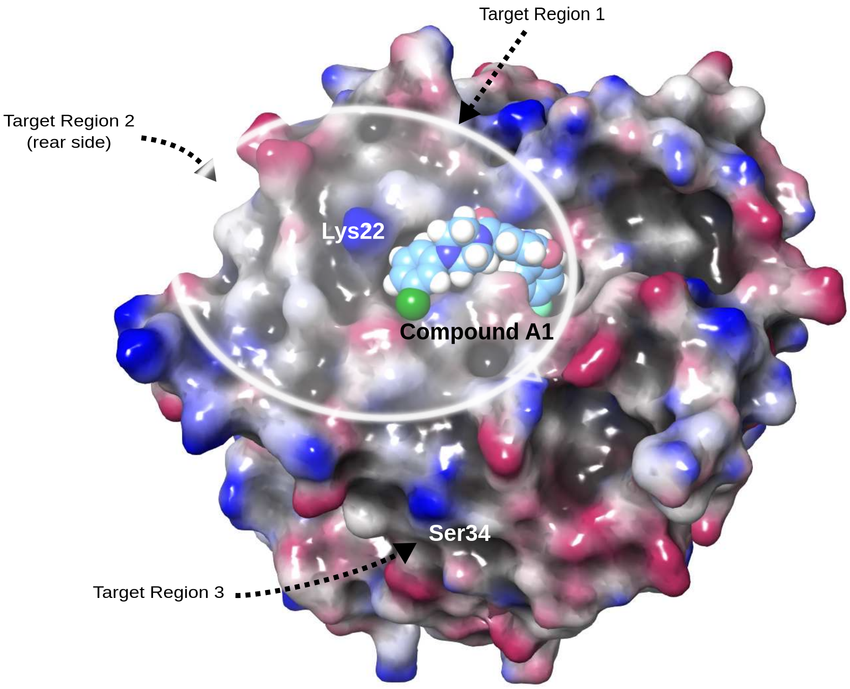


Figure 9.9: Docking pose of an example hit compound for target region 1 of EBP1. One end (containing a fluorine atom) of the compound binds into the pocket which was was formed during the MD simulations and is also present in receptor structure which was used for the dockings. The other end containing the chlorine atom bind directly besides the target residue K22.

Because this PhD project is rooted within theory and computation, experimental verifications were not planned to be included initially but rather intended to be left for continuation projects after the PhD. However, bringing a computational project to the experimental level can take a considerable amount of time, alone the step of receiving sufficient funding can take years. Therefore it seemed favorable to initiate the work on the experimental level already during the PhD (as far as possible without additional funding), in particular since first experimental results can be very helpful to receive funding for continuation projects. For this purpose the author has set up collaborations with three research groups in Boston, MA, USA. Two[4] of them are specialized on structure-based (experimental) drug discovery as well as PPIs, and moreover are working on new cancer medications by their own. Therefore these two groups seemed to be ideal for bringing our project on the experimental level. On the other hand, our new methods and software seemed to be ideal for application in several of their projects in a complementary manner. Due to the mutually beneficial effects which were to be expected the author was invited to visit the groups and worked in them for approximately half a year.

In order to test the hit compounds experimentally, the candidate compounds need to be obtained. While all of the compounds are commercially available, only approximately 5 % are in-stock compounds. The vast majority of compounds are on-demand compounds (largely part of virtual libraries of compounds which were

[4] The first is Wagner Lab at the Harvard Medical School (HMS) of Harvard University. The second group of Haribabu Arthanari is located at the Dana Farber Cancer Institute.

never synthesized before). Therefore these compounds need to be synthesized at first by medicinal chemists of the suppliers, which often is very expensive (a single true on-demand compound can cost thousands of dollars).

One of the additional advantages of carrying out virtual screenings on a very large scale is the increased number of computational hit compounds which one normally obtains, in our case several hundred to several thousand (depending on where the threshold is set). This favorable circumstance provides for example the possibility to apply additional criteria when selecting candidate compounds such as the availability by certain vendors or a specific purchasability type (such as *in-stock*).

In our case of EBP1 we have therefore clustered the hit compounds by vendors (yielding non-disjoint sets, one for each vendor). Only target region 1 was taken into account for the first experimental verifications. The by far largest cluster for this target region is is from Enamine[5], which contains approximately 25 % of all hits, and of these around 90 % are on-demand compounds. Because no funds for experiments were included in this PhD project, the author tried to establish a collaboration Enamine, which was successful. Enamine has kindly agreed to support our cancer drug design project by providing us for free in-stock and on-demand compounds of their chemical libraries (which contains approximately 30 million compounds). In the first supply round Enamine has provided us with twelve on-demand compounds, referred to via their short names Z1 to Z12.

Thermal Shift Assays. In order to verify experimentally which compounds are binding, thermal shift assays based on Thermofluor were carried out [252, 61, 185]. For this purpose the compounds needed to be prepared as well as the target protein.[6]

To be able to produce the target protein (EBP1) the gene of EBP1 (PA2G4) was obtained from the nonprofit plasmid repository Addgene[7] in a mammalian expression vector tagged with HA (pCMV-HA-Ebp1, plasmid #67792), as described in [314]. Cell cultures were grown to amplify the plasmid in two stages, after which it was purified via minipreps. Then the DNA sequence of the relevant region of the plasmid was sequenced. The backward sequencing did not work at first when using the BGH reverse primer (which is specified by Addgene), but the M13 reverse primer did work in a second sequencing trial. The sequencing confirmed that the plasmid contains the desired gene without errors. Afterwards the vector was purified and amplified by PCR, and then cloned into DH5 alpha cells (by restriction free cloning techniques). The new plasmid was isolated, sequenced, and transformed into expression cells (BL21 C43 DE3) as well as cloning cells for future purposes. Using the expression cells the protein was expressed and purified via affinity and size exclusion chromatographies.

Due to the challenging nature of PPIs such as in this case hit rates as low as

[5] Enamine Ltd, `http://www.enamine.net`

[6] All the work done in relation to the protein preparation and the thermal shift assays was carried out by Andras Boeszoermenyi and the author in the Harvard Medical School, Boston, MA, USA.

[7] `https://www.addgene.org/`

1 % can already represent a significant success with conventional virtual screening approaches (which corresponds on average to 1 in 100 tested compounds), since one is in such cases often fortunate when being able to identify only a single binder.With hit rates of this order it would be very unlikely that among the twelve Enamine compounds would be even one experimental binder. However, the hope was that the extreme scale of the virtual screening with more than 100 million compounds as well as the multistage approach (both described in chapter 5 on page 91) would lead to significantly higher hit rates. And according to the thermal shift assays this was indeed the case. Among the twelve Enamine compounds at least 4 (possibly 5) showed clear signs of binding to EBP1 in the binding assays. This number is remarkable and corresponds to a hit rate of above 30 %. And it is an indication that the theoretically predicted positive effect of the scale of the screening has indeed a considerable effect on the hit rate as hoped.

Fluorescence Microscopy. Another important question is whether the experimentally confirmed hit compounds also have the desired effects within living cancer cells. For this purpose human U2OS cancer lines were prepared and treated with the Enamine compounds.[8] As pointed out earlier in this chapter, the precise location of where small molecules need to bind in target regions 1 and 2 is not known, and therefore it could be that the binding to EBP1 occurs at

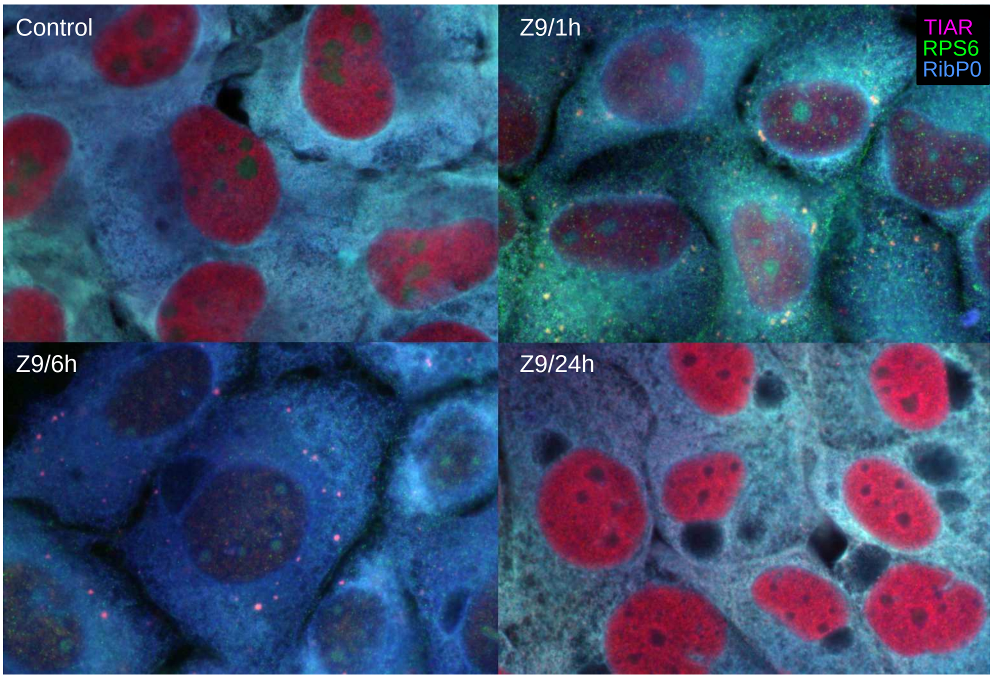


Figure 9.10: Fluoresence microscopy images of human U2OS cancer cells which were treated with compound Z9. After only one hour significant changes within the cells are already clearly visible. After 24 hours large, dark vacuoles have formed, indicating these cells are quite sick, possibly even dying.

[8] All the work done regarding the cell lines and the fluorescence microscopy was done by our collaborators (in particular Dr. Nancy Kedersha and Prof. Dr. Pavel Ivanov) in the Brigham and Women's Hospital (BWH) in Boston, MA, USA.

locations which do not suppress the desired functions of EBP1. And it seems that this phenomenon has indeed occurred. However, of the four Enamine compounds which were active in the binding assays this was only the case for one of the compounds, while the other three compounds had effects detrimental on the cancer cells which was visible in the fluorescence microscopy experiments. Fluorescence images of the cells treated with the compound Z9 are shown in Figure 9.10 on the preceding page. However, the effects on the cancer cells by Z9 do not necessarily have to come from binding to EBP1 as desired, it could in theory also be that Z9 causes these effects by some other unpredicted interactions within the cells. To find indications whether this happens or not can also be done with fluorescence microscopy in different ways. On the one hand the location of EBP1 can be traced by fluorescent anti-EBP1 antibodies when the cells are treated with Z9, with other

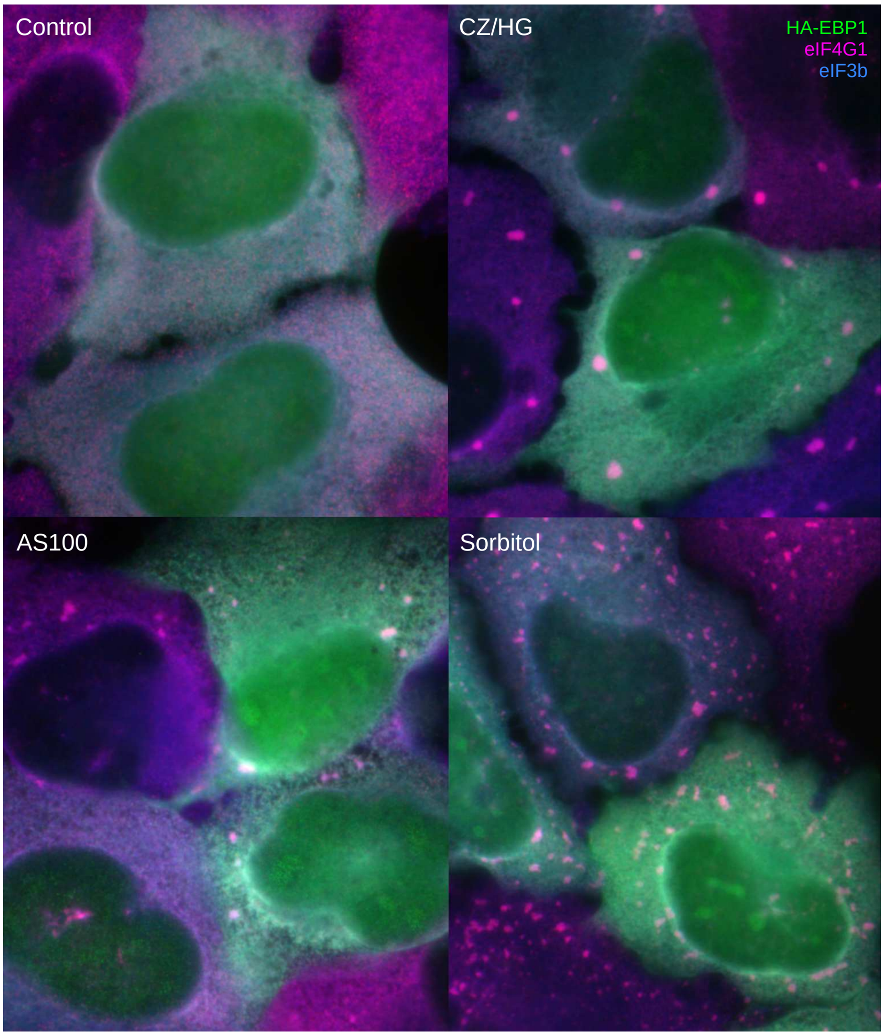


Figure 9.11: Fluorescence microscopy images of human U2OS cancer cells which were treated by different types of stresses/toxic agents. Top left: Control. Top right: Energy starvation caused by 20 µM clotrimazole in high-glucose media (CZ/HG). Bottom left: Oxidative stress induced by 100 µM sodium arsenite (AS100). Bottom right: Osmotic stress caused by 0.4 M sorbitol under hyperosmotic conditions. HA-tagged EBP1 (green) appears to aggregate within the nucleoli under all types of stresses, but is not recruited to stress granulas.

types of toxic agents, or be left untreated at all. This was done, and the tests showed that in the cancer cells which were treated with Z9 the location of EBP1 changed in a unique way, while in the cells which were treated with various other toxic agents did not exhibit a specific translocation of EPB1. Special U2OS cell lines were created for this purpose which are able to stably express the HA-tagged p48 isoform of EBP1 via the mammalian expression vector. Fluorescence images of of the effects of some of the toxic agents which were used can be seen in Figure 9.11 on the preceding page.

As can bee seen when treated by CZ/HG, AS100, and sorbitol which cause energy starvation, oxidative stress and osmotic stress, respectively, EBP1 seems to concentrate within the nuclei/nucleoli. In contrast to these results when the cancer cells are treated with Z9 the targeted protein EBP1 concentrates within stress granulas in the cytosol (see Figure 9.12). Therefore Z9 seems to have a specific effect on EBP1. And moreover the experiments indicate that EBP1 is kept out from the nuclei when treated with Z9, which was one of the effects which were aimed for since EBP1 is expected to carry out its oncogenic functions within the nuclei.

Alternatively, to find out whether EBP1 is specifically effected by Z9 other cellular structures can be monitored which are closely related to the functionality of EBP1. This was done by tracing the proteins TIAR and G3BP1. TIAR has several functions directly related to ribosomal protein expression, which is also the case for EBP1 (since it is a part of the pre-ribosomal ribonucleoprotein complexes). G3BP1 on the other hand has regulatory functions related to the global expression of proteins. Fluorescence microscopy (no images shown) has shown that U2OS cancer cells treated with Z9 causes TIAR to aggregate in stress granulas, but not G3BP1.

The above experiments indicate that Z9 has unique effects on EBP1, which

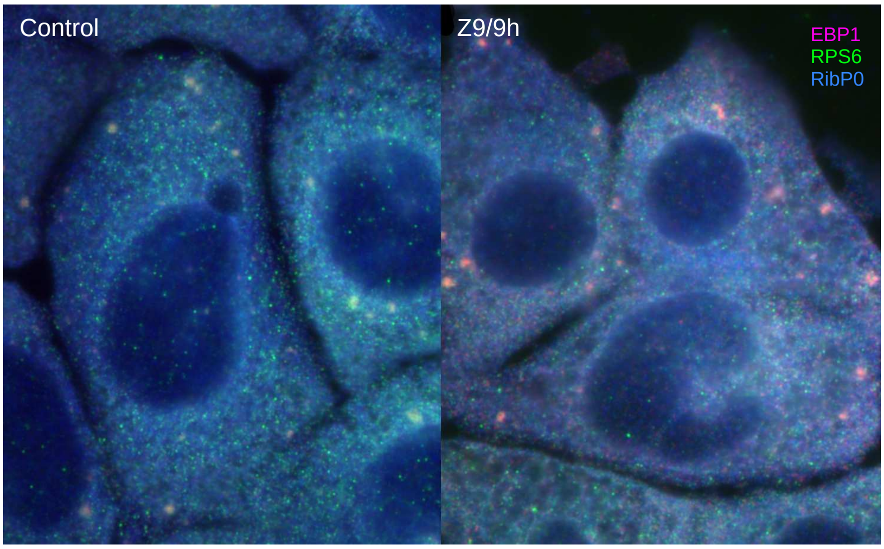


Figure 9.12: Fluoresence microscopy images of human U2OS cancer cells which were treated by compound Z9. EBP1 (magenta) appears to aggregate in stress granulas within the cytosol and thus outside the nuclei.

is in accordance with the results of the thermal shift assays which show that Z9 binds directly to this protein. Therefore the detrimental effects on the cancer cells caused by Z9 are most likely caused by inhibition of EBP1 by Z9, which was the desired effect. As mentioned earlier EBP1 was also shown to have hepatitis C enhancing functions, most likely within the nucleus, and therefore it could be that Z9 is not only a promising hit compound as an anti-cancer compound but possibly also as an anti-hepatitis compound.

Further studies will be necessary regarding Z9 as well as the other two active compounds regarding the binding to EBP1, its effects within the cells, and subsequently also the optimization of the compounds (hit expansion/lead optimization).

Chapter 10

Antibacterial Drug Discovery

> *I can't be as confident about computer science as I can about biology. Biology easily has 500 years of exciting problems to work on. It's at that level.*
>
> Donald Knuth

Contents

10.1 Introduction

According to the World Health Organization (WHO) the resistance of bacteria against antibacterials is one of the most serious threats to human health and is becoming a global emergency [334, 335, 164, 278]. The underlying cause is that bacteria and other microorganisms are able to become resistant against antibacterials/antimicrobials. The number of antibacterials is on the other hand relatively small. This becomes a serious problem when certain bacteria develop resistances against multiple antibacterials. In particular because the development of new antibacterials appears to be happening more slowly than the increase in capability of bacteria to form multidrug resistances, but also due to the circumstance that new antibacterials are often relatively similar to previous antibacterials which makes resistances against more simple [337]. Cases were already reported where people died due to infections with bacteria which were resistant to all antimicrobial drugs which were available [55]. Due to this imminent problem several countries and governments have created special programs and funding opportunities for the development of new antimicrobiotics, such as the U.S. via the *Antibiotic Development to Advance Patient Treatment Act* [1] or Germany by the *DART 2020* initiative [283], both of the year 2015.

This project is a joint endeavor between Magdalena Czuban of the Charité in Berlin and the author. The idea to apply virtual screenings to enhance the affinity of her compounds came from Magdalena Czuban, the idea to target the peptidoglycan arose during discussions and is due to us both, while the computational approach and procedure were designed and carried out by the author.

10.2 The Principle

Bacteria can be classified into four classes according to the type of their cell envelop (which is defined as the cell membrane and the cell wall) [260, 40, 80]:[1]

(i) **Gram-Positive Bacteria.** Strains of this type have relatively thick layers of peptidoglycan. These thick layers lead to positive results in the Gram staining procedure with the crystal violet dye.

(ii) **Gram-Negative Bacteria.** Strains of this type have relatively thin layers of peptidoglycan. These thin layers are not able to retain the crystal violet dye when decolorizing the bacteria in the Gram staining procedure.

(iii) **Acid-Fast Bacteria.** Acid fast-bacteria like mycobacteria also have a peptidoglycan, and they form a third class. They are negative in the Gram staining procedure but have a cell wall which is different than the one found in typical gram-negative bacteria. *Mycobacterium tuberculosis* is one of the more prominent species in this class [275].

(iv) **Bacteria Without Peptidoglycan.** There very few bacteria which do not have a peptidoglycan, primarily bacteria which do not possess a cell wall at all. Among them are the mollicutes [240] as well as the L-forms [171].

Almost every bacteria possesses a cell wall, with the notable exception of a few types. A vital function which the cell wall provides via the peptidoglycan (also called murein) is to prevent that the bacteria will burst apart due to the hydrostatic pressure inside the cytoplasm, which is roughly 20 MPa in many gram-negative bacteria, but can be as high as 200 MPa (1 bar = 10 MPa) for certain species such as the gram-positive pathogen *Staphylococcus aureus* [246, 324]. The structure of the cell wall of gram-negative bacteria, gram-positive bacteria as well as mycobacteria is illustrated in Figure 10.1 on the next page. Since the peptidoglycan is unique to bacteria and does not occur in human tissue, some of the antibacterials (e.g. most members of the β-lactam family) have the purpose to impair the peptidoglycan.The mechanism by which these antibiotics achieve their goal is by inhibiting the synthesis of the peptidoglycan.

Another, relatively new class of antimicrobia drugs are photosensitizers, which are photoactive and upon excitation with visible light of certain wavelengths form radicals which are highly reactive [140, 349, 346, 193, 110, 261, 136]. The combination of photosensitizer and activation with light is called phototherapy, which induces the creation of reactive oxygen species (ROS). Sufficient amounts of ROS are able to cause heavy damage to microorganisms causing their death. One important aspect of this approach is related to the circumstance that photosensitizers cause damage to almost any type of tissue when active, including

[1] This classification is informal and non-rigorous, but widely used in the scientific communities (in particular classes (i) and (ii)).

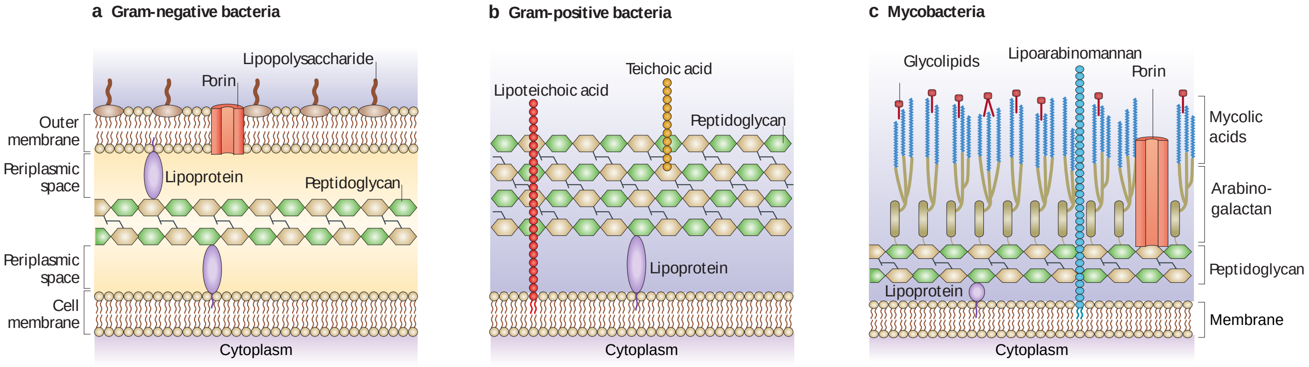


Figure 10.1: Architecture of the cell envelopes of gram-negative bacteria, gram-positive bacteria, and mycobacteria. Adapted with permission from [40].

healthy tissue. Therefore the (inactive) photosensitizers compounds should ideally be highly affine to the pathogens to focus the damage which is caused by the ROSs to these organisms. But a highly specific localization of the photosensitizers to microorganisms such bacteria is still elusive.

Due to the above problematic we had the idea to develop a small molecule which is highly specific to a relatively large class of microorgansisms, and attach the photosensitizers to it, resulting in augmented photosensitizers. For this purpose a structural target would be ideal which is present in microorganisms in abundance so that large amounts of augmented photosensizers are able to bind, while not present in human tissue to reduce damage to healthy tissues. And as outlined above, the peptidoglycan satisfies all these criteria remarkably well, and since it is present in almost any bacteria one of the largest classes of microorganisms would be affected. Therefore we decided to target the peptidoglycan, in other words to find novel compounds with high affinity to the murein.[2] The peptidoglycan seems to be one of the most difficult targets perceivable for small molecules due to its structure of long, thin polymer chains which form a three-dimensional mesh, which substantially reduces the available interaction surface area since there is no pocket at all, and not even a flat surface as in most PPIs which are already very difficult to inhibit. Furthermore, the peptidoglycan is highly flexible and dynamic which not only makes binding, but also computational approaches to predict high-affinity binders, more challenging. The binding of small molecules to the peptidoglycan alone will likely not have detrimental effects of significance on the bacteria. But upon activation of the photoactive part of the augmented compounds they can be expected to react mainly with the bacteria to which they have bound.

[2] The project was planned to together by Magdalena Czuban and the author. The idea to target the peptidoglycan came from Magdalena Czuban, while the computational approach was designed and carried out by the author.

10.3 Methods

We decided to run a large-scale structure-based virtual screening with the peptidoglycan as the target. Not only in order to identify compounds which have a high predicted binding affinity to it, but also to find out which classes of molecules might be particularly promising for targeting the peptidoglycan. The latter can be important because even though the compound library is large, it is only a drop in the ocean when considering the entire space of small chemical organic compounds (as pointed out in 5.2 on page 94). And therefore it can easily happen that of certain specific classes which are particularly suitable for the peptidoglycan only a few representatives of them are present in the screened library, and moreover that there are considerably stronger members in these classes which are not present in the library. For such cases the hit optimization/expansion step which follows the hit identification step is normally included in the drug discovery pipeline.

To the best of knowledge to the author a virtual screening of the peptidoglycan has so far never been carried out, therefore this project has some pioneering character. Since docking programs like AutoDock are based on atoms, and the same types of atoms are occurring in both the peptidoglycan and in proteins for which the docking functions were often designed, it can be expected that they work similarly well for the peptidoglycan.[3]

The three-dimensional structure of the murein was required for the structure-based drug design which we intended to carry out. This structure was long elusive due to the circumstances that suitable samples where hard to obtain, that structure determination can hardly be facilitated by crystallization due to its flexibility, and because its structure is relatively complicated which makes its determination by other methods more difficult. However, in the year 2006 a high-quality NMR structure was reported by Meroueh *et al.* [197]. Because the structure was not publicly available we have contacted the authors of the article and obtained it in the PDB format. The structure is shown in Figure 10.2 on the next page. For the virtual screening procedure a fragment of the segment shown in the figure was used which contained one full peptidoglycan monomer plus the nearby regions of the adjacent monomers for the case that a small molecule binds at the boundary region between two monomers.

The virtual screening procedure carried out contained approximately 130 million compounds. The docking program used was QuickVina 2 with an exhaustiveness level of 10. In the subsequent rescoring procedure 750 000 of the highest scoring compounds of the primary virtual screening were rescored with AutoDock Vina (exhaustiveness 4) and Smina Vinardo (exhaustiveness 10). The docking box was with a size of 45.75 Å × 27 Å × 25.5 Å the same in both the primary virtual screening and the rescoring procedure. The receptor was rigid in all cases due to the circumstance that flexible residues can only readily be modeled in a flexible way in proteins.

[3] The author was in contact with some of the developers of the used docking programs, and they have confirmed this assumption.

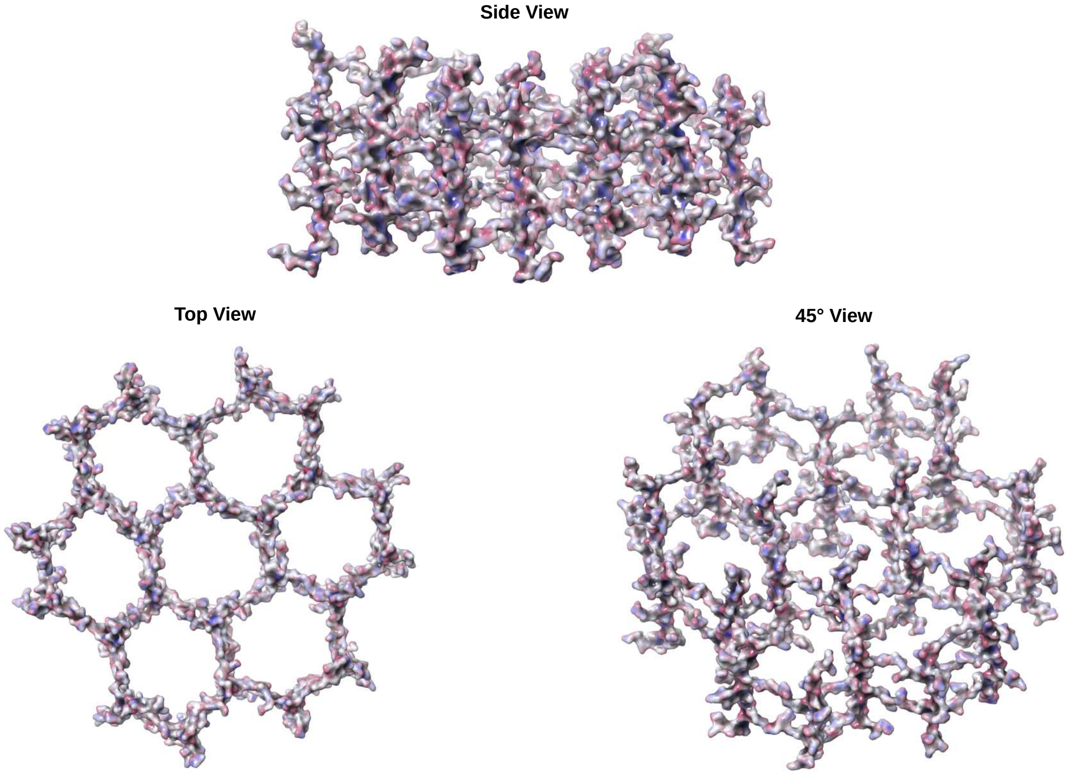


Figure 10.2: Three-dimensional structure of a polymeric peptidoglycan segment which was determined by NMR [197]. The shown structure contains approximately 12 000 atoms and has a diameter of roughly 220 Å.

10.4 Results and Discussion

Approximately 100 to 200 compounds were identified with predicted binding affinities stronger than −9 kcal/mol from the rankings obtained by the different scoring functions. Such affinities are relatively strong for a target like the peptidoglycan and the fact that the residues were not flexible and thus could not adapt to the ligands (which can have a substantial effect on the binding affinity).

One very interesting class of hit compounds which were exhibited by the virtual screening are large macrocycles whose diameter matches to the size of the strands of the peptidoglycan, allowing the cycles to fit on the strands like a ring on a finger. One example compound (referred to as B1 in this text) is shown in Figure 10.3 on the following page. The predicted binding pose of compound B1 to murein can be seen in Figure 10.4 on page 191 which involves only a small segment of the peptidoglycan.

Macrocycles have several advantages regarding physicochemical and pharmacokinetic properties. Due to the ring closure macrocycles have considerable less conformation freedom than if the same ring would be opened at one location. Increasing the rigidity generally means increasing the specificity since the small molecule is less able to adopt to other receptors. Also it can increase the binding affinity since the conformation entropy loss upon binding in trend decreases. Also, due to the cyclic structure the compounds can be protected from enzymatic

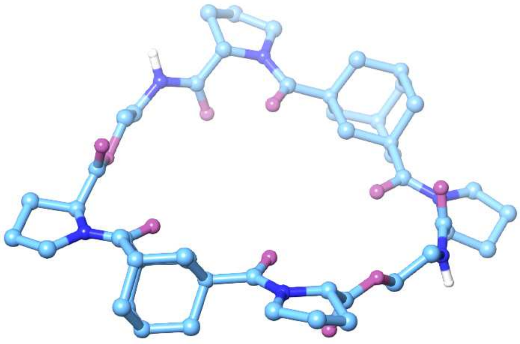


Figure 10.3: Compound B1, one of the macrocyclic hit compounds which are predicted to have a high binding affinity to the peptidoglycan.

degradation. Moreover, they can have an increased membrane permeability due to the increased likelihood of intramolecular hydrogen bonds and reduced polarity [192]. That such compounds can indeed make it to the market is demonstrated by fact that there exist already a variety of approved macrocycles [99]. Some of them are even used as antimicrobiotics such as the family of macrolides, which rather than having the purpose of binding to the peptidoglycan are inhibiting proteins resposible for the synthesis of the peptidoglycan [72]. Furthermore, some natural products are macrocycles, and they can have favorable biological properties ranging from anticancer, antifungal, or immunosuppressive effects [187, 347].

In Figure 10.2 on the preceding page where a larger segment is shown it appears as if the structure consists of a relatively regular and closed system of strands. In this case one question which arises is how could a macrocycle possibly bind one of the strands if the three-dimensional mesh structure is closed? In reality the peptidoglycan is not a perfect structure, which probably means that there are open strand ends in abundance. Moreover, the peptidoglycan also needs to be synthesized in the first place. In addition, other larger biological macromolecules need to be transported through the peptidoglycan such as fimbriae, flagella or other multiprotein complexes in order build the cell envelope, maintain it, and alter its composition during thee bacterial lifetime (e.g. if the environmental conditions change). To facilitate these transportations dedicated murein hydrolases open the petidogylcan at certain regions [309]. For these reasons macrocycles should be able to bind like a ring to the open ends of the murein mesh.

Outlook

Among the next steps which would be interesting and useful to be carried out would be experimental verifications in which at first selected hit compounds are tested for their real binding affinity to the peptidoglycan. And afterwards the true hit compounds should be attached to the photosensitizer and tested for their biological activities. For this purpose we are planning to write a dedicated grant application. Also, the computational procedure could be refined or extended with additional rescoring steps via either flexible dockings or free energy simulations (using for example QSTAR via HyperQ, possibly in combination with QUASAR via i-QI).

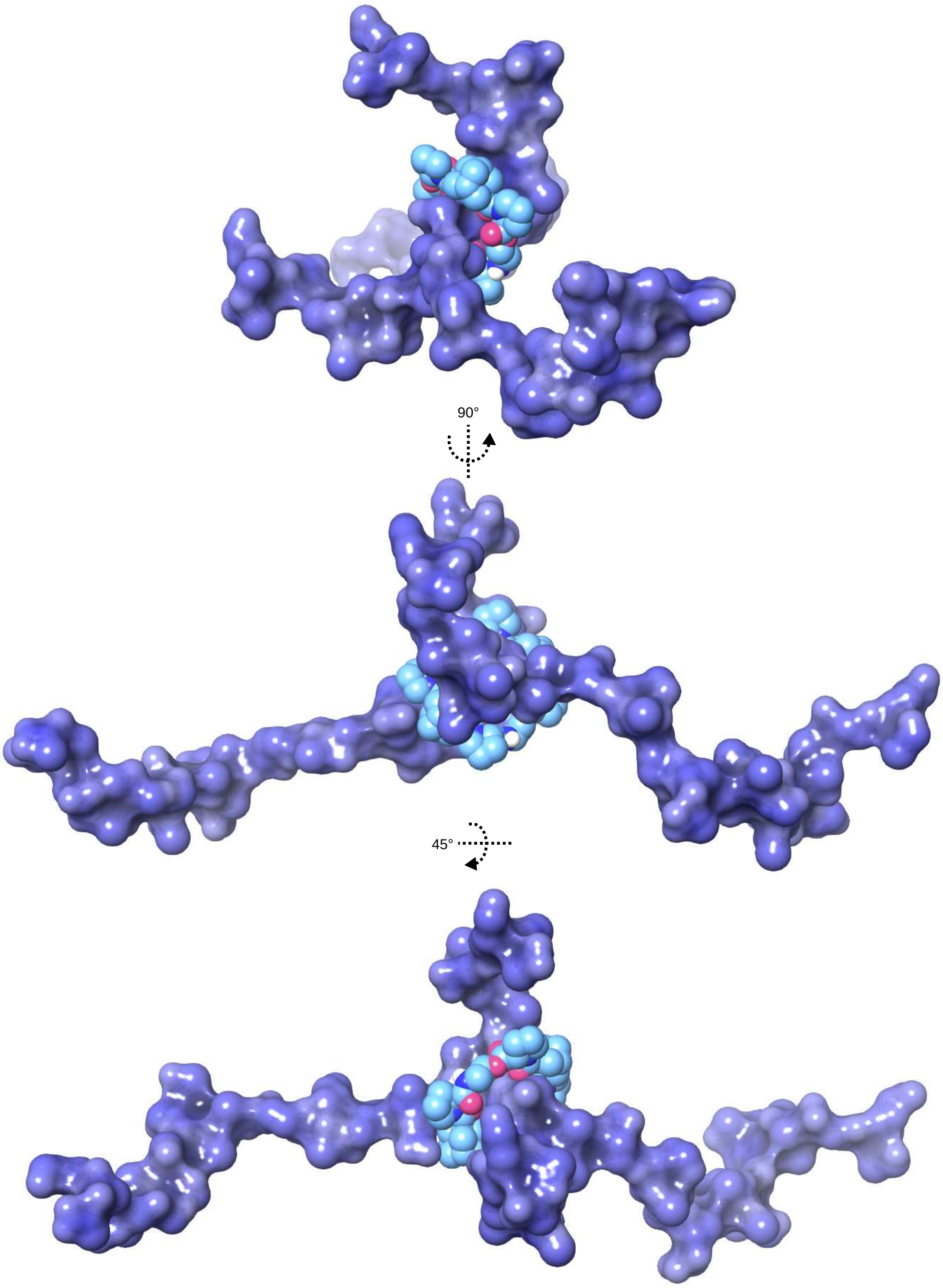


Figure 10.4: Predicted binding mode of one of the macrocyclic hit compounds (van der Waals sphere representation in light blue) to the peptidoglycan (dark blue surface) in the role of the receptor. The macrocycles diameter matches perfectly to thickness of one strand of the peptidoglycan which allows it to enclose it fully. Furthermore, it is located at a three-way junction of the peptidoglycan which is one of the regions with the highest possible surface area.

Chapter 11

Collaborative Projects

But in my opinion, all things in nature occur mathematically.
René Descartes
Correspondence with Mersenne

Contents

11.1 Introduction

The new methods and software were applied also to a number of projects from collaborating groups. The targets of these projects are the following:

- **SHP2** (Src-homology 2 domain-containing phosphatase 2): Anti-cancer drug development, hit expansion.
- **MED15** (mediator of RNA polymerase II transcription subunit 15): Anti-cancer drug development, both hit discovery and hit expansion.
- **eIF4E** (eukaryotic translation initiation factor 4E): Anti-cancer drug development, hit discovery.
- **UL50** and **UL53** (nuclear egress protein 2 and 1, respectively): Anti-herpes drug development, hit discovery.

Some of the computational results for these projects are currently under experimental verification, while others are anticipated to be investigated in the near future according to the responsible collaborators. The projects involving MED15 and SHP2 are briefly outlined below, while the projects involing eIF4E, UL50, and UL53 will be skipped due to the scope of this thesis.

11.2 SHP2

The protein Src-homology 2 domain-containing phosphatase 2 (SHP2), also known as tyrosine-protein phosphatase non-receptor type 11, is an enzyme and signaling protein with multiple roles in cells of humans and other species. Among them are the regulation of cell growth, differentiation, survival and mitosis. This protein is also involved in multiple types of diseases, such as the Noonan syndrome, LEOPARD syndrome, and several types of cancer. SHP2 related diseases are often related to a mutation in the PTPN11 proto-oncogene. Among the tumor types are in particular juvenile myelomonocytic leukemia, acute myeloid leukemia, myelodysplastic syndrome or acute lymphoblastic leukemia. First experimental drug candidates which modulate the function of SHP2 allosterically were reported in the recent years. [3, 169, 90]

SHP2, upon activation, increases the half-life of activated RAS-GTP, which is expected to be its primary mechanism of oncogenic action. Inhibition of SHP2 (via RNAi or allosteric inhibitors) reduces the viability of RTK-driven tumors [56]. Inhibition of SHP2 leads to the downstream reduction in Erk1/2 activation. RTK-driven tumors (especially those with EGFR amplifications or FLT-3 activation) are killed effectively upon SHP2 inhibition or depletion, and other types of cancer are currently under investigation.

One of the researchers working on the development of new inhibitors for SHP2 is Jonathan LaRochelle at the Dana Farber Cancer Institute in Boston. Upon his suggestion we started a new collaboration in which the author was responsible to find new analogs via computational methods for a certain experimental hit compound called JLR-1, i.e. to support the hit expansion step for this compound. The compound binds within a tunnel which is able to allosterically inhibit certain functions of SHP2. The targeted region and SHP2 are illustrated in Figure 11.1.

In order to carry out the hit expansion a customized ligand database was created which contained approximately 100 000 analogs of JLR-1. The analogs

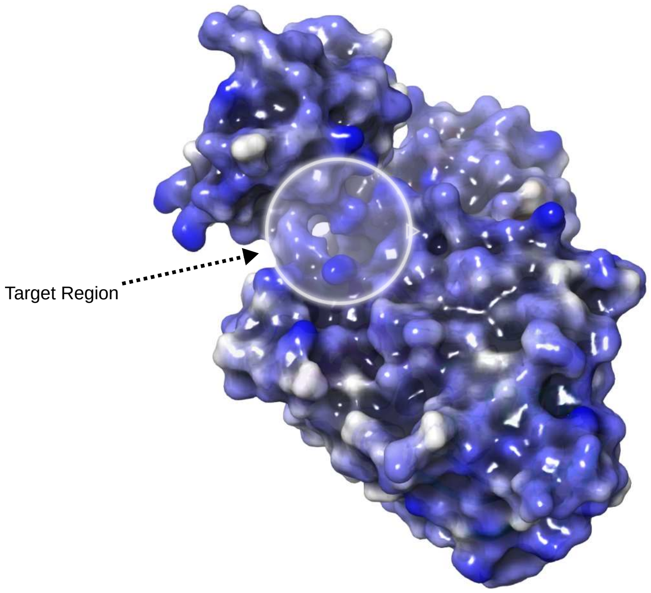


Figure 11.1: The signaling protein SHP2 and the allosteric site in form of a tunnel which is targeted (white circle). Inhibiting this region allows to allosterically modulate the activities of this protein, which in turn has detrimental effects on certain types of cancers. The surface (colored by the electrostatic potential) is based on a crystal structure from our collaborators.

were obtained by searched the entire ZINC15 database via molecular similarity measures (based on the Tanimoto and Dice similarity coefficients), as well by filtering the compounds of the MolPort database via pharmacophore modeling as made possible by PHARMIT [274]. Pharmacophore modeling was possible due to the availability of a co-crystal structure of SHP2 with one of the experimental analogs of JLR-1. This crystal structure was used as a starting point for the structure preparation step which included MD simulations of the target to obtain biologically relevant conformations. After that VFVS was used to carry out extensive docking procedures, and DATAWARRIOR was used to compute pharmacological and physico-chemical properties in order to facilitate further filtering. Experimental verifications are currently in progress.

11.3 MED15/KIX Domain

The Mediator complex, also known as the activator-recruited cofactor (ARC) complex, is a large multiprotein complex which is present in all eukariotic species. It is a primary regulator of a certain set of genes which are in particular related to lipid synthesis and homeostatis [255, 38]. The number of subunits in this multiprotein complex varies between different species, but in humans it expected to consists of 31 proteins: MED1, MED4, MED6, MED7, MED8, MED9, MED10, MED11, MED12, MED13, MED13L, MED14, MED15, MED16, MED17, MED18, MED19, MED20, MED21, MED22, MED23, MED24, MED25, MED26, MED27, MED29, MED30, MED31, CCNC, CDK8 and CDK11 [38].

In humans the Mediator complex is a transcriptional activator of the sterol regulatory element binding proteins (SREBPs) family of transcription factors [354]. Regarding the binding mechanism, the KIX-domain of the MED15 subunit of the Mediator complex is thought to bind to the transactivation domain (TAD) of the SREBPs transcription factors. The regulatory mechanism involving SREBP and the directly associated regulatory complexes are shown in Figure 11.2 on the following page.

11.3.1 Pathophysiology

SREBPs as well as the directly associated Mediator complex play critical roles in several diseases. Most of these roles are related to lipid synthesis and homeostatis, in particular involving cholesterol and fatty acids, of which SREBPs and the Mediator complex are the master regulators.

Obesity. Diet-induced obesity is directly linked with the dysregulation lipid homeostatis, and it was shown that inhibiting the interaction between MED15-KIX and SREBP-TAD can improve the lipid homeostatis in such cases [354].

Type 2 Diabetes. The disregulation of lipid homeostatis is also related to type 2 diabetes, and roles of SREBPs in this disease were confirmed [308, 100].

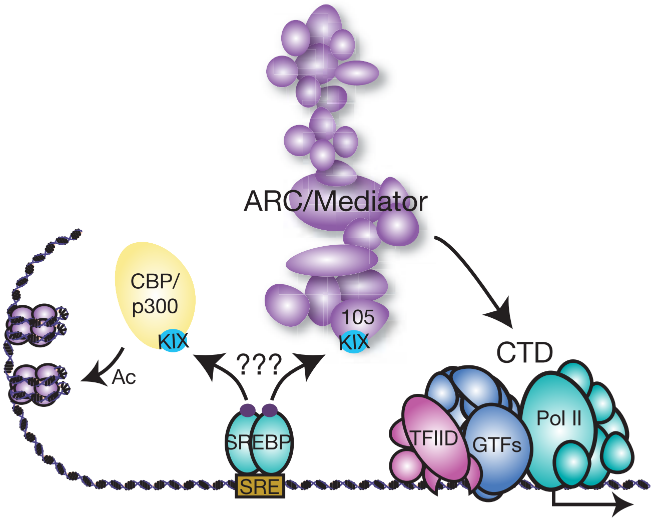


Figure 11.2: Outline of the regulatory mechanism involving SREBPs. One of the primary regulatory complexes is the Mediator complex which binds via the KIX domain of its mediator of RNA polymerase II transcription subunit 15 (MED15) subunit to the TAD domain of SREBPs. Adapted with permission from [342].

Interestingly it was shown recently that the herbal plant silymarin (milk thistle) has positive effects on diabetes by suppressing SREBP-1c which helps to prevent the accumulation of lipids in the liver [254, 145].

Cardiovascular Diseases. SREBPs are also involved in the cardiovascular homeostatis. And it was shown that in people with coronary heart disease the expression level of SREBPs is considerably increased. [340, 221]

Fungal Multidrug Resistance. Like bacteria fungi are also able to form multiple drug resistances [106]. The Mediator complex activates orthologs of the pleiotropic drug resistance transcription factor (Pdr1). These proteins contribute to the multidrug resistance in several types of fungi, including the pathogen *Candida glabrata* [211].

Cancer. In human tumors the expression levels of proteins such as fatty acid synthase are upregulated [159, 254] in order to produce sufficient fatty acid required for the tumor proliferation. It is therefore not surprising that SREBPs play an important role in a variety of cancer types [157, 254, 282, 132]. The types of cancer which are affected are currently investigated. Some of them are already known such as glioblastoma [174], and other types are currently investigated by our collaborators (unpublished) whose experiments so far have shown that various cancer types depend vitally on the activity of SREBPs and the Mediator complex.

11.3.2 Computational Drug Discovery.

The activation of lipid synthesis and related components by SREBPs can be inhibited by disrupting interaction between the TAD-domain of SREBP and the KIX-domain of MED15. That this is possible in principle was already shown as there have already a few biologically active compounds been reported [211, 354]. However, these compounds suffer either from weak binding affinities or other highly unfavorable properties. For these reasons this computational project started with the goal of identifying either new hit compounds or analogs of existing hits by structure-based approaches.[1]

Hit Identification. In order to find new hit compounds a dual-stage large scale virtual screening procedure was carried out involving 180 million compounds of a snapshot of the ZINC15 database from the end of the year 2016. As the starting structure of the KIX-domain the PDB file 2GUT was used which was published in the year 2006 and determined by NMR techniques [342]. The docking program which was used for the virtual screenings was QuickVina 2, the exhaustiveness was set to. The docking box had a size of 23.25 Å × 21.00 Å × 20.25 Å.From the primary virtual screening the 3 million highest scoring compounds were taken and used as input for the rescoring procedure in which 12 residues of the receptor were allowed to be flexible. Two docking programs were used during the rescoring procedure, AutoDock Vina as well as Smina Vinardo (both with exhaustiveness 2). The docking box size was set to 51.0 Å × 51.0 Å × 51.0 Å^2.

Tens of thousands of high-affinity compounds were obtained with the above procedure. To filter the compounds further a variety of physico-chemical and pharamcokinetic properties were computed with DataWarrior, in particular the following:

- Corrected[2] minimal[3] predicted free energy of binding: ≤ -12 kcal/mol
- Molecular weight: ≤ 700 Dalton
- logP (octanol/water): ≤ 5
- clogS: ≥ -6
- Hydrogen bond accepting atoms: ≤ -12
- Hydrogen bond accepting and donating atoms: ≥ 3
- Druglikeness measure: ≥ -10
- Tumogenic risk assessment: ≤ low
- Mutagenic risk assessment: ≤ low
- Reproductive effect risk assessment: ≤ low

Applying these filtering criteria resulted in approximately remaining 8 000 hit compounds.The large number of hits in the filtered lists made a vendor clustering

[1] The project was started during the stay of the author in the Arthanari lab of the Dana Farber Cancer Institute and the Wagner Lab of the Harvard Medical School, and continued afterwards.

[2] Docking scores of different programs were scaled to be directly comparable.

[3] The minimum binding free energy among all the docking programs, i.e. the highest predicted binding strength.

attractive which was carried out as further selection criteria.

Hit Expansion. Regarding the identification of improved analogs of an experimentally confirmed binder called C2 the ZINC database was searched for the 50 000 most similar compounds to C2 based on the Dice and Tanimoto coefficients. Regarding the receptor structure (starting from 2GUT) conformational sampling was carried out with HyperQ (energy minimization and MD simulations) and 13 different conformations selected for the docking procedures. Afterwards two types of docking runs were carried out.

(1) **Rigid Docking.** The receptor structures were rigid. All of the 50 000 identified analogs were docked. The exhaustiveness of AutoDock Vina and Smina Vinardo was set to 10.

(2) **Flexible Docking.** Of the KIX-domain 14 residues were allowed to be flexible. The most similar 1 500 compounds were docked. The exhaustiveness of AutoDock Vina and Smina Vinardo was set to 1.[4]

In each of these types each ligand was docked 5 times to each of the 13 different receptor backbone conformations with each of the docking programs AutoDock Vina as well as Smina Vinardo, giving rise to 52 different docking scenarios and 260 dockings per ligand. The docking box size varied with each docking scenario, but was on average about 25 Å × 25 Å × 25 Å in size.

Regarding the results, approximately 5 % of the 50 000 compounds of in the rigid docking procedure had a stronger predicted binding affinity than the original analog C2, i.e. around 2 500 compounds. On the other roughly 10 % of 1 500 compounds docked with a flexible receptor were predicted to have a higher binding affinity than C2, i.e. around 150 compounds.

Outlook. Experimental verifications of both the results of the hit identification as well as the hit expansion approaches are planned to be carried out in the near future by the Wagner Lab and the Arthanari Lab in Boston which are primarily working on this project.

[4] In the experience of the author higher exhaustiveness values did not improve the results significantly, rather multiple independent docking runs per ligand seemed to be more beneficial.

Chapter 12

Conclusion and Outlook

Imagination is more important than knowledge. For knowledge is limited, whereas imagination encircles the world, stimulating progress, giving birth to evolution.

Albert Einstein
What Life Means to Einstein

In this thesis novel methods were developed and implemented which are particularly useful for CADD, and they were applied in a number of drug discovery projects. Among these methods were QSTAR, a new quantum mechanical free energy method based on path integrals which was implemented in free energy suite HyperQ. The QUASAR scheme for QM/MM simulations was implemented in i-QI, which is a client for the path integral software i-PI, and which can be used in concert with HyperQ. Furthermore, Virtual Flow, a virtual screening framework was developed which currently consists of two packages, VFLP and VFVS.

Free Energy Simulations

Overview of Contributions. The HyperQ package which implements the QSTAR method allows for the first time to explicitly and conveniently include nuclear quantum effects in free energy simulations of biomolecular systems. The inclusion can significantly increase the accuracy of the results since the negligence of the quantum nature of atomic nuclei can be a major source of error in biomolecular systems, possibly as large as the error due to electronic-structure methods [50]. This is the case because nuclear quantum effects are the more pronounced the lighter the nuclei are, and hydrogen atoms, the lightest of all atoms, are the most abundant atom type in biomolecular systems. Since non-covalent ligand-receptor interactions almost always involve hydrogen atoms (for instance via hydrogen bonds or nearby solvent molecules) nuclear quantum effects can be expected to play a significant role in free energies of binding. QSTAR and HyperQ can be principally be used together with any electronic structure method available, and therefore allow to carry out alchemical free energy simulations on the highest accuracy level possible to date. For the task of computing relative binding free energies a new alchemical scheme (PEARL) was developed which is based on the serial insertion approach. One of its major advantages is that it can be used for carrying out relative binding free energy simulation with MD packages which are not dedicated for such purposes (and therefore might lack features such as softcore potentials). PEARL was implemented in HyperQ as well. HyperQis particularly suitable for the hit and lead optimization steps of CADD, but can also be used as the FEM in later steps of multistage virtual screening procedures, possibly in combination with VFVS (see below). However, since HyperQ can also be used for vanilla PIMD simulations it principally be used in the (receptor) structure preparation step as well.

Computational Expense. One of the major disadvantages of QSTAR method, and PIMD simulations in general, is the circumstance that they are computationally relatively expensive. However, their computational overhead is relatively small in comparison to electronic structure methods, which are usually several magnitudes slower than MM simulations. Moreover, while at the time of writing the computational expense is a limiting factor, the speed of computers will continue to increase over the years, and it is only a question of time until PIMD-based simulations of biomolecular systems can be carried out routinely not only for free energies, but also for other purposes such as conformational sampling. Until this time is reached, there are several ways to accelerate the PIMD simulations already now. Several advanced PIMD simulations are implemented in i-PI such as ring polymer contraction RPC. Furthermore, HyperQ can run QM/MM simulations with i-QI (see below in the paragraph about QM/MM approaches)). Also the nature of the QSTAR method is expected to provide a relatively favorable convergence behavior due to the single-topology approach.

Future Work. The QSTAR method and the HyperQ package were thoroughly tested regarding the technical functionality. They are now able to run robustly, conveniently (e.g. autonomously), and passed verification benchmarks indicating that the methods are implemented correctly.

What would be useful as a next step would be extensive tests with real systems. While models which include the quantum nature of the nuclei are in theory more accurate than corresponding methods which neglect them, in practice the situation are more intricate and challenging (for instance regarding verifications). For instance due to the sampling time (which depends also on the computational expense of the method) which plays an important role, and sufficient computational resources have to be available. Some potentials/parametrizations work better for certain systems than others. Errors can cancel in certain circumstances. All these aspects can make it a substantial challenge to find out which of the available options work best in general or for certain systems. What complicates this matter further is the vast array of options available when running free energy simulations with HyperQ. On the one hand there are plenty of options which the external MD programs make available. CPK2 alone provides thousands of options including various types of electronic structure methods, from MM force fields over semiempirical methods such as PM6 or DFTB3 to *ab initio* potentials such as second MP2. Furthermore there are the options which i-PI provides which drives the MD simulations. And then there are the various options and features which HyperQ makes available. Finally, additional clients can be used to run in concert with i-PI and CP2K, such as i-QI (see below). Extensive tests will for the above reasons consume a considerable amount of time and resources, but should be carried out and will most likely prove useful for further applications.

While the HyperQ software package is already publicly available as free and open source software, what needs to be done is the preparation of a thorough documentation, example input data, a mailing/list or forum, and a homepage which makes available all the additional resources in a convenient way.

Furthermore, there are many additional features which could be implemented in the future, in particular regarding HyperQ. New functionalities could for instance be capabilities to compute absolute binding free energies, other types of transfer free energies such as distribution coefficients, more extended support for other types of macromolecules such as RNA or DNA, full support for non-Boltzmann sampling/reweighting, or the implementation of additional free energy methods such as MBAR.

QM/MM Approaches

Overview of Contributions. When using electronic structure methods the computational speed can be dramatically increased when applying QM/MM methods. We have developed a QM/MM scheme for diffusive systems called QUASAR. QUASAR is particularly suitable for equilibrium properties like free energies and it allows the definition of arbitrary shapes QM/MM regions, which is particularly favorable for biomolecular systems. QUASAR was implemented in i-QI, a client for i-PI, which is fully supported by HyperQ.

Future Work. The QUASAR method can in principle also be implemented in other MD codes which support QM/MM simulations either directly or indirectly, such as CHARMM or NWChem. Also the QUASAR method can be extended, for instance to support additional types of constraints such as hyperplanes defining half spaces. Like HyperQ also i-QI requires a documentation and homepage.

Virtual Flow

Overview of Contributions. With the development of the Virtual Flow suite we have created a highly scalable, flexible, and efficient workflow system which allows to carry out virtual screening related tasks on computer clusters of many types and virtually any size. Two versions of Virtual Flow were implemented which share the same core technology (in particular regarding the parallelization). The first one, VFLP, is dedicated to large ligand databases. It can be used seamlessly with VFVS which is tailored for carrying out the virtual screening and rescoring procedures itself. In multistage virtual screenings VFVS can for instance also be used in concert with HyperQ, where the latter is an attractive option for the last rescoring steps (possibly together with i-QI for applying the QUASAR method). Both VFLP and VFVS were extensively tested, and after several major revisions of the underlying mechanisms are able to run reliably.

VFLP was used two times for preparing the entire ZINC15 database into a ready-to-dock format with VFVS. While the ZINC database provides molecules in a ready-to-dock format it does so only with a fraction of compounds. In the year 2014 when ZINC15 was still in beta it contained only around 10 % of the database, while in the year 2016 it were around 40 %. Two snapshots of the entire ZINC15 database, one of each of these two years, were therefore converted with VFLP. The 2014 version contained 133 million commercially available compounds, while the 2016 version contained approximately 300 million compounds.

VFVS was applied by the author in several real drug discovery projects for different purposes such as virtual screenings, rescorings, hit expansions, and thorough docking procedures.

Future Work. VFLP can be used for the creation of the publicly available ligand databases of the future. While the currently available ligand databases providing ligands in a ready-to-dock format already contain millions of molecules such as ZINC15, this number becomes vanishingly small when considering the chemical universe of small organic molecules, which is estimated to contain between 10^{60} and 10^{70} compounds (which are roughly as many as there are atoms in our entire galaxy). As pointed out in chapter 5 the scale of the screening matters. Therefore it will be beneficial if even larger molecule databases become available. One candidate database would be for example the GDBs (generated databases) created by the group of Jean Louis Reymond [29, 243]. These databases contain molecules particularly suitable for drug discovery, but only in the SMILES format rather than in a ready-to-dock format. The GDB-17 database for instance contains 166 billion compounds, and would render a promising future application for VFLP.

VFLP is currently not on the latest version of the Virtual Flow framework, and thus lacks a few features (e.g. regarding the types of batchsystems which are natively supported). Since these features are already implemented in VFVS, the update of VFLP will be relatively trivial.

The common core architecture of Virtual Flow is currently two times implemented. In one instance it is used for the preparation of ligand databases (via VFLP) while in the other is used for docking-related tasks (via VFVS). However, the architecture can in principal be used also for other types of tasks as long as the type of workflow remains the same (i.e. many jobs which are independent from each other and which can be managed as tasks in a single task list). Because such types of tasks occur frequently, a general version of Virtual Flow for arbitrarily definable tasks might be promising, where the task can be defined for instance via external scripts.

A documentation for VFLS and VFVS already exists, as well as a homepage, but it is planned to a create new and improved website for these packages.

Computational Drug Discovery

Overview of Contributions. A dedicated drug discovery project was started for which a novel protein target was selected, EBP1, which contributes to the genesis and proliferation of various types of cancer including (including certain types of skin cancer, brain cancer, or oral squamous cell carcinoma). It also contributes to the proliferation of HCV. Three different target regions were identified, and

VFVS was applied in the hit identification step of each of these regions to carry out a dual-stage virtual screening procedure involving 130 million compounds. In all cases the virtual screening has provided hundreds to thousands of promising hit compounds (depending on where the threshold is set). Experimental verifications by newly established collaborations revealed a virtual screening hit rate of above 30 %, which is remarkably high for a target involving a relatively flat PPI like in this case and without using any tailored ligand database. Further experiments showed that three of the four (possibly 5) active compounds also have inhibiting effects on living cancer cells. One of them (Z9) was investigated in detail, and it was observed that the compound causes the cancer cells to become very sick, possibly even dying. It was also possible to show that EBP1 is aggregating in the cytosol rather than the nucleus when treated with this compound. This was the desired effect, and furthermore might imply that this compound might be an dual-cancer/HCV inhibitor since EBP1 is expected to carry out its HCV enhancing activities within the nuclei.

Another joint antibiotics drug discovery project targeting the peptidoglycan was started by Magdalena Czuban and the author. It consisted of the first structure-based virtual screening approach reported targeting directly at the peptidoglycan, and yielded more than a hundred promising hit compounds. Among them was for instance one promising molecule class in the form of macrocycles. It might be useful to carry out additional rescoring steps to refine the results further, and based on them to subsequentially carry out experimental verifications. A grant application is plant for this purpose.

VFVS was also applied in a variety of other drug discovery projects which were not initially planned mainly due to the early interest by other research groups. Among the targeted receptors are ARC105 (cancer), KEAP1 (cancer), eIF4E (cancer), SHP1 (cancer), UL50 (herpes) and UL53 (herpes). HyperQ was also applied in first initial tests in two of these projects. On the computational level the results (either hit discovery, hit expansion, or lead optimization) for all these projects appeared to be promising. Experimental verifications for most of these projects are already underway or planned for the near future.

Future Work. Regarding EBP1, more experimental studies are required for the obtained experimental hit compounds, and if the results continue to be promising, hit/lead optimization procedures are among the next steps which need to be carried out. While in the first verification round only 12 compounds where tested, a second round is in preparation involving approximately 50 compounds. The second round also includes already compounds from virtual screenings/rescorings which were intended for the second and the third target regions. Regarding the other applied projects on which we have worked on, experimental verifications are also the next steps which need to be carried out.

In the future the new methods and software packages which were developed in this thesis can be applied to the discovery of new medications of other diseases as well. Hopefully not only by us, but also by other researches, which is one reason why we have tried to make the usage of the new software as convenient and automatic as possible. Due to their elaborate nature, the new methods and software might in particular be beneficial also for targeting the challenging class of PPIs. PPIs are among the most abundant and common interactions living cells, are highly specific, and play a role in most known diseases. Therefore they are likely belonging to the future of drug discovery.

Bibliography

[1] 114th U.S. Congress (2015-2016). *H.R.2629 - Antibiotic Development to Advance Patient Treatment Act.* 2015.

[2] Abreu, R. M. V., Froufe, H. J. C., Queiroz, M. J. R. P., and Ferreira, I. C. F. R. "MOLA: A bootable, self-configuring system for virtual screening using AutoDock4/Vina on computer clusters". In: *Journal of Cheminformatics* 2.1 (2010), p. 10. doi: 10.1186/1758-2946-2-10.

[3] Aceto, N., Sausgruber, N., Brinkhaus, H., Gaidatzis, D., Martiny-Baron, G., Mazzarol, G., Confalonieri, S., Quarto, M., Hu, G., Balwierz, P. J., Pachkov, M., Elledge, S. J., Nimwegen, E. van, Stadler, M. B., and Bentires-Alj, M. "Tyrosine phosphatase SHP2 promotes breast cancer progression and maintains tumor-initiating cells via activation of key transcription factors and a positive feedback signaling loop". In: *Nature Medicine* 18.4 (Apr. 2012), pp. 529–537. doi: 10.1038/nm.2645.

[4] Acevedo, O. and Jorgensen, W. L. "Cope elimination: Elucidation of solvent effects from QM/MM simulations". In: *Journal of the American Chemical Society* 128.18 (2006), pp. 6141–6146. doi: 10.1021/ja057523x.

[5] Al, M. J., Feenstra, T. L., and Hout, B. A. van. "Optimal allocation of resources over health care programmes: dealing with decreasing marginal utility and uncertainty". In: *Health Economics* 14.7 (July 2005), pp. 655–667. doi: 10.1002/hec.973.

[6] *Alchemistry.org.* url: http://www.alchemistry.org[Accessed:2018-03-16]. 2018.

[7] Aldeghi, M., Heifetz, A., Bodkin, M. J., Knapp, S., and Biggin, P. C. "Accurate calculation of the absolute free energy of binding for drug molecules". In: *Chem. Sci.* 7.1 (2016), pp. 207–218. doi: 10.1039/C5SC02678D.

[8] Alhossary, A., Handoko, S. D., Mu, Y., and Kwoh, C. K. "Fast, accurate, and reliable molecular docking with QuickVina 2". In: *Bioinformatics* 31.13 (2015), pp. 2214–2216. doi: 10.1093/bioinformatics/btv082.

[9] André, J.-M. "The Alliance of Newton's Apple and Schrödinger's Cat". In: *Chemistry International* March-Apri (2014), pp. 2–7.

[10] Antoci, S. "The Third Way to Quantum Mechanics is the Forgotten First". In: *Annales de la Fondation Louis de Broglie* 21.3 (1996), pp. 349–368. arXiv: 9704028v1 [arXiv:physics].

[11] Antoci, S. "Wentzel's Path Integrals". In: *International Journal of Theoretical Physics* 37.1 (1998), pp. 531–535. doi: 10.1023/A:1026628515300.

[12] Baba, N. and Akaho, E. "VSDK: Virtual screening of small molecules using AutoDock Vina on Windows platform". In: *Bioinformation* 6.10 (Aug. 2011), pp. 387–388. doi: 10.6026/97320630006387.

[13] Badyal, D. and Desai, C. "Animal use in pharmacology education and research: The changing scenario". In: *Indian Journal of Pharmacology* 46.3 (2014), p. 257. doi: 10.4103/0253-7613.132153.

[14] Baker, C. M., Anisimov, V. M., and MacKerell, A. D. "Development of CHARMM polarizable force field for nucleic acid bases based on the classical drude oscillator model". In: *Journal of Physical Chemistry B* 115.3 (2011), pp. 580–596. doi: 10.1021/jp1092338. arXiv: NIHMS150003.

[15] Barducci, A., Bonomi, M., and Parrinello, M. "Metadynamics". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.5 (Sept. 2011), pp. 826–843. doi: 10.1002/wcms.31.

[16] Becke, A. D. "Density-functional thermochemistry. III. The role of exact exchange". In: *The Journal of Chemical Physics* 98.7 (1993), pp. 5648–5652. doi: 10.1063/1.464913. arXiv: z0024.

[17] Beddell, C. R., Goodford, P. J., Norrington, F. E., Wilkinson, S., and Wootton, R. "Compounds Designed To Fit A Site Of Known Structure In Human Haemoglobin". In: *British Journal of Pharmacology* 57.2 (June 1976), pp. 201–209. doi: 10.1111/j.1476-5381.1976.tb07468.x.

[18] Beierlein, F., Lanig, H., Schürer, G., Horn, A. H. C., and Clark, T. "Quantum mechanical/molecular mechanical (QM/MM) docking: an evaluation for known test systems". In: *Molecular Physics* 101.15 (2003), pp. 2469–2480. doi: 10.1080/0026897031000092940.

[19] Ben-Naim, A. and Marcus, Y. "Solvation thermodynamics of nonionic solutes". In: *The Journal of Chemical Physics* 81.1984 (1984), p. 2016. doi: 10.1063/1.447824.

[20] Bennett, C. H. "Efficient estimation of free energy differences from Monte Carlo data". In: *Journal of Computational Physics* 22.2 (1976), pp. 245–268. doi: 10.1016/0021-9991(76)90078-4.

[21] Bergeler, M., Mizuno, H., Fron, E., and Harvey, J. N. "QM/MM-Based Calculations of Absorption and Emission Spectra of LSSmOrange Variants". In: *The Journal of Physical Chemistry B* 120.49 (Dec. 2016), pp. 12454–12465. doi: 10.1021/acs.jpcb.6b09815.

[22] Berman, H., Henrick, K., and Nakamura, H. "Announcing the worldwide Protein Data Bank". In: *Nature Structural & Molecular Biology* 10.12 (Dec. 2003), pp. 980–980. doi: 10.1038/nsb1203-980.

[23] Bernstein, N. and Mones, L. "Tests of an Adaptive QM/MM Calculation on Free Energy Profiles of Chemical Reactions in Solution". In: *The Journal of Physical Chemistry B* (2013).

[24] Bernstein, N., Várnai, C., Solt, I., Winfield, S. a., Payne, M. C., Simon, I., Fuxreiter, M., and Csányi, G. "QM/MM simulation of liquid water with an adaptive quantum region." In: *Physical Chemistry Chemical Physics* 14.2 (2012), pp. 646–56. doi: 10.1039/c1cp22600b.

[25] Best, R. B., Zhu, X., Shim, J., Lopes, P. E., Mittal, J., Feig, M., and MacKerell, A. D. "Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain χ1and χ2Dihedral Angles". In: *Journal of Chemical Theory and Computation* 8.9 (2012), pp. 3257–3273. doi: 10.1021/ct300400x. arXiv: NIHMS150003.

[26] BiochemLabSolutions Co. *Protein-ligand docking and in-silico virtual screening for windows.* url: http://biochemlabsolutions.com/Molecule_Docking.html [Accessed: 2018-02-12]. 2010.

[27] BIOVA Co. (fomerly Accelrys). *Molecular Simulations With Biova Discovery Studio: Datasheet.* url: http://accelrys.com/products/datasheets/simulation.pdf [Accessed: 2018-02-12].

[28] Blank, J., Exner, P., and Havlicek, M. *Hilbert Space Operators in Quantum Physics.* 2nd ed. Vol. 1. 2008.

[29] Blum, L. C. and Reymond, J. L. "970 Million druglike small molecules for virtual screening in the chemical universe database GDB-13". In: *Journal of the American Chemical Society* 131.25 (2009), pp. 8732–8733. doi: 10.1021/ja902302h.

[30] Böckmann, M., Doltsinis, N. L., and Marx, D. "Adaptive Switching of Interaction Potentials in the Time Domain: An Extended Lagrangian Approach Tailored to Transmute Force Field to QM/MM Simulations and Back". In: *Journal of Chemical Theory and Computation* 11.6 (June 2015), pp. 2429–2439. doi: 10.1021/acs.jctc.5b00142.

[31] Bohacek, R. S., McMartin, C., and Guida, W. C. "The art and practice of structure-based drug design: A molecular modeling perspective". In: *Medicinal Research Reviews* 16.1 (1996), pp. 3–50. doi: 10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6.

[32] Boresch, S. and Bruckner, S. "Avoiding the van der Waals endpoint problem using serial atomic insertion". In: *Journal of Computational Chemistry* 32.11 (Aug. 2011), pp. 2449–2458. doi: 10.1002/jcc.21829.

[33] Boresch, S. and Karplus, M. "The Role of Bonded Terms in Free Energy Simulations. 2. Calculation of Their Influence on Free Energy Differences of Solvation". In: *The Journal of Physical Chemistry A* 103.1 (Jan. 1999), pp. 119–136. doi: 10.1021/jp981629f.

[34] Boresch, S., Tettinger, F., Leitgeb, M., and Karplus, M. "Absolute Binding Free Energies: A Quantitative Approach for Their Calculation". In: *The Journal of Physical Chemistry B* 107.35 (Sept. 2003), pp. 9535–9551. doi: 10.1021/jp0217839.

[35] Born, M. “Volumen und Hydratationswärme der Ionen”. In: *Zeitschrift für Physik* 1.1 (Feb. 1920), pp. 45–48. doi: 10.1007/BF01881023.

[36] Born, M., Heisenberg, W., and Jordan, P. “Zur Quantenmechanik. II.” In: *Zeitschrift für Physik* 35.8-9 (Aug. 1926), pp. 557–615. doi: 10.1007/BF01379806.

[37] Born, M. and Jordan, P. “Zur Quantenmechanik”. In: *Zeitschrift für Physik* 34.1 (Dec. 1925), pp. 858–888. doi: 10.1007/BF01328531.

[38] Bourbon, H. M., Aguilera, A., Ansari, A. Z., Asturias, F. J., Berk, A. J., Bjorklund, S., Blackwell, T. K., Borggrefe, T., Carey, M., Carlson, M., Conaway, J. W., Conaway, R. C., Emmons, S. W., Fondell, J. D., Freedman, L. P., Fukasawa, T., Gustafsson, C. M., Han, M., He, X., Herman, P. K., Hinnebusch, A. G., Holmberg, S., Holstege, F. C., Jaehning, J. A., Kim, Y. J., Kuras, L., Leutz, A., Lis, J. T., Meisterernest, M., Naar, A. M., Nasmyth, K., Parvin, J. D., Ptashne, M., Reinberg, D., Ronne, H., Sadowski, I., Sakurai, H., Sipiczki, M., Sternberg, P. W., Stillman, D. J., Strich, R., Struhl, K., Svejstrup, J. Q., Tuck, S., Winston, F., Roeder, R. G., and Kornberg, R. D. “A unified nomenclature for protein subunits of mediator complexes linking transcriptional regulators to RNA polymerase II”. In: *Molecular Cell* 14.5 (2004), pp. 553–557. doi: 10.1016/j.molcel.2004.05.011.

[39] Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M., and Karplus, M. “CHARMM: The biomolecular simulation program”. In: *Journal of Computational Chemistry* 30.10 (July 2009), pp. 1545–1614. doi: 10.1002/jcc.21287.

[40] Brown, L., Wolf, J. M., Prados-Rosales, R., and Casadevall, A. “Through the wall: extracellular vesicles in Gram-positive bacteria, mycobacteria and fungi”. In: *Nature Reviews Microbiology* 13.10 (Oct. 2015), pp. 620–630. doi: 10.1038/nrmicro3480. arXiv: 15334406.

[41] Bufalo, R., Pimentel, B. M., and Zambrano, G. E. R. “Path integral quantization of generalized quantum electrodynamics”. In: *Physical Review D* 83.4 (Feb. 2011), p. 045007. doi: 10.1103/PhysRevD.83.045007. arXiv: 1008.3181.

[42] Bullock, C. W., Jacob, R. B., McDougal, O. M., Hampikian, G., and Andersen, T. “Dockomatic - automated ligand creation and docking”. In: *BMC Research Notes* 3.1 (2010), p. 289. doi: 10.1186/1756-0500-3-289.

[43] Bullock, C., Cornia, N., Jacob, R., Remm, A., Peavey, T., Weekes, K., Mallory, C., Oxford, J. T., Mcdougal, O. M., and Andersen, T. L. "DockoMatic 2 . 0 : High Throughput Inverse Virtual Screening and Homology Modeling". In: *Journal of chemical information and modeling* (2013).

[44] Bulo, R. E., Michel, C., Fleurat-Lessard, P., and Sautet, P. "Multiscale Modeling of Chemistry in Water: Are We There Yet?" In: *Journal of Chemical Theory and Computation* 9.12 (Dec. 2013), pp. 5567–5577. doi: 10.1021/ct4005596.

[45] Bulo, R. E., Ensing, B., Sikkema, J., and Visscher, L. "Toward a Practical Method for Adaptive QM/MM Simulations". In: *Journal of Chemical Theory and Computation* 5.9 (Sept. 2009), pp. 2212–2221. doi: 10.1021/ct900148e.

[46] Cao, J. and Voth, G. a. "The formulation of quantum statistical mechanics based on the Feynman path centroid density. I. Equilibrium properties". In: *The Journal of Chemical Physics* 100.7 (1994), p. 5093. doi: 10.1063/1.467175.

[47] Cavill, S. *Number of atoms in the universe.* url: https://educationblog.oup.com/secondary/maths/numbers-of-atoms-in-the-universe [Accessed: 2018-02-12]. Oxford University Press, 2015.

[48] Ceriotti, M., Bussi, G., and Parrinello, M. "Nuclear Quantum Effects in Solids Using a Colored-Noise Thermostat". In: *Physical Review Letters* 103.3 (July 2009), p. 030603. doi: 10.1103/PhysRevLett.103.030603.

[49] Ceriotti, M., Manolopoulos, D. E., and Parrinello, M. "Accelerating the convergence of path integral dynamics with a generalized Langevin equation". In: *Journal of Chemical Physics* 134.8 (2011). doi: 10.1063/1.3556661. arXiv: arXiv:1202.4093v1.

[50] Ceriotti, M., More, J., and Manolopoulos, D. E. "I-PI: A Python interface for ab initio path integral molecular dynamics simulations". In: *Computer Physics Communications* 185.3 (2014), pp. 1019–1026. doi: 10.1016/j.cpc.2013.10.027. arXiv: 1402.1045.

[51] Chandler, D. and Wolynes, P. G. "Exploiting the isomorphism between quantum theory and classical statistical mechanics of polyatomic fluids". In: *The Journal of Chemical Physics* 74.7 (Apr. 1981), pp. 4078–4095. doi: 10.1063/1.441588.

[52] Chaskar, P., Zoete, V., and Ro, U. F. "Toward On-The-Fly Quantum Mechanical / Molecular Mechanical (QM / MM) Docking : Development and Benchmark of a Scoring Function". In: *J. Chem. Inf. Model.* (2014).

[53] Chaskar, P., Zoete, V., and Röhrig, U. F. "Toward on-the-fly quantum mechanical/molecular mechanical (QM/MM) docking: Development and benchmark of a scoring function". In: *Journal of Chemical Information and Modeling* 54.11 (2014), pp. 3137–3152. doi: 10.1021/ci5004152.

[54] Chen, H.-M., Liu, B.-F., Huang, H.-L., Hwang, S.-F., and Ho, S.-Y. "SODOCK: Swarm optimization for highly flexible protein–ligand docking". In: *Journal of Computational Chemistry* 28.2 (Jan. 2007), pp. 612–623. doi: 10.1002/jcc.20542.

[55] Chen, L., Todd, R., Kiehlbauch, J., Walters, M., and Kallen, A. "Notes from the Field : Pan-Resistant New Delhi Metallo-Beta-Lactamase-Producing Klebsiella pneumoniae — Washoe County, Nevada, 2016". In: *MMWR. Morbidity and Mortality Weekly Report* 66.1 (Jan. 2017), p. 33. doi: 10.15585/mmwr.mm6601a7.

[56] Chen, Y.-N. P., LaMarche, M. J., Chan, H. M., Fekkes, P., Garcia-Fortanet, J., Acker, M. G., Antonakos, B., Chen, C. H.-T., Chen, Z., Cooke, V. G., Dobson, J. R., Deng, Z., Fei, F., Firestone, B., Fodor, M., Fridrich, C., Gao, H., Grunenfelder, D., Hao, H.-X., Jacob, J., Ho, S., Hsiao, K., Kang, Z. B., Karki, R., Kato, M., Larrow, J., La Bonte, L. R., Lenoir, F., Liu, G., Liu, S., Majumdar, D., Meyer, M. J., Palermo, M., Perez, L., Pu, M., Price, E., Quinn, C., Shakya, S., Shultz, M. D., Slisz, J., Venkatesan, K., Wang, P., Warmuth, M., Williams, S., Yang, G., Yuan, J., Zhang, J.-H., Zhu, P., Ramsey, T., Keen, N. J., Sellers, W. R., Stams, T., and Fortin, P. D. "Allosteric inhibition of SHP2 phosphatase inhibits cancers driven by receptor tyrosine kinases." In: *Nature* 535.7610 (2016), pp. 148–52. doi: 10.1038/nature18621. arXiv: arXiv:1507.02142v2.

[57] Chipot, C. "Frontiers in free-energy calculations of biological systems". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 4.1 (2014), pp. 71–89. doi: 10.1002/wcms.1157.

[58] Chipot, C. and Pohorille, A. *Free Energy Calculations*. Vol. 86. Springer, 2007. doi: 10.1007/978-3-540-38448-9.

[59] Cho, A. E. and Rinaldo, D. "Extension of QM/MM docking and its applications to metalloproteins". In: *Journal of Computational Chemistry* 30.16 (Dec. 2009), pp. 2609–2616. doi: 10.1002/jcc.21270.

[60] Cierpicki, T. and Grembecka, J. "Targeting protein-protein interactions in hematologic malignancies: still a challenge or a great opportunity for future therapies?" In: *Immunological Reviews* 263.1 (Jan. 2015), pp. 279–301. doi: 10.1111/imr.12244. arXiv: 15334406.

[61] Ciulli, A. and Abell, C. "Fragment-based approaches to enzyme inhibition". In: *Current Opinion in Biotechnology* 18.6 (Dec. 2007), pp. 489–496. doi: 10.1016/j.copbio.2007.09.003.

[62] Claussen, H., Buning, C., Rarey, M., and Lengauer, T. "FlexE: efficient molecular docking considering protein structure variations." In: *Journal of molecular biology* 308.2 (2001), pp. 377–95. doi: 10.1006/jmbi.2001.4551.

[63] Cohen, S. "A strategy for the chemotherapy of infectious disease". In: *Science* 197.4302 (July 1977), pp. 431–432. doi: 10.1126/science.195340.

[64] Collignon, B., Schulz, R., Smith, J. C., and Baudry, J. "Task-parallel message passing interface implementation of Autodock4 for docking of very large databases of compounds using high-performance super-computers". In: *Journal of Computational Chemistry* 32.6 (Apr. 2011), pp. 1202–1209. DOI: 10.1002/jcc.21696.

[65] Corbeil, C. R., Englebienne, P., Yannopoulos, C. G., Chan, L., Das, S. K., Bilimoria, D., L'Heureux, L., and Moitessier, N. "Docking Ligands into Flexible and Solvated Macromolecules. 2. Development and Application of Fitted 1.5 to the Virtual Screening of Potential HCV Polymerase Inhibitors". In: *Journal of Chemical Information and Modeling* 48.4 (Apr. 2008), pp. 902–909. DOI: 10.1021/ci700398h.

[66] Craig, I. R. and Manolopoulos, D. E. "Quantum statistics and classical mechanics: Real time correlation functions from ring polymer molecular dynamics". In: *The Journal of Chemical Physics* 121.8 (Aug. 2004), pp. 3368–3373. DOI: 10.1063/1.1777575.

[67] Cramer, C. J. *Essentials of Computational Chemistry Theories and Models.* 2nd ed. Vol. 42. 2. 2004, pp. 334–342. DOI: 10.1021/ci010445m.

[68] Daniell, P. J. "Integrals in An Infinite Number of Dimensions". In: *The Annals of Mathematics* 20.4 (July 1919), p. 281. DOI: 10.2307/1967122.

[69] Di Muzio, E., Toti, D., and Polticelli, F. "DockingApp: a user friendly interface for facilitated docking simulations with AutoDock Vina". In: *Journal of Computer-Aided Molecular Design* 31.2 (Feb. 2017), pp. 213–218. DOI: 10.1007/s10822-016-0006-1.

[70] DiMasi, J. A., Grabowski, H. G., and Hansen, R. W. "Innovation in the pharmaceutical industry: New estimates of R&D costs". In: *Journal of Health Economics* 47 (May 2016), pp. 20–33. DOI: 10.1016/j.jhealeco.2016.01.012.

[71] Ding, Y., Fang, Y., Feinstein, W. P., Ramanujam, J., Koppelman, D. M., Moreno, J., Brylinski, M., and Jarrell, M. "GeauxDock: A novel approach for mixed-resolution ligand docking using a descriptor-based force field". In: *Journal of Computational Chemistry* 36.27 (2015), pp. 2013–2026. DOI: 10.1002/jcc.24031.

[72] Dinos, G. P. "The Macrolide Antibiotic Renaissance". In: *British Journal of Pharmacology* 174.18 (Sept. 2017), pp. 2967–2983. DOI: 10.1111/bph.13936.

[73] Dirac, P. A. M. "A new notation for quantum mechanics". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 35.03 (July 1939), p. 416. DOI: 10.1017/S0305004100021162.

[74] Dirac, P. A. M. "On the Theory of Quantum Mechanics". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 112.762 (Oct. 1926), pp. 661–677. DOI: 10.1098/rspa.1926.0133.

[75] Dirac, P. A. M. “The Lagrangian in Quantum Mechanics”. In: *Feynman’s Thesis — A New Approach to Quantum Theory*. World Scientific, Aug. 2005, pp. 111–119. doi: 10.1142/9789812567635_0003.

[76] Drago, A. “A dozen formulations of quantum mechanics : a mutual comparison according to several criteria”. In: *Atti del XXXIV Congresso SISFA*. 2014.

[77] Drago, A. and Federico, N. “The three formulations of quantum mechanics founded on the alternative choices”. In: *Atti del XXXV Convegno annuale SISFA*. 2015.

[78] Durrant, J. D. and McCammon, J. A. “NNScore 2.0: A Neural-Network Receptor–Ligand Scoring Function”. In: *Journal of Chemical Information and Modeling* 51.11 (Nov. 2011), pp. 2897–2903. doi: 10.1021/ci2003889.

[79] Edwards, W. “The theory of decision making”. In: *Psychological Bulletin* 51.4 (1954), pp. 380–417. doi: 10.1037/h0053870.

[80] Egan, A. J. F. “Bacterial outer membrane constriction”. In: *Molecular Microbiology* 107.6 (Mar. 2018), pp. 676–687. doi: 10.1111/mmi.13908.

[81] Einstein, A. “Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen”. In: *Annalen der Physik* 322.8 (1905), pp. 549–560. doi: 10.1002/andp.19053220806.

[82] Ellingson, S. R., Smith, J. C., and Baudry, J. “VinaMPI: Facilitating multiple receptor high-throughput virtual docking on high-performance computers”. In: *Journal of Computational Chemistry* 34.25 (2013), pp. 2212–2221. doi: 10.1002/jcc.23367.

[83] European Medicines Agency. *List of substances and products subject to worksharing for signal management*. url: http://www.ema.europa.eu/ema/pages/includes/document/open_document.jsp?webContentId=WC500226389 [Accessed: 2018-04-03]. 2018.

[84] Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D. M., Forman, D., and Bray, F. “Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012”. In: *International Journal of Cancer* 136.5 (Mar. 2015), E359–E386. doi: 10.1002/ijc.29210.

[85] Feynman, R. P. *Space-time approach to non-relativistic quantum mechanics*. 1948. doi: 10.1103/RevModPhys.20.367.

[86] Feynman, R. P. and Kleinert, H. “Effective classical partition functions”. In: *Physical Review A* 34.6 (1986), pp. 5080–5084. doi: 10.1103/PhysRevA.34.5080.

[87] Feynman, R. P. “Simulating physics with computers”. In: *International Journal of Theoretical Physics* 21.6-7 (1982), pp. 467–488. doi: 10.1007/BF02650179. arXiv: 9508027 [quant-ph].

[88] Feynman, R. P. and Hibbs, A. R. *Quantum Mechanics and Path Integrals.* Dover, 1965.

[89] Feynman, R. P. *Feynman's Thesis - A New Approach to Quantum Theory.* Ed. by Brown, L. M. World Scientific Publishing Co. Pte. Ltd., 2005, pp. 1–119. doi: 10.1142/9789812567635.

[90] Fodor, M., Price, E., Wang, P., Lu, H., Argintaru, A., Chen, Z., Glick, M., Hao, H.-X., Kato, M., Koenig, R., LaRochelle, J. R., Liu, G., McNeill, E., Majumdar, D., Nishiguchi, G. A., Perez, L. B., Paris, G., Quinn, C. M., Ramsey, T., Sendzik, M., Shultz, M. D., Williams, S. L., Stams, T., Blacklow, S. C., Acker, M. G., and LaMarche, M. J. "Dual Allosteric Inhibition of SHP2 Phosphatase". In: *ACS Chemical Biology* (Jan. 2018), acschembio.7b00980. doi: 10.1021/acschembio.7b00980.

[91] Frenkel, D. and Smit, B. "Understanding Molecular Simulation: From Algorithms to Applications". In: (2002), p. 664. doi: 10.1063/1.881812.

[92] Gale, J. D. and Rohl, a. L. "The General Utility Lattice Program (GULP)". In: *Molecular Simulation* 29.5 (2003), pp. 291–341. doi: 10.1080/0892702031000104887.

[93] Gaus, M., Cui, Q., and Elstner, M. "DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB)". In: *Journal of Chemical Theory and Computation* 7.4 (Apr. 2011), pp. 931–948. doi: 10.1021/ct100684s.

[94] Gaussian Inc. *ONIOM.* url: http://gaussian.com/oniom/ [Accessed: 2018-02-12].

[95] Genheden, S., Ryde, U., and Söderhjelm, P. "Binding affinities by alchemical perturbation using QM/MM with a large QM system and polarizable MM model". In: *Journal of Computational Chemistry* 36.28 (2015), pp. 2114–2124. doi: 10.1002/jcc.24048.

[96] Gibbons, G. W. and Hawking, S. W. "Action integrals and partition functions in quantum gravity". In: *Physical Review D* 15.10 (May 1977), pp. 2752–2756. doi: 10.1103/PhysRevD.15.2752.

[97] Gillan, M. J. "Quantum-classical crossover of the transition rate in the damped double well". In: *Journal of Physics C: Solid State Physics* 20.24 (Aug. 1987), pp. 3621–3641. doi: 10.1088/0022-3719/20/24/005.

[98] Gilson, M., Given, J., Bush, B., and McCammon, J. "The statistical-thermodynamic basis for computation of binding affinities: a critical review". In: *Biophysical Journal* 72.3 (Mar. 1997), pp. 1047–1069. doi: 10.1016/S0006-3495(97)78756-3.

[99] Giordanetto, F. and Kihlberg, J. "Macrocyclic Drugs and Clinical Candidates: What Can Medicinal Chemists Learn from Their Properties?" In: *Journal of Medicinal Chemistry* 57.2 (Jan. 2014), pp. 278–295. doi: 10.1021/jm400887j.

[100] Girard, J. and Lafontan, M. "Impact of visceral adipose tissue on liver metabolism and insulin resistance. Part II: Visceral adipose tissue production and liver metabolism". In: *Diabetes and Metabolism* 34.5 (2008), pp. 439–445. doi: 10.1016/j.diabet.2008.04.002.

[101] Glimm, J. and Jaffe, A. *Quantum Physics*. New York, NY: Springer New York, 1987. doi: 10.1007/978-1-4612-4728-9.

[102] Golze, D., Iannuzzi, M., Nguyen, M.-T., Passerone, D., and Hutter, J. "Simulation of Adsorption Processes at Metallic Interfaces: An Image Charge Augmented QM/MM Approach". In: *Journal of Chemical Theory and Computation* 9.11 (Nov. 2013), pp. 5086–5097. doi: 10.1021/ct400698y.

[103] Götz, A. W., Clark, M. A., and Walker, R. C. "An extensible interface for QM/MM molecular dynamics simulations with AMBER". In: *Journal of Computational Chemistry* 35.2 (Jan. 2014), pp. 95–108. doi: 10.1002/jcc.23444.

[104] Groenewold, H. J. "On the principles of elementary quantum mechanics". In: *Physica* 12.7 (Oct. 1946), pp. 405–460. doi: 10.1016/S0031-8914(46)80059-4.

[105] Gulbahar, B. "Quantum Path Computing". In: *arXiv* (2017), pp. 1–40. arXiv: 1709.00735.

[106] Gulshan, K. and Moye-Rowley, W. S. "Multidrug Resistance in Fungi". In: *Eukaryotic Cell* 6.11 (Nov. 2007), pp. 1933–1942. doi: 10.1128/EC.00254-07.

[107] Hajduk, P. J. and Sauer, D. R. "Statistical Analysis of the Effects of Common Chemical Substituents on Ligand Potency". In: *Journal of Medicinal Chemistry* 51.3 (Feb. 2008), pp. 553–564. doi: 10.1021/jm070838y.

[108] Hall, B. C. *Quantum Theory for Mathematicians*. Vol. 267. Graduate Texts in Mathematics. New York, NY: Springer New York, 2013, pp. 1–566. doi: 10.1007/978-1-4614-7116-5.

[109] Hall, R. W. and Berne, B. J. "Nonergodicity in path integral molecular dynamics". In: *The Journal of Chemical Physics* 81.8 (Oct. 1984), pp. 3641–3643. doi: 10.1063/1.448112.

[110] Hamblin, M. R. "Antimicrobial photodynamic inactivation: a bright new technique to kill resistant microbes". In: *Current Opinion in Microbiology* 33 (Oct. 2016), pp. 67–73. doi: 10.1016/j.mib.2016.06.008.

[111] Hanahan, D. and Weinberg, R. A. "The Hallmarks of Cancer". In: *Cell* 100.1 (Jan. 2000), pp. 57–70. doi: 10.1016/S0092-8674(00)81683-9.

[112] Handoko, S. D., Ouyang, X., Su, C. T. T., Kwoh, C. K., and Ong, Y. S. "QuickVina: Accelerating AutoDock Vina using gradient-based heuristics for global optimization". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 9.5 (2012), pp. 1266–1272. doi: 10.1109/TCBB.2012.82.

[113] Hartle, J. B. and Hawking, S. W. “Path-integral derivation of black-hole radiance”. In: *Physical Review D* 13.8 (Apr. 1976), pp. 2188–2203. doi: 10.1103/PhysRevD.13.2188.

[114] Harvey, M. J., Giupponi, G., and Fabritiis, G. D. “ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale”. In: *Journal of Chemical Theory and Computation* 5.6 (June 2009), pp. 1632–1639. doi: 10.1021/ct9000685. arXiv: 0902.0827.

[115] Hawking, S. W. “Quantum gravity and path integrals”. In: *Physical Review D* 18.6 (Sept. 1978), pp. 1747–1753. doi: 10.1103/PhysRevD.18.1747.

[116] Hayik, S. A., Dunbrack, R., and Merz, K. M. J. “Mixed Quantum Mechanics/Molecular Mechanics Scoring Function To Predict Protein-Ligand Binding Affinity”. In: *Journal of Chemical Theory and Computation* 6.10 (Oct. 2010), pp. 3079–3091. doi: 10.1021/ct100315g.

[117] Heisenberg, W. “Über quantentheoretische Umdeutung kinematischer und mechanischer Beziehungen.” In: *Zeitschrift für Physik* 33.1 (Dec. 1925), pp. 879–893. doi: 10.1007/BF01328377.

[118] Hermans, J. and Shankar, S. “The Free Energy of Xenon Binding to Myoglobin from Molecular Dynamics Simulation”. In: *Israel Journal of Chemistry* 27.2 (1986), pp. 225–227. doi: 10.1002/ijch.198600032.

[119] Heyden, A., Lin, H., and Truhlar, D. G. “Adaptive partitioning in combined quantum mechanical and molecular mechanical calculations of potential energy functions for multiscale simulations”. In: *J. Phys. Chem. B* 111.9 (2007), pp. 2231–2241. doi: 10.1021/jp0673617.

[120] Holden, H., Karlsen, K., Lie, K.-A., and Risebro, N. H. *Splitting Methods for Partial Differential Equations with Rough Solutions*. Zuerich, Switzerland: European Mathematical Society Publishing House, Apr. 2010. doi: 10.4171/078.

[121] Homeyer, N. and Gohlke, H. “FEW: A workflow tool for free energy calculations of ligand binding”. In: *Journal of Computational Chemistry* 34.11 (2013), pp. 965–973. doi: 10.1002/jcc.23218.

[122] Hu, H. and Yang, W. “Free energies of chemical reactions in solution and in enzymes with ab initio quantum mechanics/molecular mechanics methods.” In: *Annu. Rev. Phys. Chem.* 59 (2008), pp. 573–601. doi: 10.1146/annurev.physchem.59.032607.093618.

[123] Hu, H. and Yang, W. “Development and application of ab initio QM/MM methods for mechanistic simulation of reactions in solution and in enzymes”. In: *Journal of Molecular Structure: THEOCHEM* 898.1-3 (Mar. 2009), pp. 17–30. doi: 10.1016/j.theochem.2008.12.025.

[124] Hu, H. and Yang, W. “Dual-topology/dual-coordinate free-energy simulation using QM/MM force field”. In: *Journal of Chemical Physics* 123.4 (2005). doi: 10.1063/1.1990113.

[125] Huber, L., Grabowski, B., Militzer, M., Neugebauer, J., and Rottler, J. "A QM/MM approach for low-symmetry defects in metals". In: *Computational Materials Science* 118 (June 2016), pp. 259–268. doi: 10.1016/j.commatsci.2016.03.028.

[126] Humbel, S., Sieber, S., and Morokuma, K. "The IMOMO method: Integration of different levels of molecular orbital approximations for geometry optimization of large systems: Test for n-butane conformation and SN2 reaction: RCl+Cl-". In: *The Journal of Chemical Physics* 105.5 (1996), p. 1959. doi: 10.1063/1.472065.

[127] Humphrey, W., Dalke, A., and Schulten, K. "VMD: Visual molecular dynamics". In: *Journal of Molecular Graphics* 14.1 (Feb. 1996), pp. 33–38. doi: 10.1016/0263-7855(96)00018-5. arXiv: arXiv:1503.05249v1.

[128] Humphreys, D. D., Friesner, R. A., and Berne, B. J. "A multiple-time-step Molecular Dynamics algorithm for macromolecules". In: *Journal of Physical Chemistry* 98.27 (1994), pp. 6885–6892. doi: 10.1021/j100078a035.

[129] Imhof, P. "Computational Study of Absorption Spectra of the Photoconvertible Fluorescent Protein EosFP in Different Protonation States". In: *Journal of Chemical Theory and Computation* 8.11 (Nov. 2012), pp. 4828–4836. doi: 10.1021/ct300706r.

[130] Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., and Coleman, R. G. "ZINC: A free tool to discover chemistry for biology". In: *Journal of Chemical Information and Modeling* 52.7 (2012), pp. 1757–1768. doi: 10.1021/ci3001277.

[131] Isborn, C. M., Götz, A. W., Clark, M. A., Walker, R. C., and Martínez, T. J. "Electronic Absorption Spectra from MM and ab Initio QM/MM Molecular Dynamics: Environmental Effects on the Absorption Spectrum of Photoactive Yellow Protein". In: *Journal of Chemical Theory and Computation* 8.12 (Dec. 2012), pp. 5092–5106. doi: 10.1021/ct3006826.

[132] Jeon, T.-I. and Osborne, T. F. "SREBPs: metabolic integrators in physiology and metabolism". In: *Trends in Endocrinology & Metabolism* 23.2 (Feb. 2012), pp. 65–72. doi: 10.1016/j.tem.2011.10.004.

[133] Jiang, X., Kumar, K., Hu, X., Wallqvist, A., and Reifman, J. "DOVIS 2.0: an efficient and easy to use parallel virtual screening tool based on AutoDock 4.0". In: *Chemistry Central Journal* 2.1 (2008), p. 18. doi: 10.1186/1752-153X-2-18.

[134] John, C., Spura, T., Habershon, S., and Kühne, T. D. "Quantum ring-polymer contraction method: Including nuclear quantum effects at no additional computational cost in comparison to ab initio molecular dynamics". In: *Physical Review E* 93.4 (Apr. 2016), p. 043305. doi: 10.1103/PhysRevE.93.043305. arXiv: 1512.08206.

[135] Jorgensen, W. L., Buckner, J. K., Boudon, S., and Tirado-Rives, J. "Efficient computation of absolute free energies of binding by computer simulations. Application to the methane dimer in water". In: *The Journal of Chemical Physics* 89.1988 (1988), p. 3742. doi: doi:10.1063/1.454895.

[136] K. Sharma, S., Dai, T., B. Kharkwal, G., Huang, Y.-Y., Huang, L., J. Bil De Arce, V., P. Tegos, G., and R. Hamblin, M. "Drug Discovery of Antimicrobial Photosensitizers Using Animal Models". In: *Current Pharmaceutical Design* 17.13 (May 2011), pp. 1303–1319. doi: 10.2174/138161211795703735. arXiv: NIHMS150003.

[137] Kac, M. "On distributions of certain Wiener functionals". In: *Transactions of the American Mathematical Society* 65.1 (Jan. 1949), pp. 1–1. doi: 10.1090/S0002-9947-1949-0027960-X.

[138] Kanaan, N., Crehuet, R., and Imhof, P. "Mechanism of the Glycosidic Bond Cleavage of Mismatched Thymine in Human Thymine DNA Glycosylase Revealed by Classical Molecular Dynamics and Quantum Mechanical/-Molecular Mechanical Calculations". In: *The Journal of Physical Chemistry B* 119.38 (Sept. 2015), pp. 12365–12380. doi: 10.1021/acs.jpcb.5b05496.

[139] Kapil, V., VandeVondele, J., and Ceriotti, M. "Accurate molecular dynamics and nuclear quantum effects at low cost by multiple steps in real and imaginary time: Using density functional theory to accelerate wavefunction methods". In: *The Journal of Chemical Physics* 144.5 (Feb. 2016), p. 054111. doi: dl. arXiv: 1512.00176.

[140] Kashef, N., Huang, Y.-Y., and Hamblin, M. R. "Advances in antimicrobial photodynamic inactivation at the nanoscale". In: *Nanophotonics* 6.5 (Jan. 2017), pp. 853–879. doi: 10.1515/nanoph-2016-0189.

[141] Kato, T. "On the Trotter-Lie product formula". In: *Proceedings of the Japan Academy* 50.9 (1974), pp. 694–698. doi: 10.3792/pja/1195518790.

[142] Kato, T. "Trotter's Product Formula for Some Nonlinear Semigroups". In: *Proceedings of the Symposium on Nonlinear Evolution Equations* (1978).

[143] Kerdcharoen, T., Liedl, K. R., and Rode, B. M. "A QM/MM simulation method applied to the solution of Li+ in liquid ammonia". In: *Chemical Physics* 211.1-3 (Nov. 1996), pp. 313–323. doi: 10.1016/0301-0104(96)00152-8.

[144] Kerdcharoen, T. and Morokuma, K. "ONIOM-XS: An extension of the ONIOM method for molecular simulation in condensed phase". In: *Chemical Physics Letters* 355.3-4 (2002), pp. 257–262. doi: 10.1016/S0009-2614(02)00210-5.

[145] Kheiripour, N., Karimi, J., Khodadadi, I., Tavilani, H., Goodarzi, M. T., and Hashemnia, M. "Silymarin prevents lipid accumulation in the liver of rats with type 2 diabetes via sirtuin1 and SREBP-1c". In: *Journal of Basic and Clinical Physiology and Pharmacology* (Feb. 2018), pp. 1–8. doi: 10.1515/jbcpp-2017-0122.

[146] Kirkpatrick, P. and Ellis, C. "Chemical space". In: *Nature* 432.7019 (Dec. 2004), pp. 823–823. doi: 10.1038/432823a.

[147] Kirkwood, J. G. "Statistical mechanics of fluid mixtures". In: *Journal Of Chemical Physics* 3.1935 (1935), pp. 300–313. doi: 10.1063/1.1749657.

[148] Kleinert, H. *Path Integrals in Quantum Mechanics, Statistics, Polymer Physics, and Financial Markets*. 5th. World Scientific Publishing Co. Pte. Ltd., 2009, p. 1624. doi: 10.1142/9789814273572.

[149] Ko, H. R., Kim, C. K., and Ahn, J.-Y. "Phosphorylation of the N-terminal domain of p48 Ebp1 by CDK2 is required for tumorigenic function of p48". In: *Molecular Carcinogenesis* 54.11 (Nov. 2015), pp. 1283–1291. doi: 10.1002/mc.22203.

[150] Koes, D. R., Baumgartner, M. P., and Camacho, C. J. "Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise". In: *Journal of Chemical Information and Modeling* 53.8 (2013), pp. 1893–1904. doi: 10.1021/ci300604z. arXiv: NIHMS150003.

[151] König, G., Hudson, P. S., Boresch, S., and Woodcock, H. L. "Multi-scale free energy simulations: An efficient method for connecting classical MD simulations to QM or QM/MM free energies using non-Boltzmann Bennett reweighting schemes". In: *Journal of Chemical Theory and Computation* 10.4 (2014), pp. 1406–1419. doi: 10.1021/ct401118k.

[152] König, G., Pickard IV, F. C., Mei, Y., and Brooks, B. R. "Predicting hydration free energies with a hybrid QM/MM approach: An evaluation of implicit and explicit solvation models in SAMPL4". In: *Journal of Computer-Aided Molecular Design* 28.3 (2014), pp. 245–257. doi: 10.1007/s10822-014-9708-4.

[153] König, G., Pickard, F. C., Huang, J., Simmonett, A. C., Tofoleanu, F., Lee, J., Dral, P. O., Prasad, S., Jones, M., Shao, Y., Thiel, W., and Brooks, B. R. "Calculating distribution coefficients based on multi-scale free energy simulations: an evaluation of MM and QM/MM explicit solvent simulations of water-cyclohexane transfer in the SAMPL5 challenge". In: *Journal of Computer-Aided Molecular Design* 30.11 (Nov. 2016), pp. 989–1006. doi: 10.1007/s10822-016-9936-x.

[154] Kowalinski, E., Bange, G., Bradatsch, B., Hurt, E., Wild, K., and Sinning, I. "The crystal structure of Ebp1 reveals a methionine aminopeptidase fold as binding platform for multiple interactions". In: *FEBS Letters* 581.23 (2007), pp. 4450–4454. doi: 10.1016/j.febslet.2007.08.024.

[155] Kratz, E. G., Walker, A. R., Lagardère, L., Lipparini, F., Piquemal, J.-P., and Andrés Cisneros, G. "LICHEM: A QM/MM program for simulations with multipolar and polarizable force fields". In: *Journal of Computational Chemistry* 37.11 (Apr. 2016), pp. 1019–1029. doi: 10.1002/jcc.24295.

[156] Krishnan, R. and Pople, J. A. "Approximate fourth-order perturbation theory of the electron correlation energy". In: *International Journal of Quantum Chemistry* 14.1 (July 1978), pp. 91–100. doi: 10.1002/qua.560140109.

[157] Krycer, J. R., Sharpe, L. J., Luu, W., and Brown, A. J. "The Akt–SREBP nexus: cell signaling meets lipid metabolism". In: *Trends in Endocrinology & Metabolism* 21.5 (May 2010), pp. 268–276. doi: 10.1016/j.tem.2010.01.001.

[158] Kubař, T., Welke, K., and Groenhof, G. "New QM/MM implementation of the DFTB3 method in the gromacs package". In: *Journal of Computational Chemistry* 36.26 (Oct. 2015), pp. 1978–1989. doi: 10.1002/jcc.24029.

[159] Kuhajda, F. P. "Fatty-acid synthase and human cancer: new perspectives on its role in tumor biology". In: *Nutrition* 16.3 (Mar. 2000), pp. 202–208. doi: 10.1016/S0899-9007(99)00266-X.

[160] Kuhn, M., Mering, C. von, Campillos, M., Jensen, L. J., and Bork, P. "STITCH: interaction networks of chemicals and proteins". In: *Nucleic Acids Research* 36.Database (Dec. 2007), pp. D684–D688. doi: 10.1093/nar/gkm795.

[161] Kuhn, M., Szklarczyk, D., Franceschini, A., Campillos, M., Mering, C. von, Jensen, L. J., Beyer, A., and Bork, P. "STITCH 2: an interaction network database for small molecules and proteins". In: *Nucleic Acids Research* 38.suppl_1 (Jan. 2010), pp. D552–D556. doi: 10.1093/nar/gkp937.

[162] Kuhn, M., Szklarczyk, D., Franceschini, A., Mering, C. von, Jensen, L. J., and Bork, P. "STITCH 3: zooming in on protein-chemical interactions". In: *Nucleic Acids Research* 40.D1 (Jan. 2012), pp. D876–D880. doi: 10.1093/nar/gkr1011.

[163] Kuhn, M., Szklarczyk, D., Pletscher-Frankild, S., Blicher, T. H., Mering, C. von, Jensen, L. J., and Bork, P. "STITCH 4: integration of protein–chemical interactions with user data". In: *Nucleic Acids Research* 42.D1 (Jan. 2014), pp. D401–D407. doi: 10.1093/nar/gkt1207.

[164] Lai, C.-C., Lee, K., Xiao, Y., Ahmad, N., Veeraraghavan, B., Thamlikitkul, V., Tambyah, P. A., Nelwan, R., Shibl, A. M., Wu, J.-J., Seto, W.-H., and Hsueh, P.-R. "High burden of antimicrobial drug resistance in Asia". In: *Journal of Global Antimicrobial Resistance* 2.3 (Sept. 2014), pp. 141–147. doi: 10.1016/j.jgar.2014.02.007.

[165] Laino, T., Mohamed, F., Laio, A., and Parrinello, M. "An efficient linear-scaling electrostatic coupling for treating periodic boundary conditions in QM/MM simulations". In: *Journal of Chemical Theory and Computation* 2.5 (2006), pp. 1370–1378. doi: 10.1021/ct6001169.

[166] Laino, T., Mohamed, F., Laio, A., and Parrinello, M. “An efficient real space multigrid QM/MM electrostatic coupling”. In: *Journal of Chemical Theory and Computation* 1.6 (2005), pp. 1176–1184. doi: 10.1021/ct050123f.

[167] Laio, A., VandeVondele, J., and Rothlisberger, U. “A Hamiltonian electrostatic coupling scheme for hybrid Car-Parrinello molecular dynamics simulations”. In: *Journal of Chemical Physics* 116.16 (2002), pp. 6941–6947. doi: 10.1063/1.1462041.

[168] Laitinen, O. H., Hytönen, V. P., Nordlund, H. R., and Kulomaa, M. S. “Genetically engineered avidins and streptavidins”. In: *Cellular and Molecular Life Sciences* 63.24 (Dec. 2006), pp. 2992–3017. doi: 10.1007/s00018-006-6288-z.

[169] LaRochelle, J. R., Fodor, M., Xu, X., Durzynska, I., Fan, L., Stams, T., Chan, H. M., LaMarche, M. J., Chopra, R., Wang, P., Fortin, P. D., Acker, M. G., and Blacklow, S. C. “Structural and Functional Consequences of Three Cancer-Associated Mutations of the Oncogenic Phosphatase SHP2”. In: *Biochemistry* 55.15 (Apr. 2016), pp. 2269–2277. doi: 10.1021/acs.biochem.5b01287.

[170] Le Trong, I., Wang, Z., Hyre, D. E., Lybrand, T. P., Stayton, P. S., and Stenkamp, R. E. “Streptavidin and its biotin complex at atomic resolution”. In: *Acta Crystallographica Section D Biological Crystallography* 67.9 (Sept. 2011), pp. 813–821. doi: 10.1107/S0907444911027806.

[171] Leaver, M., Domínguez-Cuevas, P., Coxhead, J. M., Daniel, R. A., and Errington, J. “Life without a wall or division machine in Bacillus subtilis”. In: *Nature* 457.7231 (Feb. 2009), pp. 849–853. doi: 10.1038/nature07742.

[172] Lee, F. X. “Path Integrals in Lattice Quantum Chromodynamics”. In: *Path Integrals — New Trends and Perspectives*. World Scientific, Nov. 2008, pp. 352–358. doi: 10.1142/9789812837271_0050. arXiv: 0710.4103.

[173] Lee, H., Heo, L., Lee, M. S., and Seok, C. “GalaxyPepDock: a protein-peptide docking tool based on interaction similarity and energy optimization.” In: *Nucleic acids research* 43.May (2015), W431–5. doi: 10.1093/nar/gkv495.

[174] Lewis, C. A., Brault, C., Peck, B., Bensaad, K., Griffiths, B., Mitter, R., Chakravarty, P., East, P., Dankworth, B., Alibhai, D., Harris, A. L., and Schulze, A. “SREBP maintains lipid biosynthesis and viability of cancer cells under lipid- and oxygen-deprived conditions and defines a gene signature associated with poor survival in glioblastoma multiforme”. In: *Oncogene* 34.40 (Oct. 2015), pp. 5128–5140. doi: 10.1038/onc.2014.439.

[175] Li, G. and Cui, Q. “pKa Calculations with QM/MM Free Energy Perturbations”. In: *The Journal of Physical Chemistry B* 107.51 (Dec. 2003), pp. 14521–14528. doi: 10.1021/jp0356158.

[176] Li, G., Zhang, X., and Cui, Q. “Free Energy Perturbation Calculations with Combined QM/MM Potentials Complications, Simplifications, and Applications to Redox Potential Calculations”. In: *The Journal of Physical Chemistry B* 107.33 (2003), pp. 8643–8653. doi: 10.1021/jp034286g.

[177] Li, H., Callahan, K., Mackerell, A. D., and Roux, B. “Development of the Charmm Drude Polarizable Force Field for the Study of Ion Interactions in Biological Systems”. In: *Biophysical Journal* 104.2 (2013), 507a. doi: 10.1016/j.bpj.2012.11.2799.

[178] Lie, S. and Engel, F. *Theorie der Transformationsgruppen I.* 1888.

[179] Lindsay, J. *The origins of alchemy in Graeco-Roman Egypt.* Muller, 1970, p. 452.

[180] Liu, S., Wang, L., and Mobley, D. L. “Is Ring Breaking Feasible in Relative Binding Free Energy Calculations?” In: *Journal of Chemical Information and Modeling* 55.4 (Apr. 2015), pp. 727–735. doi: 10.1021/acs.jcim.5b00057.

[181] Liu, S., Wu, Y., Lin, T., Abel, R., Redmann, J. P., Summa, C. M., Jaber, V. R., Lim, N. M., and Mobley, D. L. “Lead optimization mapper: Automating free energy calculations for lead optimization”. In: *Journal of Computer-Aided Molecular Design* 27.9 (2013), pp. 755–770. doi: 10.1007/s10822-013-9678-y.

[182] Liu, Y., Lu, G., Chen, Z., and Kioussis, N. “An improved QM/MM approach for metals”. In: *Modelling and Simulation in Materials Science and Engineering* 15.3 (Apr. 2007), pp. 275–284. doi: 10.1088/0965-0393/15/3/006.

[183] Liu, Y., Zhao, L., Li, W., Zhao, D., Song, M., and Yang, Y. “FIPSDock: A new molecular docking technique driven by fully informed swarm optimization algorithm”. In: *Journal of Computational Chemistry* 34.1 (2013), pp. 67–75. doi: 10.1002/jcc.23108.

[184] Liu, Z., Ahn, J.-Y., Liu, X., and Ye, K. “Ebp1 isoforms distinctively regulate cell survival and differentiation.” In: *Proceedings of the National Academy of Sciences of the United States of America* 103.29 (2006), pp. 10917–22. doi: 10.1073/pnas.0602923103.

[185] Lo, M.-C., Aulabaugh, A., Jin, G., Cowling, R., Bard, J., Malamas, M., and Ellestad, G. “Evaluation of fluorescence-based thermal shift assays for hit identification in drug discovery”. In: *Analytical Biochemistry* 332.1 (Sept. 2004), pp. 153–159. doi: 10.1016/j.ab.2004.04.031.

[186] Loeffler, H. H., Michel, J., and Woods, C. “FESetup: Automating Setup for Alchemical Free Energy Simulations”. In: *Journal of Chemical Information and Modeling* 55.12 (Dec. 2015), pp. 2485–2490. doi: 10.1021/acs.jcim.5b00368.

[187] Madsen, C. M. and Clausen, M. H. “Biologically Active Macrocyclic Compounds - from Natural Products to Diversity-Oriented Synthesis”. In: *European Journal of Organic Chemistry* 2011.17 (June 2011), pp. 3107–3115. doi: 10.1002/ejoc.201001715.

[188] Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., and Simmerling, C. “ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB”. In: *Journal of Chemical Theory and Computation* 11.8 (Aug. 2015), pp. 3696–3713. doi: 10.1021/acs.jctc.5b00255. arXiv: 15334406.

[189] Markland, T. E. and Manolopoulos, D. E. “A refined ring polymer contraction scheme for systems with electrostatic interactions”. In: *Chemical Physics Letters* 464.4-6 (2008), pp. 256–261. doi: 10.1016/j.cplett.2008.09.019.

[190] Markland, T. E. and Manolopoulos, D. E. “An efficient ring polymer contraction scheme for imaginary time path integral simulations”. In: *Journal of Chemical Physics* 129.2 (2008). doi: 10.1063/1.2953308.

[191] Marsalek, O. and Markland, T. E. “Ab initio molecular dynamics with nuclear quantum effects at classical cost: Ring polymer contraction for density functional theory”. In: *Journal of Chemical Physics* 144.5 (2016). doi: 10.1063/1.4941093. arXiv: 1512.00473.

[192] Marsault, E. and Peterson, M. L. “Macrocycles Are Great Cycles: Applications, Opportunities, and Challenges of Synthetic Macrocycles in Drug Discovery”. In: *Journal of Medicinal Chemistry* 54.7 (Apr. 2011), pp. 1961–2004. doi: 10.1021/jm1012374.

[193] Martinez De Pinillos Bayona, A., Mroz, P., Thunshelle, C., and Hamblin, M. R. “Design features for optimization of tetrapyrrole macrocycles as antimicrobial and anticancer photosensitizers”. In: *Chemical Biology & Drug Design* 89.2 (Feb. 2017), pp. 192–206. doi: 10.1111/cbdd.12792.

[194] Marx, D. and Hutter, J. *Ab Initio Molecular Dynamics*. Cambridge: Cambridge University Press, 2009, p. 579. doi: 10.1017/CBO9780511609633.

[195] Maseras, F. and Morokuma, K. “IMOMM: A new integrate ab initio + molecular mechanics geometry optimization scheme of equilibrium structures and transition states”. In: *Journal of Computational Chemistry* 16.9 (Sept. 1995), pp. 1170–1179. doi: 10.1002/jcc.540160911.

[196] Mei, Y., Zhang, P., Zuo, H., Clark, D., Xia, R., Li, J., Liu, Z., and Mao, L. “Ebp1 activates podoplanin expression and contributes to oral tumorigenesis.” In: *Oncogene* February (2013), pp. 1–12. doi: 10.1038/onc.2013.354.

[197] Meroueh, S. O., Bencze, K. Z., Hesek, D., Lee, M., Fisher, J. F., Stemmler, T. L., and Mobashery, S. "Three-dimensional structure of the bacterial cell wall peptidoglycan." In: *Proceedings of the National Academy of Sciences of the United States of America* 103.12 (Mar. 2006), pp. 4404–4409. doi: 10.1073/pnas.0510182103.

[198] Merz, K. M. J., Ringe, D., and Reynolds, C. H. *Drug Design: Structure- and Ligand-Based Approaches.* 2010, p. 289. doi: 10.1017/CBO9780511730412.

[199] Metz, S., Kästner, J., Sokol, A. A., Keal, T. W., and Sherwood, P. "ChemShell - A modular software package for QM/MM simulations". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 4.2 (2014), pp. 101–110. doi: 10.1002/wcms.1163.

[200] Min, D., Chen, M., Zheng, L., Jin, Y., Schwartz, M. A., Sang, Q. X. A., and Yang, W. "Enhancing QM/MM molecular dynamics sampling in explicit environments via an orthogonal-space-random-walk-based strategy". In: *Journal of Physical Chemistry B* 115.14 (2011), pp. 3924–3935. doi: 10.1021/jp109454q.

[201] Mishra, P., Dixit, U., Pandey, A. K., Upadhyay, A., and Pandey, V. N. "Modulation of HCV replication and translation by ErbB3 binding protein1 isoforms". In: *Virology* 500.October 2016 (Jan. 2017), pp. 35–49. doi: 10.1016/j.virol.2016.10.006.

[202] Mobley, D. L., Heinzelmann, G., Henriksen, N. M., and Gilson., M. K. "Predicting binding free energies: Frontiers and benchmarks (a perpetual review)". In: (2017). doi: 10.5281/zenodo.839047.

[203] Monie, T. P., Perrin, A. J., Birtley, J. R., Sweeney, T. R., Karakasiliotis, I., Chaudhry, Y., Roberts, L. O., Matthews, S., Goodfellow, I. G., and Curry, S. "Structural insights into the transcriptional and translational roles of Ebp1." In: *The EMBO journal* 26.17 (2007), pp. 3936–44. doi: 10.1038/sj.emboj.7601817.

[204] Monticelli, L., Salonen, E., Johansson, M. P., Kaila, V. R. I., and Sundholm, D. *Biomolecular Simulations: Methods and Protocols.* Ed. by Monticelli, L. and Salonen, E. Vol. 924. Humana Press, 2013, pp. 3–27. doi: 10.1007/978-1-62703-017-5.

[205] Morrone, J. A., Markland, T. E., Ceriotti, M., and Berne, B. J. "Efficient multiple time scale molecular dynamics: Using colored noise thermostats to stabilize resonances". In: *The Journal of Chemical Physics* 134 (2011). doi: 10.1063/1.3518369.

[206] Muddana, H. S., Fenley, A. T., Mobley, D. L., and Gilson, M. K. "The SAMPL4 host-guest blind prediction challenge: An overview". In: *Journal of Computer-Aided Molecular Design* 28.4 (2014), pp. 305–317. doi: 10.1007/s10822-014-9735-1.

[207] Mukerjee, M. "Speaking for the Animals". In: *Scientific American* 291.2 (Aug. 2004), pp. 96–97. doi: 10.1038/scientificamerican0804-96.

[208] Mullard, A. “2015 FDA drug approvals”. In: *Nature Reviews Drug Discovery* 15.2 (2016), pp. 73–76. doi: 10.1038/nrd.2016.15.

[209] Munos, B. “Lessons from 60 years of pharmaceutical innovation.” In: *Nature reviews. Drug discovery* 8.12 (2009), pp. 959–968. doi: 10.1038/nrd2961.

[210] NIH Center For Macromolecular Modeling & Bioinformatics. *Hybrid QM/MM NAMD*. url: http://www.ks.uiuc.edu/~rcbernardi/QMMM/ [Accessed: 2018-02-12].

[211] Nishikawa, J. L., Boeszoermenyi, A., Vale-Silva, L. A., Torelli, R., Posteraro, B., Sohn, Y.-J., Ji, F., Gelev, V., Sanglard, D., Sanguinetti, M., Sadreyev, R. I., Mukherjee, G., Bhyravabhotla, J., Buhrlage, S. J., Gray, N. S., Wagner, G., Näär, A. M., and Arthanari, H. “Inhibiting fungal multidrug resistance by disrupting an activator–Mediator interaction”. In: *Nature* 530.7591 (Feb. 2016), pp. 485–489. doi: 10.1038/nature16963. arXiv: 15334406.

[212] O’Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. “Open Babel: An open chemical toolbox”. In: *Journal of Cheminformatics* 3.1 (2011), pp. 1–14. doi: 10.1186/1758-2946-3-33.

[213] Object Management Group (OMG). *OMG Unified Modeling Language TM (OMG UML) Superstructure - Version 2.2*. url: http://www.omg.org/spec/UML/2.2/Superstructure [Accessed: 2018-02-12]. 2009.

[214] OECD. *Health at a Glance 2017*. Health at a Glance. OECD Publishing, Nov. 2017. doi: 10.1787/health_glance-2017-en.

[215] Orlando, A. and Jorgensen, W. L. “Advances in quantum and molecular mechanical (QM/MM) simulations for organic and enzymatic reactions”. In: *Accounts of Chemical Research* 43.1 (2010), pp. 142–151. doi: 10.1021/ar900171c. arXiv: NIHMS150003.

[216] Ostermeir, K. and Zacharias, M. “Advanced replica-exchange sampling to study the flexibility and plasticity of peptides and proteins”. In: *Biochimica et Biophysica Acta - Proteins and Proteomics* 1834.5 (2013), pp. 847–853. doi: 10.1016/j.bbapap.2012.12.016.

[217] Pearlman, D. “A Comparison of Alternative Approaches to Free Energy Calculations”. In: *The Journal of Physical Chemistry* 98.5 (Feb. 1994), pp. 1487–1493. doi: 10.1021/j100056a020.

[218] Pederzoli, M., Sobek, L., Brabec, J., Kowalski, K., Cwiklik, L., and Pittner, J. “Fluorescence of PRODAN in water: A computational QM/MM MD study”. In: *Chemical Physics Letters* 597 (Mar. 2014), pp. 57–62. doi: 10.1016/j.cplett.2014.02.031.

[219] Peguiron, A., Colombi Ciacchi, L., De Vita, A., Kermode, J. R., and Moras, G. “Accuracy of buffered-force QM/MM simulations of silica”. In: *The Journal of Chemical Physics* 142.6 (Feb. 2015), p. 064116. doi: 10.1063/1.4907786.

[220] Penney, M. D., Koh, D. E., and Spekkens, R. W. “Quantum circuit dynamics via path integrals: Is there a classical action for discrete-time paths?” In: *New Journal of Physics* 19.7 (July 2017), p. 073006. doi: 10.1088/1367-2630/aa61ba. arXiv: 1604.07452.

[221] Pérez-Belmonte, L. M., Moreno-Santos, I., Cabrera-Bueno, F., Sánchez-Espín, G., Castellano, D., Such, M., Crespo-Leiro, M. G., Carrasco-Chinchilla, F., Alonso-Pulpón, L., López-Garrido, M., Ruiz-Salas, A., Becerra-Muñoz, V. M., Gómez-Doblas, J. J., Teresa-Galván, E. de, and Jiménez-Navarro, M. “Expression of Sterol Regulatory Element-Binding Proteins in epicardial adipose tissue in patients with coronary artery disease and diabetes mellitus: preliminary study”. In: *International Journal of Medical Sciences* 14.3 (2017), pp. 268–274. doi: 10.7150/ijms.17821.

[222] Pernía, J. R. J., Ruggiero, G. D., and Williams, I. H. “QM/MM kinetic isotope effects for chloromethane hydrolysis in water”. In: *Journal of Physical Organic Chemistry* 26.12 (Dec. 2013), pp. 1058–1065. doi: 10.1002/poc.3144.

[223] Persistence of Vision Raytracer Pty. Ltd. *POV-Ray (Persistence of Vision Raytracer) Homepage.* url: http://www.povray.org/ [Accessed:2018-03-16]. 2018.

[224] Pezeshki, S., Davis, C., Heyden, A., and Lin, H. “Adaptive-Partitioning QM/MM Dynamics Simulations: 3. Solvent Molecules Entering and Leaving Protein Binding Sites”. In: *Journal of Chemical Theory and Computation* 10.11 (Nov. 2014), pp. 4765–4776. doi: 10.1021/ct500553x.

[225] Pezeshki, S. and Lin, H. “Molecular dynamics simulations of ion solvation by flexible-boundary QM/MM: On-the-fly partial charge transfer between QM and MM subsystems”. In: *Journal of Computational Chemistry* 35.24 (2014), pp. 1778–1788. doi: 10.1002/jcc.23685.

[226] Piana, S., Lindorff-Larsen, K., Dirks, R. M., Salmon, J. K., Dror, R. O., and Shaw, D. E. “Evaluating the Effects of Cutoffs and Treatment of Long-range Electrostatics in Protein Folding Simulations”. In: *PLoS ONE* 7.6 (June 2012). Ed. by Verma, C., e39918. doi: 10.1371/journal.pone.0039918.

[227] Pitera, J. W. and Gunsteren, W. F. van. “A Comparison of Non-Bonded Scaling Approaches for Free Energy Calculations”. In: *Molecular Simulation* 28.1-2 (Jan. 2002), pp. 45–65. doi: 10.1080/08927020211973.

[228] Pontén, F., Jirström, K., and Uhlen, M. “The Human Protein Atlas-a tool for pathology”. In: *The Journal of Pathology* 216.4 (Dec. 2008), pp. 387–393. doi: 10.1002/path.2440.

[229] Prakhov, N. D., Chernorudskiy, A. L., and Gainullin, M. R. “VSDocker: A tool for parallel high-throughput virtual screening using AutoDock on Windows-based computer clusters”. In: *Bioinformatics* 26.10 (2010), pp. 1374–1375. doi: 10.1093/bioinformatics/btq149.

[230] Praprotnik, M., Delle Site, L., and Kremer, K. "Adaptive resolution molecular-dynamics simulation: Changing the degrees of freedom on the fly". In: *Journal of Chemical Physics* 123.22 (2005). doi: 10.1063/1.2132286. arXiv: 0510223 [cond-mat].

[231] Praprotnik, M., Delle Site, L., and Kremer, K. "Adaptive resolution scheme for efficient hybrid atomistic-mesoscale molecular dynamics simulations of dense liquids". In: *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 73.6 (2006), pp. 1–12. doi: 10.1103/PhysRevE.73.066701.

[232] Pyzer-Knapp, E. O., Suh, C., Gómez-Bombarelli, R., Aguilera-Iparraguirre, J., and Aspuru-Guzik, A. "What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery". In: *Annual Review of Materials Research* 45.1 (July 2015), pp. 195–216. doi: 10.1146/annurev-matsci-070214-020823. arXiv: 1406.3533.

[233] Quhe, R., Nava, M., Tiwary, P., and Parrinello, M. "Path integral metadynamics". In: *Journal of Chemical Theory and Computation* 11.4 (2015), pp. 1383–1388. doi: 10.1021/ct501002a.

[234] Rathore, R., Sumakanth, M., Reddy, M., Reddanna, P., Rao, A., Erion, M., and Reddy, M. "Advances in Binding Free Energies Calculations: QM/MM-Based Free Energy Perturbation Method for Drug Design". In: *Current Pharmaceutical Design* 19.26 (June 2013), pp. 4674–4686. doi: 10.2174/1381612811319260002.

[235] Ravindranath, P. A., Forli, S., Goodsell, D. S., Olson, A. J., and Sanner, M. F. "AutoDockFR: Advances in Protein-Ligand Docking with Explicitly Specified Binding Site Flexibility". In: *PLOS Computational Biology* 11.12 (Dec. 2015). Ed. by Fetrow, J. S., e1004586. doi: 10.1371/journal.pcbi.1004586.

[236] Reddy, M. R. and Erion, M. D., eds. *Free Energy Calculations in Rational Drug Design.* 2001, p. 384.

[237] Reed, M. and Simon, B. *Methods of Modern Mathematical Physics. Volume 1: Functional Analysis.* 1980.

[238] Robert Koch-Institut. *Krebs in Deutschland Häufigkeiten und Trends Gesundheitsberichterstattung des Bundes Beiträge zur Gesundheitsberichterstattung des Bundes Krebs in Deutschland Häufigkeiten und Trends.* Vol. 7. 2010, p. 121.

[239] Rode, B. M. and Hofer, T. S. "How to access structure and dynamics of solutions: The capabilities of computational methods (Special Topic Article)". In: *Pure and Applied Chemistry* 78.3 (Jan. 2006), pp. 525–539. doi: 10.1351/pac200678030525.

[240] Rottem, S. "Interaction of Mycoplasmas With Host Cells". In: *Physiological Reviews* 83.2 (Apr. 2003), pp. 417–432. doi: 10.1152/physrev.00030.2002.

[241] Roux, B., Nina, M., Pomès, R., and Smith, J. “Thermodynamic stability of water molecules in the bacteriorhodopsin proton channel: a molecular dynamics free energy perturbation study”. In: *Biophysical Journal* 71.2 (Aug. 1996), pp. 670–681. doi: 10.1016/S0006-3495(96)79267-6.

[242] Rowley, C. N. and Roux, B. “The Solvation Structure of Na + and K + in Liquid Water Determined from High Level ab Initio Molecular Dynamics Simulations”. In: *Journal of Chemical Theory and Computation* 8.10 (Oct. 2012), pp. 3526–3535. doi: 10.1021/ct300091w.

[243] Ruddigkeit, L., Van Deursen, R., Blum, L. C., and Reymond, J. L. “Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17”. In: *Journal of Chemical Information and Modeling* 52.11 (2012), pp. 2864–2875. doi: 10.1021/ci300415d.

[244] Sandeep, G., Nagasree, K. P., Hanisha, M., and Kumar, M. M. K. “AUDocker LE: A GUI for virtual screening with AUTODOCK Vina”. In: *BMC Research Notes* 4.1 (2011), p. 445. doi: 10.1186/1756-0500-4-445.

[245] Sander, T., Freyss, J., Korff, M. von, and Rufener, C. “DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis”. In: *Journal of Chemical Information and Modeling* 55.2 (Feb. 2015), pp. 460–473. doi: 10.1021/ci500588j.

[246] Schaechter, M. and Lederberg, J. *Desk Encyclopedia of Microbiology.* 2004, p. 1149.

[247] Schrödinger, LLC. *Maestro: The completely reimagined all-purpose molecular modeling environment.* url: https://www.schrodinger.com/maestro[Accessed:2018-03-16]. 2018.

[248] Schrödinger, E. “Quantisierung als Eigenwertproblem”. In: *Annalen der Physik* 384.4 (1926), pp. 361–376. doi: 10.1002/andp.19263840404. arXiv: 1.Konstante,D.{\&}Rtets,V.D.M.N(g,{\%})=.361âĂ§376.

[249] Schroedinger LLC. *A High-Performance QM/MM Program.* url: https://www.schrodinger.com/qsite [Accessed: 2018-02-12].

[250] Schulman, L. S. *Techniques and applications of path integration.* John Wiley & Sons, 2005, p. 416.

[251] Schwenk, C. F., Loeffler, H. H., and Rode, B. M. “Structure and Dynamics of Metal Ions in Solution: QM/MM Molecular Dynamics Simulations of Mn 2+ and V 2+”. In: *Journal of the American Chemical Society* 125.6 (Feb. 2003), pp. 1618–1624. doi: 10.1021/ja0286831.

[252] Semisotnov, G. V., Rodionova, N. A., Razgulyaev, O. I., Uversky, V. N., Gripas’, A. F., and Gilmanshin, R. I. “Study of the "molten globule" intermediate state in protein folding by a hydrophobic fluorescent probe”. In: *Biopolymers* 31.1 (Jan. 1991), pp. 119–128. doi: 10.1002/bip.360310111.

[253] Senn, H. M. and Thiel, W. “QM/MM methods for biological systems”. In: *Atomistic Approaches in Modern Biology: from Quantum Chem. to Molecular Simulations* 268.November 2006 (2007), pp. 173–290. doi: 10.1007/128_2006_084.

[254] Shao, W. and Espenshade, P. J. “Expanding Roles for SREBP in Metabolism”. In: *Cell Metabolism* 16.4 (Oct. 2012), pp. 414–419. doi: 10.1016/j.cmet.2012.09.002. arXiv: NIHMS150003.

[255] Shimano, H. “Sterol regulatory element-binding proteins (SREBPs): transcriptional regulators of lipid synthetic genes”. In: *Progress in Lipid Research* 40.6 (Nov. 2001), pp. 439–452. doi: 10.1016/S0163-7827(01)00010-8.

[256] Shin, W. H. and Seok, C. “GalaxyDock: Protein-ligand docking with flexible protein side-chains”. In: *Journal of Chemical Information and Modeling* 52.12 (2012), pp. 3225–3232. doi: 10.1021/ci300342z.

[257] Shin, W.-H., Christoffer, C. W., and Kihara, D. “In silico structure-based approaches to discover protein-protein interaction-targeting drugs”. In: *Methods* 131 (Dec. 2017), pp. 22–32. doi: 10.1016/j.ymeth.2017.08.006.

[258] Shirts, M. R. and Pande, V. S. “Solvation free energies of amino acid side chain analogs for common molecular mechanics water models”. In: *The Journal of Chemical Physics* 122.13 (Apr. 2005), p. 134508. doi: 10.1063/1.1877132.

[259] Shoichet, B. K. “Virtual screening of chemical libraries.” In: *Nature* 432.7019 (2004), pp. 862–865. doi: 10.1038/nature03197.

[260] Silhavy, T. J., Kahne, D., and Walker, S. “The Bacterial Cell Envelope”. In: *Cold Spring Harbor Perspectives in Biology* 2.5 (May 2010), a000414–a000414. doi: 10.1101/cshperspect.a000414.

[261] Sperandio, F., Huang, Y.-y., and Hamblin, M. “Antimicrobial Photodynamic Therapy to Kill Gram-negative Bacteria”. In: *Recent Patents on Anti-Infective Drug Discovery* 8.2 (July 2013), pp. 108–120. doi: 10.2174/1574891X113089990012.

[262] Springer, C., Adalsteinsson, H., Young, M. M., Kegelmeyer, P. W., and Roe, D. C. “PostDOCK: A structural, empirical approach to scoring protein ligand complexes”. In: *Journal of Medicinal Chemistry* 48.22 (2005), pp. 6821–6831. doi: 10.1021/jm0493360.

[263] Squatrito, M., Mancino, M., Donzelli, M., Areces, L. B., and Draetta, G. F. “EBP1 is a nucleolar growth-regulating protein that is part of pre-ribosomal ribonucleoprotein complexes”. In: *Oncogene* 23.25 (May 2004), pp. 4454–4465. doi: 10.1038/sj.onc.1207579.

[264] Stanton, R. V., Dixon, S. L., and Merz, K. M. J. “General Formulation for a Quantum Free Energy Perturbation Study”. In: *The Journal of Physical Chemistry* 99.27 (July 1995), pp. 10701–10704. doi: 10.1021/j100027a004.

[265] Statistisches Bundesamt. "Statistisches Bundesamt Gesundheit Todesursachen in Deutschland Gestorbene in Deutschland an ausgewählten Todesursachen". In: *Wirtschaft Und Statistik* 2.0 (2015).

[266] Steinbrecher, T., Mobley, D. L., and Case, D. A. "Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations". In: *The Journal of Chemical Physics* 127.21 (Dec. 2007), p. 214108. doi: 10.1063/1.2799191.

[267] Stephens, P. J., Devlin, F. J., Chabalowski, C. F., and Frisch, M. J. "Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields". In: *The Journal of Physical Chemistry* 98.45 (Nov. 1994), pp. 11623–11627. doi: 10.1021/j100096a001.

[268] Sterling, T. and Irwin, J. J. "ZINC 15 - Ligand Discovery for Everyone". In: *Journal of Chemical Information and Modeling* 55.11 (2015), pp. 2324–2337. doi: 10.1021/acs.jcim.5b00559.

[269] Stewart, J. *MOPAC Manual - QMMM*. url: http://openmopac.net/manual/QMMM.html [Accessed: 2018-02-12].

[270] Strang, G. "On the Construction and Comparison of Difference Schemes". In: *SIAM Journal on Numerical Analysis* 5.3 (Sept. 1968), pp. 506–517. doi: 10.1137/0705041.

[271] Strocchi, F. *An Introduction to the Mathematical Structure of Quantum Mechanics: A Short Course for Mathematicians*. World Scientific, 2008.

[272] Styer, D. F., Balkin, M. S., Becker, K. M., Burns, M. R., Dudley, C. E., Forth, S. T., Gaumer, J. S., Kramer, M. A., Oertel, D. C., Park, L. H., Rinkoski, M. T., Smith, C. T., and Wotherspoon, T. D. "Nine formulations of quantum mechanics". In: *American Journal of Physics* 70.3 (Mar. 2002), pp. 288–297. doi: 10.1119/1.1445404.

[273] Suchanski, J., Tejchman, A., Zacharski, M., Piotrowska, A., Grzegrzolka, J., Chodaczek, G., Nowinska, K., Rys, J., Dziegiel, P., Kieda, C., and Ugorski, M. "Podoplanin increases the migration of human fibroblasts and affects the endothelial cell network formation: A possible role for cancer-associated fibroblasts in breast cancer progression". In: *PLOS ONE* 12.9 (Sept. 2017). Ed. by Ahmad, A., e0184970. doi: 10.1371/journal.pone.0184970.

[274] Sunseri, J. and Koes, D. R. "Pharmit: interactive exploration of chemical space". In: *Nucleic Acids Research* 44.W1 (July 2016), W442–W448. doi: 10.1093/nar/gkw287.

[275] Sutcliffe, I. C. and Harrington, D. J. "Lipoproteins of Mycobacterium tuberculosis : an abundant and functionally diverse class of cell envelope components". In: *FEMS Microbiology Reviews* 28.5 (Nov. 2004), pp. 645–659. doi: 10.1016/j.femsre.2004.06.002.

[276] Svensson, M., Humbel, S., Froese, R. D. J., Matsubara, T., Sieber, S., and Morokuma, K. “ONIOM: A Multilayered Integrated MO + MM Method for Geometry Optimizations and Single Point Energy Predictions. A Test for Diels-Alder Reactions and Pt(P(t-Bu) 3) 2 + H 2 Oxidative Addition”. In: *Journal of Physical Chemistry* 100.50 (1996), pp. 19357–19363. doi: 10.1021/jp962071j.

[277] Szklarczyk, D., Santos, A., Mering, C. von, Jensen, L. J., Bork, P., and Kuhn, M. “STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data”. In: *Nucleic Acids Research* 44.D1 (Jan. 2016), pp. D380–D384. doi: 10.1093/nar/gkv1277.

[278] Tacconelli, E., Carrara, E., Savoldi, A., Harbarth, S., Mendelson, M., Monnet, D. L., Pulcini, C., Kahlmeter, G., Kluytmans, J., Carmeli, Y., Ouellette, M., Outterson, K., Patel, J., Cavaleri, M., Cox, E. M., Houchens, C. R., Grayson, M. L., Hansen, P., Singh, N., Theuretzbacher, U., Magrini, N., Aboderin, A. O., Al-Abri, S. S., Awang Jalil, N., Benzonana, N., Bhattacharya, S., Brink, A. J., Burkert, F. R., Cars, O., Cornaglia, G., Dyar, O. J., Friedrich, A. W., Gales, A. C., Gandra, S., Giske, C. G., Goff, D. A., Goossens, H., Gottlieb, T., Guzman Blanco, M., Hryniewicz, W., Kattula, D., Jinks, T., Kanj, S. S., Kerr, L., Kieny, M.-P., Kim, Y. S., Kozlov, R. S., Labarca, J., Laxminarayan, R., Leder, K., Leibovici, L., Levy-Hara, G., Littman, J., Malhotra-Kumar, S., Manchanda, V., Moja, L., Ndoye, B., Pan, A., Paterson, D. L., Paul, M., Qiu, H., Ramon-Pardo, P., Rodríguez-Baño, J., Sanguinetti, M., Sengupta, S., Sharland, M., Si-Mehand, M., Silver, L. L., Song, W., Steinbakk, M., Thomsen, J., Thwaites, G. E., Meer, J. W. van der, Van Kinh, N., Vega, S., Villegas, M. V., Wechsler-Fördös, A., Wertheim, H. F. L., Wesangula, E., Woodford, N., Yilmaz, F. O., and Zorzet, A. “Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis”. In: *The Lancet Infectious Diseases* 18.3 (Mar. 2018), pp. 318–327. doi: 10.1016/S1473-3099(17)30753-3.

[279] Takahashi, H., Omi, A., Morita, A., and Matubayasi, N. “Simple and exact approach to the electronic polarization effect on the solvation free energy: Formulation for quantum-mechanical molecular-mechanical system and its applications to aqueous solutions”. In: *Journal of Chemical Physics* 136.21 (2012). doi: 10.1063/1.4722347.

[280] Takenaka, N., Kitamura, Y., Koyano, Y., and Nagaoka, M. “The number-adaptive multiscale QM/MM molecular dynamics simulation: Application to liquid water”. In: *Chemical Physics Letters* 524 (2012), pp. 56–61. doi: 10.1016/j.cplett.2011.12.053.

[281] Takhtajan, L. A. *Quantum Mechanics for Mathematicians.* AMS, 2008, p. 387.

[282] Thakur, J. K., Yadav, A., and Yadav, G. “Molecular recognition by the KIX domain and its role in gene regulation”. In: *Nucleic Acids Research* 42.4 (Feb. 2014), pp. 2112–2125. doi: 10.1093/nar/gkt1147.

[283] The German Goverment (2015). *DART 2020.* 2015.

[284] The Nobel Foundation. *The Nobel Prize in Chemistry 2013.* url: http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2013/ [Accessed: 2018-01-18]. Nobel Media AB, 2014.

[285] The Nobel Foundation. *The Nobel Prize in Physics 1932.* url: http://www.nobelprize.org/nobel_prizes/physics/laureates/1932/ [Accessed: 2018-01-18]. Nobel Media AB, 2014.

[286] *The RDKit Documentation.* url: http://www.rdkit.org/docs/index.html [Accessed: 2018-02-12]. 2017.

[287] Thellamurege, N. M., Cui, F., and Li, H. “Quantum mechanical/molecular mechanical/continuum style solvation model: Time-dependent density functional theory”. In: *Journal of Chemical Physics* 139.8 (2013). doi: 10.1063/1.4819139.

[288] Thellamurege, N. M., Si, D., Cui, F., and Li, H. “Quantum mechanical/molecular mechanical/continuum style solvation model: Second order Møller-Plesset perturbation theory”. In: *Journal of Chemical Physics* 140.17 (2014). doi: 10.1063/1.4873344.

[289] Thellamurege, N. M., Si, D., Cui, F., Zhu, H., Lai, R., and Li, H. “Quan Pol: A full spectrum and seamless QM/MM program”. In: *Journal of Computational Chemistry* 34.32 (2013), pp. 2816–2833. doi: 10.1002/jcc.23435.

[290] Tirado-Rives, J. and Jorgensen, W. L. “Performance of B3LYP Density Functional Methods for a Large Set of Organic Molecules”. In: *Journal of Chemical Theory and Computation* 4.2 (Feb. 2008), pp. 297–306. doi: 10.1021/ct700248k.

[291] Torrie, G. M. and Valleau, J. P. “Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling”. In: *Journal of Computational Physics* 23.2 (1977), pp. 187–199. doi: 10.1016/0021-9991(77)90121-8.

[292] Tovchigrechko, A. and Vakser, I. A. “GRAMM-X public web server for protein-protein docking”. In: *Nucleic Acids Research* 34.Web Server (July 2006), W310–W314. doi: 10.1093/nar/gkl206.

[293] Tozzini, V. “Coarse-grained models for proteins”. In: *Current Opinion in Structural Biology* 15.2 (Apr. 2005), pp. 144–150. doi: 10.1016/j.sbi.2005.02.005.

[294] Trott, O. and Olson, A. J. “AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.” In: *Journal of computational chemistry* 31.2 (Jan. 2010), pp. 455–61. doi: 10.1002/jcc.21334.

[295] Trotter, H. F. "Approximation of semi-groups of operators". In: *Pacific Journal of Mathematics* 8.4 (Dec. 1958), pp. 887–919. doi: 10.2140/pjm.1958.8.887.

[296] Tuckerman, M. E. "Path Integration via Molecular Dynamics". In: *NIC Series: Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms* 10.August 2009 (2002), pp. 269–298.

[297] Tuckerman, M. E. *Statistical Mechanics: Theory and Molecular Simulation.* Oxford University Press, 2010, p. 720.

[298] U.S. Food and Drug Administration. *FDA Approved Drug Products.* url: https://www.accessdata.fda.gov/scripts/cder/daf/ [Accessed: 2018-02-12]. 2018.

[299] Ugorski, M., Dziegiel, P., and Suchanski, J. "Podoplanin - a small glycoprotein with many faces." In: *American journal of cancer research* 6.2 (2016), pp. 370–86.

[300] Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., Wernerus, H., Björling, L., and Ponten, F. "Towards a knowledge-based Human Protein Atlas". In: *Nature Biotechnology* 28.12 (Dec. 2010), pp. 1248–1250. doi: 10.1038/nbt1210-1248. arXiv: pubmed:21139605.

[301] Valiev, M., Bylaska, E., Govind, N., Kowalski, K., Straatsma, T., Van Dam, H., Wang, D., Nieplocha, J., Apra, E., Windus, T., and Jong, W. de. "NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations". In: *Computer Physics Communications* 181.9 (Sept. 2010), pp. 1477–1489. doi: 10.1016/j.cpc.2010.04.018.

[302] Van Norman, G. A. "Drugs, Devices, and the FDA: Part 1". In: *JACC: Basic to Translational Science* 1.3 (Apr. 2016), pp. 170–179. doi: 10.1016/j.jacbts.2016.03.002.

[303] Van Voorhis, T. and Head-Gordon, M. "Two-body coupled cluster expansions". In: *The Journal of Chemical Physics* 115.11 (Sept. 2001), pp. 5033–5040. doi: 10.1063/1.1390516.

[304] Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I., and Mackerell, A. D. "CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields". In: *Journal of Computational Chemistry* 31.4 (2010), pp. 671–690. doi: 10.1002/jcc.21367. arXiv: NIHMS150003.

[305] Vanommeslaeghe, K. and MacKerell, A. D. "CHARMM additive and polarizable force fields for biophysics and computer-aided drug design." In: *Biochimica et biophysica acta* 1850.5 (May 2015), pp. 861–71. doi: 10.1016/j.bbagen.2014.08.004.

[306] Vanommeslaeghe, K. and MacKerell, A. D. "Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing". In: *Journal of Chemical Information and Modeling* 52.12 (Dec. 2012), pp. 3144–3154. doi: 10.1021/ci300363c. arXiv: NIHMS150003.

[307] Vanommeslaeghe, K., Raman, E. P., and MacKerell, A. D. "Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges". In: *Journal of Chemical Information and Modeling* 52.12 (Dec. 2012), pp. 3155–3168. doi: 10.1021/ci3003649.

[308] Vergès, B. "New insight into the pathophysiology of lipid abnormalities in type 2 diabetes." In: *Diabetes & metabolism* 31.5 (2005), pp. 429–439. doi: 10.1016/S1262-3636(07)70213-6.

[309] Vollmer, W. and Bertsche, U. "Murein (peptidoglycan) structure, architecture and biosynthesis in Escherichia coli". In: *Biochimica et Biophysica Acta (BBA) - Biomembranes* 1778.9 (Sept. 2008), pp. 1714–1734. doi: 10.1016/j.bbamem.2007.06.007.

[310] Vreven, T., Byun, K. S., Komáromi, I., Dapprich, S., Montgomery, J. A., Morokuma, K., and Frisch, M. J. "Combining Quantum Mechanics Methods with Molecular Mechanics Methods in ONIOM". In: *Journal of Chemical Theory and Computation* 2.3 (May 2006), pp. 815–826. doi: 10.1021/ct050289g.

[311] Vreven, T. and Morokuma, K. "On the application of the IMOMO (integrated molecular orbital + molecular orbital) method". In: *Journal of Computational Chemistry* 21.16 (2000), pp. 1419–1432. doi: 10.1002/1096-987X(200012)21:16<1419::AID-JCC1>3.0.CO;2-C.

[312] Waller, M. P., Kumbhar, S., and Yang, J. "A density-based adaptive quantum mechanical/molecular mechanical method". In: *ChemPhysChem* 15.15 (2014), pp. 3218–3225. doi: 10.1002/cphc.201402105.

[313] Wan, S., Stote, R. H., and Karplus, M. "Calculation of the aqueous solvation energy and entropy, as well as free energy, of simple polar solutes". In: *The Journal of Chemical Physics* 121.19 (2004), p. 9539. doi: 10.1063/1.1789935.

[314] Wang, G. Z., Wolf, D., and Goff, S. P. "EBP1, a novel host factor involved in primer binding site-dependent restriction of moloney murine leukemia virus in embryonic cells." In: *Journal of virology* 88.3 (2014), pp. 1825–9. doi: 10.1128/JVI.02578-13.

[315] Wang, H., Hartmann, C., Schütte, C., and Delle Site, L. "Grand-Canonical-like Molecular-Dynamics Simulations by Using an Adaptive-Resolution Technique". In: *Physical Review X* 3.1 (Mar. 2013), p. 011018. doi: 10.1103/PhysRevX.3.011018. arXiv: arXiv:1301.4802v1.

[316] Wang, J., Shao, Q., Cossins, B. P., Shi, J., Chen, K., and Zhu, W. "Thermodynamics calculation of protein–ligand interactions by QM/MM polarizable charge parameters". In: *Journal of Biomolecular Structure and Dynamics* 1102.March 2015 (2015), pp. 1–14. doi: `10.1080/07391102.2015.1019928`.

[317] Wang, L., Fried, S. D., Boxer, S. G., and Markland, T. E. "Quantum delocalization of protons in the hydrogen-bond network of an enzyme active site". In: *Proceedings of the National Academy of Sciences* 111.52 (Dec. 2014), pp. 18454–18459. doi: `10.1073/pnas.1417923111`. arXiv: `1501.0241`.

[318] Warshel, A. and Karplus, M. "Calculation of ground and excited state potential surfaces of conjugated molecules. I. Formulation and parametrization". In: *Journal of the American Chemical Society* 94.16 (1972), pp. 5612–5625. doi: `10.1021/ja00771a014`.

[319] Warshel, A. and Levitt, M. "Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme". In: *Journal of Molecular Biology* 103.2 (May 1976), pp. 227–249. doi: `10.1016/0022-2836(76)90311-9`.

[320] Watanabe, H. C., Kubař, T., and Elstner, M. "Size-Consistent Multipartitioning QM/MM: A Stable and Efficient Adaptive QM/MM Method". In: *Journal of Chemical Theory and Computation* 10.10 (Oct. 2014), pp. 4242–4252. doi: `10.1021/ct5005593`.

[321] Wentzel, G. "Zur Quantenoptik". In: *Zeitschrift für Physik* 22.1 (Dec. 1924), pp. 193–199. doi: `10.1007/BF01328122`.

[322] Wentzel, G. "Zur Quantentheorie des Röntgenbremsspektrums". In: *Zeitschrift für Physik* 27.1 (Dec. 1924), pp. 257–284. doi: `10.1007/BF01328033`.

[323] Weyl, H. "Quantenmechanik und Gruppentheorie". In: *Zeitschrift für Physik* 46.1-2 (Nov. 1927), pp. 1–46. doi: `10.1007/BF02055756`.

[324] Whatmore, A. M. and Reed, R. H. "Determination of turgor pressure in Bacillus subtilis: a possible role for K+ in turgor regulation". In: *Journal of General Microbiology* 136.12 (Dec. 1990), pp. 2521–2526. doi: `10.1099/00221287-136-12-2521`.

[325] Wicki, A. and Christofori, G. "The potential role of podoplanin in tumour invasion". In: *British Journal of Cancer* 96.1 (Jan. 2007), pp. 1–5. doi: `10.1038/sj.bjc.6603518`.

[326] Wiener, N. "The Average of an Analytic Functional". In: *Proceedings of the National Academy of Sciences* 7.9 (Sept. 1921), pp. 253–260. doi: `10.1073/pnas.7.9.253`.

[327] Wiener, N. "The Average of an Analytic Functional and the Brownian Movement". In: *Proceedings of the National Academy of Sciences* 7.10 (Oct. 1921), pp. 294–298. doi: `10.1073/pnas.7.10.294`.

[328] Wigner, E. "On the Quantum Correction For Thermodynamic Equilibrium". In: *Physical Review* 40.5 (June 1932), pp. 749–759. doi: 10.1103/PhysRev.40.749.

[329] Wigner, E. P. "The unreasonable effectiveness of mathematics in the natural sciences. Richard courant lecture in mathematical sciences delivered at New York University, May 11, 1959". In: *Communications on Pure and Applied Mathematics* 13.1 (Feb. 1960), pp. 1–14. doi: 10.1002/cpa.3160130102.

[330] Wishart Research Group. *DrugBank*. url: https://www.drugbank.ca/ [Accessed: 2018-02-12]. 2017.

[331] Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., Pon, A., Knox, C., and Wilson, M. "DrugBank 5.0: a major update to the DrugBank database for 2018". In: *Nucleic Acids Research* 46.D1 (Jan. 2018), pp. D1074–D1082. doi: 10.1093/nar/gkx1037.

[332] Woodcock, H. L., Hodoščeк, M., Gilbert, A. T. B., Gill, P. M. W., Schaefer, H. F., and Brooks, B. R. "Interfacing Q-Chem and CHARMM to perform QM/MM reaction path calculations". In: *Journal of Computational Chemistry* 28.9 (2007), pp. 1485–1502. doi: 10.1002/jcc.20587.

[333] Woods, C. J., Manby, F. R., and Mulholland, A. J. "An efficient method for the calculation of quantum mechanics/molecular mechanics free energies". In: *Journal of Chemical Physics* 128.1 (2008). doi: 10.1063/1.2805379.

[334] World Health Organization. *Antibiotic resistance: Fact Sheet*. url: http://www.who.int/mediacentre/factsheets/antibiotic-resistance/en/[Accessed:2018-03-16]. 2017.

[335] World Health Organization. *Antimicrobial Resistance: Fact Sheet*. url: http://www.who.int/mediacentre/factsheets/fs194/en/[Accessed:2018-03-16]. 2018.

[336] World Health Organization. *Genes and Human Disease*. url: http://www.who.int/genomics/public/geneticdiseases/en/index2.html [Accessed: 2018-02-12].

[337] World Helath Organization. "Antibacterial Agents in Clinical Development: An Analysis of the Antibacterial Clinical Development Pipeline, Including Tuberculosis". In: 2017.

[338] Wu, X.-p., Gagliardi, L., and Truhlar, D. G. "Combined quantum mechanical and molecular mechanical method for metal–organic frameworks: proton topologies of NU-1000". In: *Physical Chemistry Chemical Physics* 20.3 (2018), pp. 1778–1786. doi: 10.1039/C7CP06751H.

[339] wwPDB. *Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description - Version 3.30*. URL: http://www.wwpdb.org/documentation/file-format-content/format33/v3.3.html [Accessed: 2018-02-12]. 2012.

[340] Xiaoping, Z. and Fajun, Y. "Regulation of SREBP-Mediated Gene Expression." In: *Sheng Wu Wu Li Hsueh Bao* 28.4 (2012), pp. 287–294. DOI: 10.3724/SP.J.1260.2012.20034.

[341] Xu, M. and Lill, M. A. "Induced fit docking, and the use of QM/MM methods in docking". In: *Drug Discovery Today: Technologies* 10.3 (2013), e411–e418. DOI: 10.1016/j.ddtec.2013.02.003.

[342] Yang, F., Vought, B. W., Satterlee, J. S., Walker, A. K., Jim Sun, Z.-Y., Watts, J. L., DeBeaumont, R., Mako Saito, R., Hyberts, S. G., Yang, S., Macol, C., Iyer, L., Tjian, R., Heuvel, S. van den, Hart, A. C., Wagner, G., and Näär, A. M. "An ARC/Mediator subunit required for SREBP control of cholesterol and lipid homeostasis". In: *Nature* 442.7103 (Aug. 2006), pp. 700–704. DOI: 10.1038/nature04942.

[343] Yang, W., Cui, Q., Min, D., and Li, H. "QM/MM Alchemical Free Energy Simulations: Challenges and Recent Developments". In: *Journal of Chemical Theory and Computation*. Vol. 10. 4. Apr. 2010, pp. 51–62. DOI: 10.1016/S1574-1400(10)06004-4.

[344] Yesselman, J. D., Price, D. J., Knight, J. L., and Brooks, C. L. "MATCH: An atom-typing toolset for molecular mechanics force fields". In: *Journal of Computational Chemistry* 33.2 (2012), pp. 189–202. DOI: 10.1002/jcc.21963.

[345] Yin, J., Henriksen, N. M., Slochower, D. R., Shirts, M. R., Chiu, M. W., Mobley, D. L., and Gilson, M. K. "Overview of the SAMPL5 host–guest challenge: Are we doing better?" In: *Journal of Computer-Aided Molecular Design* 31.1 (Jan. 2017), pp. 1–19. DOI: 10.1007/s10822-016-9974-4.

[346] Yin, R. and Hamblin, M. "Antimicrobial Photosensitizers: Drug Discovery Under the Spotlight". In: *Current Medicinal Chemistry* 22.18 (June 2015), pp. 2159–2185. DOI: 10.2174/0929867322666150319120134.

[347] Yu, M., Wang, C., Kyle, A. F., Jakubec, P., Dixon, D. J., Schrock, R. R., and Hoveyda, A. H. "Synthesis of macrocyclic natural products by catalyst-controlled stereoselective ring-closing metathesis". In: *Nature* 479.7371 (Nov. 2011), pp. 88–93. DOI: 10.1038/nature10563.

[348] Yu, W., He, X., Vanommeslaeghe, K., and MacKerell, A. D. "Extension of the CHARMM general force field to sulfonyl-containing compounds and its utility in biomolecular simulations". In: *Journal of Computational Chemistry* 33.31 (Dec. 2012), pp. 2451–2468. DOI: 10.1002/jcc.23067.

[349] Yuan, Y., Liu, Z.-Q., Jin, H., Sun, S., Liu, T.-J., Wang, X., Fan, H.-J., Hou, S.-K., and Ding, H. “Photodynamic antimicrobial chemotherapy with the novel amino acid-porphyrin conjugate 4I: In vitro and in vivo studies”. In: *PLOS ONE* 12.5 (May 2017). Ed. by Hamblin, M., e0176529. doi: 10.1371/journal.pone.0176529.

[350] Zhang, H., Williams, P. S., Zborowski, M., and Chalmers, J. J. “Binding affinities/avidities of antibody–antigen interactions: Quantification and scale-up implications”. In: *Biotechnology and Bioengineering* 95.5 (Dec. 2006), pp. 812–829. doi: 10.1002/bit.21024.

[351] Zhang, S., Kumar, K., Jiang, X., Wallqvist, A., and Reifman, J. “DOVIS: an implementation for high-throughput virtual screening using AutoDock.” In: *BMC bioinformatics* 9 (2008), p. 126. doi: 10.1186/1471-2105-9-126.

[352] Zhang, X., Wong, S. E., and Lightstone, F. C. “Message passing interface and multithreading hybrid for parallel molecular docking of large databases on petascale high performance computing machines”. In: *Journal of Computational Chemistry* 34.11 (2013), pp. 915–927. doi: 10.1002/jcc.23214.

[353] Zhang, Y., Lu, Y., Zhou, H., Lee, M., Liu, Z., Hassel, B. A., and Hamburger, A. W. “Alterations in cell growth and signaling in ErbB3 binding protein-1 (Ebp1) deficient mice”. In: *BMC Cell Biology* 9.1 (2008), p. 69. doi: 10.1186/1471-2121-9-69.

[354] Zhao, X., Xiaoli, Zong, H., Abdulla, A., Yang, E. S. T., Wang, Q., Ji, J.-Y., Pessin, J. E., Das, B. C., and Yang, F. “Inhibition of SREBP Transcriptional Activity by a Boron-Containing Compound Improves Lipid Homeostasis in Diet-Induced Obesity”. In: *Diabetes* 63.7 (July 2014), pp. 2464–2473. doi: 10.2337/db13-0835.

[355] Zheng, M., Kuriappan, J. A., and Waller, M. P. “Toward more efficient density-based adaptive QM/MM methods”. In: *International Journal of Quantum Chemistry* 117.6 (Mar. 2017), e25336. doi: 10.1002/qua.25336.

[356] Zheng, M. and Waller, M. P. “Adaptive quantum mechanics/molecular mechanics methods”. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 6.August (2016), pp. 369–385. doi: 10.1002/wcms.1255.

[357] Zwanzig, R. W. “High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases”. In: *The Journal of Chemical Physics* 22.8 (Aug. 1954), pp. 1420–1426. doi: 10.1063/1.1740409.

Appendices

Appendix A

Comments

- The following figures were (partially) created with MAESTRO [247] (version 11.3.015): 5.4 on page 96, 7.18 on page 147, 7.19 on page 147, 9.2 on page 170, 9.3 on page 171, 9.5 on page 174, 9.6 on page 175, 9.8 on page 178, 9.9 on page 179, 10.2 on page 189, 10.4 on page 191, 10.3 on page 190, and 11.1 on page 194.

- The following figures were created with VMD [127] (version 1.9.2): 6.3 on page 112, and 6.4 on page 113.

- The following figures were created with POV-Ray [223] (version 3.7.0): 1.5 on page 7, 7.14 on page 141, and 7.15 on page 143.

- The primary literature which was used for the theory in chapter 2 is [28], [297], [271], [281], and [108]. The author of this thesis has used elements of all of them, partially merged and rewritten the used contents in a consistent manner, and reformulated several expressions and concepts. Therefore it was nearly impossible to specify for each sentence or piece of information which of the sources has contributed to each of them. The author of this thesis does not claim to have contributed new knowledge of significance in chapter 2, rather it contains prior knowledge of significance to this thesis.

- The primary literature which was used for the basic theory in section 3.2 is [253], and for section 3.3 [356]. The author of this thesis does not claim to have contributed new knowledge of significance in these sections, rather it contains prior knowledge of significance to this thesis.

- The basic principles and concepts of section 4.2 are for instance described in [6, 58]. The author of this thesis has rewritten some of the concepts in a different way or in a different notation.

Appendix B

Configuration Files

B.1 HyperQ Configuration Files

The two types of configurations files which are used by HyperQ are listed below, the main configuration file and the batchsystem configuration file. HyperQ provides several default configuration files for the batchsystem module, dependent of the job type. Shown below is the general default configuration file which is job independent.

Listing B.1: HyperQ Main Configuration File (default settings).

```
***********************************************************************************************
*****************************************   General   *****************************************
***********************************************************************************************


workflow_id=ApplicationX_RunY
# Used within the socket names as well as the job names in the batchsystem to avoid clashes
     between parallel running instances of hyper-Q.
# In particular relevant for the batchsystem, for the socket-files a time-stamp mechanism is used
      as additional protection mechanism.
# Up to 50 characters are supported (provided the job IDs have less than 10 digits). The first
     character should be an alphabetic letter (required by some batchsystems)

verbosity_runtime=standard
# Possible values: standard, debug
# This option mainly effects the screen output and the logfiles

verbosity_nonruntime=standard
# Possible values: standard, debug
# This option affects the preparation scripts for setting up the basic workflow files (before the
      workflow is running)

temperature=298
# In Kelvin
# Used during all simulations and calculations where a temperature is needed (i.e. EQ, MD, FEC)
# Input files for the simulations are adjusted automatically by replacing the term '
     temperature_placeholder'



***********************************************************************************************
**********************************   Structure Preparation   **********************************
***********************************************************************************************

******************************************* General *******************************************

input_file_format=smi
# Currently supported are smi, pdbqt, sdf, mol2_2d_h, pdb_3d_h
# The format should be in lowercase letters
# If the subsystem is RLS, then the coordinates of the ligands have to be provided in the input
     files to suit the coordinates of the receptor.
# Ligand names should not contain the underscore "_" letter
# Resname doesn't matter for the ligand, it will be changed to LIG, as well as the chain which
     will be set to L
```

```
receptor_mode=common
# Possible values: individual, common
# If set to ’individual’, then a receptor has to be in the receptor folder for each ligand with
    the same name as the ligand
# Only needed if the subsystem includes a receptor (e.g. RLS)

receptor_basename=
# Only required if receptor_mode=common and if the simulations include a receptor (e.g. RLS)
# If receptor_FFparameter_source=folder, then there is no need to include in the receptor
    filenames the words ’unique’ or ’typed’. This is assumed implicitly, and filenames including
     these terms will be created.

receptor_type=P
# Possible values: P (Protein), H (Host)
# Only required if the simulations include a receptor (e.g. RLS)

box_edge_length_L=10
waterbox_padding_size_LS=10
waterbox_padding_size_RLS=10

ligand_FFparameter_source=MATCH
# Possible choices: folder, MATCH
# If set to folder, then in the folder input-files/ligands/FF there need to be the files for each
     ligand:
#   * ligand_name.rtf
#   * ligand_name.prm
# The prm files need to have a proper END statement at the end of the file

receptor_FFparameter_source=MATCH
# Possible choices: folder, MATCH
# Only relevant if receptor_type=H
# If set to folder, then in the folder input-files/receptor there need to be for each receptor (
    common or all the individual ones) the following files:
#   * receptorname.pdb
#   * receptorname.rtf
#   * receptorname.prm
# The prm files need to have a proper END statement at the end of the file
# If receptor_FFparameter_source=folder, then there is no need to include in the receptor
    filenames the words ’unique’ or ’typed’. This is assumed implicitly, and filenames including
     these terms will be created.


******************************************** LOMAP ********************************************

lomap_mol2_folder=mol2-Li
# For lomap, relative path name w.r.t. to input-files/ligands

draw_pairwise_mcs=false
# Possible values: false or true (case insensitive)

lomap_ncpus=6

mcs_time=1
# The time which lomap can take for finding the MCS (maximal common substructure) per molecule-
    pair



***********************************************************************************************
***************************   Simulation Settings (OPT, EQ, MD, CE)  **************************
***********************************************************************************************

******************************************** General ********************************************

nbeads=4
# Total number of path integral beads per particle
# Minimum value: 1
# Is only relevant for MD simulations and CE

cell_dimensions_scaling_factor_L=2
```

```
cell_dimensions_scaling_factor_LS=2
cell_dimensions_scaling_factor_RLS=2
# This value will be used for the variables cell_dimensions_scaled_rounded and
     cell_dimensions_scaled_odd_rounded in the simulation software input files (e.g. cp2k)

signpostings_activate=true
# Possible values: false, true

signpostings_minimum_waiting_time=180
# Possible values: Positive integers including zero
# Unit: Seconds

signpostings_dispersion_time_maximum=180
# Possible values: Positive integers (excluding 0)
# Unit: Seconds


****************************************** TD Cycles ******************************************

tdw_count_total=4
# The total number of thermodynamic windows
# The following condition has to be satisfied: tdw_count_total = tdw_count_msp_transformation +
     tdw_count_es_transformation_initial + tdw_count_es_transformation_final
# Minimum value: 1

tdw_count_es_transformation_initial=0
tdw_count_es_transformation_final=0
# Possible values: any positive int

tdc_es_tds_configurations_system1=1
tdc_es_tds_configurations_system2=1
# Possible values: Colon separated list of configurations (e.g. 1:0.5:0 if
     tdcycle_es_transformation_type=lambda or 0:2:4 if tdcycle_es_transformation_type=hq)
# The number of values specified in both variables together has to be either
#     * equal to tdw_count_total + 1
#     * equal to tdw_count_es_transformation_initial/final
# Usually
#     * the first value for tdc_es_tds_configurations_system1 and the last value
#            of tdc_es_tds_configurations_system2 are set to 1
# Usually if the variables tdw_count_es_transformation_initial/final have the length
# tdw_count_total, then
#     * the number of different values for system 1 equals tdw_count_es_transformation_initial + 1
#     * the number of different values for system 2 equals tdw_count_es_transformation_final + 1
#     * the values for system 1 change at the beginning during the first
     tdw_count_es_transformation_initial + 1 indices
#     * the values for system 2 change at the end during the last
     tdw_count_es_transformation_final + 1 indices
# The order of values specified corresponds to the direction from the starting state (TDS 1) to
     the final state (last TDS)

es_transformation_atoms_to_transform=dawn
# Possible values:
# * dao : dummy atoms only
# * dawn : dummy atoms with neighbors (directly bonded atoms)
# * ligand : (the entire) ligand

tdcycle_msp_transformation_type=hq
# Possible options: hq, lambda
# If using lambda, use the same input file for cp2k with the ending lambda in the variables below
      for cp2k, for both the sys1 and sys2 files (Todo : improve description)

tdw_count_msp_transformation=4
# The total number of thermodynamic windows during which the two molecular systems are
     transformed into each other
# Minimum value: 1

tdcycle_si_activate=true
# Possible values: true, false

tdcycle_si_hydrogen_single_step=false
```

```
# Possible values: true, false
# Only relevant if tdcycle_si_activate=true

tdcycle_si_separate_neighbors=true
# Possible values: true, false
# Only relevant if tdcycle_si_activate=true

tdcycle_si_consider_branches=true
# Possible values: true, false
# Only relevant if tdcycle_si_activate=true and tdcycle_si_separate_neighbors=true


********************************************* CP2K ********************************************

# At least one of the two CP2K input folders for each GEO_OPT, EQ, MD and CE has to contain the
     following files:
# When used for GEO_OPT:
# * main.opt.lambda
# * main.opt.sys1
# * main.opt.sys2
# When used for EQ:
# * main.eq.lambda
# * main.eq.sys1
# * main.eq.sys2
# When used for MD simulations:
# * main.ipi.lambda
# * main.ipi.sys1
# * main.ipi.sys2
# Optional for GEO_OPT, EQ, MD, and CE are files of the form sub.*
# Possible variables in the cp2k input files: cell_dimensions_full_rounded,
     cell_dimensions_half_rounded, cell_dimensions_odd_rounded, tdsname, lambda_value

# Ligand (L) input folder
inputfolder_cp2k_opt_general_L=hqf.general.type1
inputfolder_cp2k_eq_general_L=hqf.general.type1
inputfolder_cp2k_md_general_L=hqf.general.type1
inputfolder_cp2k_ce-bp_general_L=hqf.general.type1
inputfolder_cp2k_opt_specific_L=hqf.specific.MM
inputfolder_cp2k_eq_specific_L=hqf.specific.MM
inputfolder_cp2k_md_specific_L=hqf.specific.MM
inputfolder_cp2k_ce-high_specific_L=hqf.specific.MM
# the ce-high folders are only used during the CE if the free energy method is a two level method
      (NBB,

# Ligand, solvent (LS) input folder
inputfolder_cp2k_opt_general_LS=hqf.general.type1
inputfolder_cp2k_eq_general_LS=hqf.general.type1
inputfolder_cp2k_md_general_LS=hqf.general.type1
inputfolder_cp2k_ce_general_LS=hqf.general.type1
inputfolder_cp2k_opt_specific_LS=hqf.specific.MM
inputfolder_cp2k_eq_specific_LS=hqf.specific.MM
inputfolder_cp2k_md_specific_LS=hqf.specific.MM
inputfolder_cp2k_ce-bp_specific_LS=hqf.specific.MM

# Protein, ligand, solvent (RLS) input folder
inputfolder_cp2k_opt_general_RLS=hqf.general.type1
inputfolder_cp2k_eq_general_RLS=hqf.general.type1
inputfolder_cp2k_md_general_RLS=hqf.general.type1
inputfolder_cp2k_ce_general_RLS=hqf.general.type1
inputfolder_cp2k_opt_specific_RLS=hqf.specific.MM
inputfolder_cp2k_eq_specific_RLS=hqf.specific.MM
inputfolder_cp2k_md_specific_RLS=hqf.specific.MM
inputfolder_cp2k_ce-bp_specific_RLS=hqf.specific.MM

cp2k_random_seed=random
# Possible values: random, <integer>
# Sets the seed of the prng in the cp2k main input files (opt and eq)
# If an integer is specified, it will be used as the random seed
# If 'random' is specified, a random integer will be generated and used (recommended). This can
     be useful in particular when a simulation crashes during the run and has to be restarted.
```

```
cp2k_command=cp2k-5.1.sopt
#cp2k-5.1.sopt
#cp2k_command=/home/wagner/cgorgulla/Dropbox/Software/installed/bin/cp2k.ssmp
#cp2k_command=/programs/x86_64-linux/cp2k/4.1/exe/Linux-x86-64-gfortran/cp2k.ssmp
#cp2k_command=/nmr/programs/cp2k-4.1/cp2k.popt


********************************************* i-PI ********************************************

# Ligand (L) input files
inputfile_ipi_md_L=ipi.in.md.nvt.xml
inputfile_ipi_ce_L=ipi.in.ce.cp2k.xml

# Ligand, solvent (LS) input files
inputfile_ipi_md_LS=ipi.in.md.nvt.xml
inputfile_ipi_ce_LS=ipi.in.ce.cp2k.xml

# Protein, ligand, solvent (RLS) input files
inputfile_ipi_md_RLS=ipi.in.md.nvt.xml
inputfile_ipi_ce_RLS=ipi.in.ce.cp2k.xml

ipi_set_randomseed=true
# Possible values: true, false
# Replaces the seed of the prng in the i-pi input files with a random number


******************************************** i-QI *********************************************

# Ligand (L) input files
inputfile_iqi_md_L=
inputfile_iqi_constraints_L=
# Normally not needed because iqi is usually not used for L only simulations

# Ligand, solvent (LS) input files
inputfile_iqi_md_LS=iqi.in.main.LS.xml
inputfile_iqi_constraints_LS=iqi.in.sub.constraints.LS.xml

# Protein, ligand, solvent (RLS) input files
inputfile_iqi_md_RLS=iqi.in.main.RLS.xml
inputfile_iqi_constraints_RLS=iqi.in.sub.constraints.RLS.xml


************************************ Geometry Optimization ************************************

opt_continue=true
# Possible values (case insensitive): true, false
# If true, then
#   * during _pro_
#       * non-existent TDS-folders will be newly created
#       * and existing TDS-folders which contain already restart files will be prepared for the
     next run provided the there are no opt-pp-files for this TDS yet, in which case it will be
     skipped
#       * the general MSP folder will only be newly prepared if at least one of the files system.
     a1c1.[uc]-atoms is present
#   * during _rop_
#       * the simulation will only be started if there are no opt-pp-files yet for this TDS
#   * during _ppo_
#       * the pp  will only be carried out if there are no opt-pp-files yet for this TDS
# If false, then
#   * during _pro_
#       * the opt/msp_name/subsystem general folder will be newly prepared
#       * each TDS folder will be newly prepared (and wiped if existent)
# If the input files have changed and should be updated, this setting should be set either to
     false, otherwise one needs to delete the folders which should be newly prepared

opt_programs_L=cp2k
opt_programs_LS=cp2k
opt_programs_RLS=cp2k
# possible: cp2k
```

```
opt_max_steps_L=500
opt_max_steps_LS=500
opt_max_steps_RLS=500
# Input files for the simulations are adjusted automatically by replacing the term '
    opt_max_steps_placeholder'

opt_trajectory_stride_L=10
opt_trajectory_stride_LS=10
opt_trajectory_stride_RLS=10
# Input files for the simulations are adjusted automatically by replacing the term '
    opt_trajectory_stride_placeholder'

opt_restart_stride_L=100
opt_restart_stride_LS=100
opt_restart_stride_RLS=100
# Input files for the simulations are adjusted automatically by replacing the term '
    opt_restart_stride_placeholder'

opt_type_L=MM
opt_type_LS=MM
opt_type_RLS=MM
# Possible values: MM, QM, QMMM
# Currently obsolete

opt_timeout_L=100
opt_timeout_LS=100
opt_timeout_RLS=100
# Unit: seconds
# This value is used to detect whether the optimization is completed by checking if the file is
    still changing or not,
# because sometimes a program (in particular CP2K is done but the program doesn't terminate)


********************************* Equilibration Simulations *********************************

eq_activate=true
# Possible values: true, false
# If false, _pmd_ will use the optimization output files, and _eq_ will not be run in hq-pipes
# If true, _pmd_ will use the equilibration output files, and _eq_ can be run in hq-pipes

eq_continue=true
# Possible values: true, false
# Details: Same as for opt_continue

eq_programs_L=cp2k
eq_programs_LS=cp2k
eq_programs_RLS=cp2k
# possible: cp2k

eq_type_L=MM
eq_type_LS=MM
eq_type_RLS=MM
# Possible values: MM, QM, QMMM
# Currently obsolete

eq_total_steps_L=100000
eq_total_steps_LS=100000
eq_total_steps_RLS=100000
# Input files for the simulations are adjusted automatically by replacing the term '
    eq_total_steps_placeholder'

eq_trajectory_stride_L=1000
eq_trajectory_stride_LS=1000
eq_trajectory_stride_RLS=1000
# The stride w.r.t. the time steps when snapshot files are written out
# Input files for the simulations are adjusted automatically by replacing the term '
    eq_trajectory_stride_placeholder'

eq_restart_stride_L=1000
```

```
eq_restart_stride_LS=1000
eq_restart_stride_RLS=1000
# The stride w.r.t. the time steps when restart files are written out
# Input files for the simulations are adjusted automatically by replacing the term '
     eq_restart_stride_placeholder'

eq_timeout_L=100
eq_timeout_LS=100
eq_timeout_RLS=100
# Unit: seconds
# This value is used to detect whether the equilibration is completed by checking if the file is
     still changing or not,
# because sometimes a program (in particular CP2K is done but the program doesn't terminate)


**************************************** MD Simulations ***************************************

md_continue=true
# Possible values (case insensitive): true, false
# If true, then during _prm_ non-existent MD-folders will be newly created, and existing ones
     which have already ipi restart files will be prepared for the next run (without deleting the
      previous runs)
# If false, then the md/msp_name/subsystem will be newly prepared (and will be wiped if already
     existent)
# If the input files have changed and should be updated, this setting should be set either to
     false, otherwise one needs to delete the folders which should be newly prepared
# This setting has mainly effects on _prm_, only very few on _rmd_

md_programs_L=ipi-cp2k
md_programs_LS=ipi-cp2k
md_programs_RLS=ipi-cp2k
# Possible values: ipi-cp2k, ipi-cp2k-iqi
# Will also be used for the CE (due to the restart files requiring the same clients)

md_type_L=MM
md_type_LS=MM
md_type_RLS=MM
# Possible values: QMMM, MM, QM
# Currently obsolete

md_total_steps_L=100000000
md_total_steps_LS=100000000
md_total_steps_RLS=100000000
# Input files for the simulations are adjusted automatically by replacing the term '
     md_total_steps_placeholder'

md_restart_stride_L=1000
md_restart_stride_LS=1000
md_restart_stride_RLS=1000
# The stride w.r.t. the time steps during the MD simulations when restart/property/cell files are
      written out for the later CEs
# Input files for the simulations are adjusted automatically by replacing the term '
     md_restart_stride_placeholder' (should be used at least for the above mentioned three types
     of files)

md_trajectory_centroid_stride_L=10000
md_trajectory_centroid_stride_LS=10000
md_trajectory_centroid_stride_RLS=10000
# The stride w.r.t. the time steps during the MD simulations when trajectory files for the
     centroids are written out
# Input files for the simulations are adjusted automatically by replacing the term '
     md_trajectory_centroid_stride_placeholder'

md_trajectory_beads_stride_L=10000
md_trajectory_beads_stride_LS=10000
md_trajectory_beads_stride_RLS=10000
# The stride w.r.t. the time steps during the MD simulations when trajectory files for the beads
     are written out
# Input files for the simulations are adjusted automatically by replacing the term '
     md_trajectory_beads_stride_placeholder'
```

```
md_forces_stride_L=10000
md_forces_stride_LS=10000
md_forces_stride_RLS=10000
# The stride w.r.t. the time steps during the MD simulations when force files are written out
# Input files for the simulations are adjusted automatically by replacing the term '
    md_forces_stride_placeholder'

md_keep_logfiles=false
# Possible values: false, true
# true: Default output files are kept
# false: No screen and related output files are kept
# Does not affect the verbosity settings of the simulation input files

md_timeout_L=1000
md_timeout_LS=1000
md_timeout_RLS=1000
# Here should be taken into account that between the start of i-pi and CP2K there can be a time
    delay of a few seconds (usually around 5 seconds)
# Timeout starts counting after CP2K (and possibly iqi) has been started


************************************** Cross-Evaluations *************************************

ce_continue=true
# Possible values (case insensitive): true, false
# If true, then * during _rce_ snapshots will be skipped for which the file ipi.out.properties
    already exists and contains a line with property values
#               * during _prc_ only snapshots will be prepared which are not already prepared (e.
    g. no finished snapshots will be overwritten)
# If false, then * during _rce_ every snapshot will be computed regardless of whether they have
    been computed before or not
#                * during _prc_ all snapshots will be newly prepared (and the ce/msp_name/
    subsystem folder will be wiped if already existent)
# If the input files have changed and should be updated, this setting should be set either to
    false, otherwise one needs to delete the folders which should be newly prepared

ce_verbosity=normal
# Possible values: normal, debug
# normal: Only essential output files of the simulation programs (i-Pi, CP2K, i-QI) are kept
# debug: All the simulation output files are kept (not recommended for large scale production
    runs)

ce_type_L=MM
ce_type_LS=MM
ce_type_RLS=MM
# If umbrella_sampling=false, then this setting should be the same as md_type (since the MD files
     are copied)
# Possible values: QMMM, MM, QM
# Currently obsolete

ce_first_restart_ID_L=1
ce_first_restart_ID_LS=1
ce_first_restart_ID_RLS=1

ce_stride_L=1
ce_stride_LS=1
ce_stride_RLS=1
# The stride which is applied during the CE, i.e. only every <stride_ce>th restart file of ipi
    from the MD simulations is used for the CEs
# The restart files are indexed starting at one (1, 2, 3, ...)

ce_timeout_L=100
ce_timeout_LS=100
ce_timeout_RLS=100
# Here should be taken into account that between the start of i-pi and CP2K there can be a time
    delay of a few seconds (usually around 5 seconds, but sometimes up to hundreds of seconds,
    either due to overloaded systems or because CP2k hangs in the beginning which sometimes
    happens)
# Timeout starts counting after CP2K (and possibly iqi) has been started
# The snapshot run will terminate as soon as the result is there, not until the timeout is
    reached


******************************************** FEC ********************************************

umbrella_sampling=false
method=BAR
# Possible values: BAR (more are on the way)
# These setting will affect the cross evaluations and the free energy computations

fec_first_snapshot_index_L=1
fec_first_snapshot_index_LS=1
fec_first_snapshot_index_RLS=1
fec_stride_L=1
fec_stride_LS=1
fec_stride_RLS=1
# The stride which is applied before carrying out the FEC, i.e. only every <stride_fec>th
    snapshots of the cross evaluation is used
# The snapshot folder in the cross evaluation folders are indexed starting by 1

C_absolute_tolerance=0.1
delta_F_min=-1000
delta_F_max=1000


************************************** Parallelization **************************************

ncpus_cp2k_opt_L=1
ncpus_cp2k_opt_LS=1
ncpus_cp2k_opt_RLS=1
ncpus_cp2k_eq_L=1
ncpus_cp2k_eq_LS=1
ncpus_cp2k_eq_RLS=1
ncpus_cp2k_md_L=1
ncpus_cp2k_md_LS=1
ncpus_cp2k_md_RLS=1
ncpus_cp2k_ce_L=1
ncpus_cp2k_ce_LS=1
ncpus_cp2k_ce_RLS=1

fes_opt_parallel_max_L=10
fes_opt_parallel_max_LS=10
fes_opt_parallel_max_RLS=10
fes_eq_parallel_max_L=10
fes_eq_parallel_max_LS=10
fes_eq_parallel_max_RLS=10
fes_md_parallel_max_L=10
fes_md_parallel_max_LS=10
fes_md_parallel_max_RLS=10
fes_ce_parallel_max_L=1
fes_ce_parallel_max_LS=1
fes_ce_parallel_max_RLS=1

command_prefix_bs_subjob=srun -N 1 -n 1
# Cray/HLRN: aprun -cc none -n 1
# Slurm: srun -N 1 -n 1
command_prefix_bs_task=bash
command_prefix_gen_run_one_pipe_sub=bash
command_prefix_opt_run_one_opt=bash
command_prefix_eq_run_one_eq=bash
command_prefix_md_run_one_md=bash
command_prefix_ce_run_one_snapshot=bash
# Can contain spaces, e.g. options to the prefix command
# Common prefixes are: bash, aprun, setsid, ...
# Should not be empty. 'bash' can be used if no other prefix is needed for a bash script for
    instance
# setsid ist not needed for command_prefix_bs_task, because the tasks are automatically run in
    their own process groups and setsid would therefore not have much effect
```

```
***************************************** Batchsystem *****************************************

batchsystem=slurm
# Supported: lsf, slurm, sge, mtp (for MOAB/torque/pbs)

tasks_parallel_delay_time=10
# Possible values: Non-negative integer
# Unit: Seconds
# Summary:  The waiting time before the next task of the same subjob file is started. This is
    useful if different tasks are preparing files in the same directory
#           (usually common files for the same MSP) to avoid race conditions which can cause
    HyperQ to fail
#           The value is used during the job creation with hq_bs_prepare_jobfiles.sh
# Recommended value: At least 10 seconds

minimum_task_time=5
# Possible values: Positive integer
# Unit: Seconds
# If the runtime of the tasks of a subjob is less than the minimum_task_time for at least one of
    the tasks, an internal error is raised.
# The task_parallel_delay_time is not counted as runtime of the tasks
# Useful for detecting immediate failures (preventing of the start of the task at all). We cannot
     rely on the exit codes of the tasks.
# Recommended value: 5 seconds (each hqf_gen_run_one_pipe.sh runs at least for 15 seconds due to
    a short initial sleep)
```

Listing B.2: HyperQ Batchsystem Configuration File (default settings).

```
***********************************************************************************************
****************************************    General    ****************************************
***********************************************************************************************

job_initial_sleeping_time_max=120
# In seconds
# Sleeping a random amount of time to avoid race conditions when jobs are starting and
    simultaneously
# Relevant if the batchsystem starts pending jobs simultaneously
# Not relevant for multiple tasks per subjob since we disperse them already in a controlled
    manner (see the tasks_parallel_delay_time option in HQ config file)



***********************************************************************************************
************************************    Job Termination    ************************************
***********************************************************************************************

*****************************************   General   *****************************************

# Order of precedence of job termination scenarios
# 1) signals_type1
# 2) signals_type2
# 3) signals_type3
# 4) errors_job (but they are deactivated during other error and signal responses)
# 5) errors_subjob
# 6) errors_pipeline
# 7) job_success

terminate_job=false
# Possible values: true, false
# If true, the job current job will be terminated without failure as soon as possible.
# Running simulations will be terminated, and pipelines will not progress to the next step.
# Also new jobs will not be submitted.
# This setting can be changed during runtime.

prevent_new_job_submissions=false
# Possible values: true, false
# This option takes precedence over other options which specify if new jobs should be started at
    the end of a job (i.e. it overrides them)
# Old jobs will check this setting at the end before submitting new jobs (as specified by other
    options).
# Can be changed during runtime

job_success_actions=submit_new_job
# Possible values: exit, prepare_new_job, submit_new_job
# Can be unset (i.e. having no values)
# Multiple values can be specified by using colons: value1:value2:...
# The 'exit' action has no effect, can be used a s placeholder
# If no action is specified, the job will simply exit
# Can be changed during runtime

job_success_new_job_jtl=next
# Possible values: same, next, [a-j]
# When set to 'same' the current jtl will be retained
# When set to 'next' the current jtl will simply be changed to the next letter in the alphabet.
    The highest letter supported is j.


*************************************   Signal Handling   *************************************

signals_type1=10                             # time signal
signals_type2=1:2:3:9:12:15:18               # termination signal, slurm uses 15 and 18 for
    preempting jobs
signals_type3=
# Multiple signals can be specified by using colons as a delimiter: signal1:signal2:...
# The signal can be anything which BASH can trap, i.e.
#       * Signal numbers
#       * Signal names
# See kill -l for a complete list of signal numbers and names
# Changes of this setting will have no effect during the runtime of the jobs (only the initial
    value will be used)

signals_type1_response=prepare_new_job:submit_new_job
signals_type2_response=prepare_new_job
signals_type3_response=exit
# Possible values: exit, prepare_new_job, submit_new_job
# Can be unset (i.e. having no values)
# Multiple values can be specified by using colons as delimiters
# The 'exit' action has no effect, can be used a s placeholder
# If no action is specified, the job will simply exit
# Can be changed during runtime

signals_type1_new_job_jtl=same
signals_type2_new_job_jtl=same
signals_type3_new_job_jtl=same
# Possible values: same, next, [a-j]
# When set to 'same' the current jtl will be retained
# When set to 'next' the current jtl will simply be changed to the next letter in the alphabet.
    The highest letter supported is j.
# Can be changed during runtime


**************************************   Error Handling   *************************************

errors_job_response=exit
errors_subjob_response=exit
errors_pipeline_response=exit
# Possible values: exit, ignore, prepare_new_job, submit_new_job
# The 'exit' action has no effect, can be used a s placeholder
# If no action is specified, the job will simply exit
# If the 'ignore' action is specified, all other actions will be skipped and the error ignored.
    Otherwise the job will be terminated
# Can be changed during runtime, except the errors_subjob_response

errors_job_new_job_jtl=same
errors_subjob_new_job_jtl=same
errors_pipeline_new_job_jtl=same
# Possible values: same, next, [a-j]
# When set to 'same' the current jtl will be retained
# When set to 'next' the current jtl will simply be changed to the next letter in the alphabet.
    The highest letter supported is j.
# Can be changed during runtime



***********************************************************************************************
******************************    Job Resource Configuration    *******************************
***********************************************************************************************

partition=shared
# also called queue

walltime=07-00:00:00
# format for slurm: dd-hh:mm:ss
# format for MTP: hh:mm:ss
# format for SGE: hh:mm:ss
# format for LSF: hh:mm
# for all: always fill up with two digits per field (used be the job scripts)

nodes_per_job=1
# Not available for LSF and SGE (is always set to 1)
# It is not recommended to change this value during runtime

cpus_per_subjob=1
# SLURM: sets the cpus-per-task variable (task = step)
# In LSF this corresponds to the number of slots per node
# Not yet available for SGE (always set to 1)
# Can be changed during runtime
```

```
memory_per_cpu=2G
# Used by SLURM, SGE
# Format for SLURM: <size[units:M/G/T]> (default units are M)
# Format for SGE: size[K/M/G]

memory_per_job=2000MB
# Used by LSF, MTP. Not needed for the HLRN (since always entire nodes will be allocated)
# Format for LSF: <size in MB>
# Format for MTP:<size><unit:MB/GB>
```

B.2 VFVS Configuration File

Listing B.3: VFVS Example Configuration File.

```
***********************************************************************************************
*******************************   Batchsystem Configuration    ********************************
***********************************************************************************************

job_letter=r
# One alphabetic character (i.e. a letter from a-z or A-Z)
# Should not be changed during runtime, and be the same for all joblines
# Required when running VF several times on the same cluster to distinguish the jobs in the
     batchsystem

batchsystem=SLURM
# Possible values: SLURM, MTP (for MOAB/TORQUE/PBS), LSF

partition=serial_requeue
# also called queue

timelimit=07-00:00:00
# format for slurm: dd-hh:mm:ss
# format for MTP: hh:mm:ss
# format for SGE: hh:mm:ss
# format for LSF: hh:mm
# for all: always fill up with two digits per field (used be the job scripts)

nodes_per_job=1
# equals steps per job bacause we start one step per node
# not yet available for LSF and SGE (is always set to  1)
# Should not be changed during runtime, and be the same for all joblines

cpus_per_step=6
# sets the slurm cpus-per-task variable (task = step) in SLURM
# in LSF this corresponds to the number of slots per node
# Should not be changed during runtime, and be the same for all joblines
# Not yet available for SGE (always set to 1)

queues_per_step=6
# sets the number of queues/processes per step
# Should not be changed during runtime, and be the same for all joblines
# Not yet available for SGE (always set to 1)

cpus_per_queue=1
# Should equal the number: cpus-per-step/queues-per-step
# Should not be changed during runtime, and be the same for all joblines
# Not yet available for SGE (always set to 1)


***********************************************************************************************
************************************   Workflow Options    ************************************
***********************************************************************************************

ligands_per_queue=200000
# used as a limit of ligands per running queue (small values only useful for testing)

ligands_todo_per_queue=80
# used as a limit of ligands for the to-do lists
# this value should be divisible by the next setting "ligands_todo_per_refilling_step"

ligands_per_refilling_step=20
# fill the to-do files of the queues with this number of ligands per refill step
# a number roughly equal to the average of number of ligands per collection is recommended

collection_folder=../../../../collections/eIF4E_DS-s1_vs1.first-1000000.ps-50
# slash at the end is not required (optional)
# relative pathname is required w.r.t. the folder tools/
```

```
verbosity=normal
# Verbosity level in the log files
# Possible values: normal, debug

error_sensitivity=normal
# Possible values: normal, high
# high sets the shell options "-uo pipefail". Not recommended for production runs, useful mainly
     for debugging. Pipefails often occur with tar combined with head/tail in pipes, which are
     not an actual problem.
# The u-option will always lead to a direct exit of the shell script when an unset variable is
     going to be used.

error_response=fail
# Affects most errors, but not all (e.g. not the u-option of the shell)
# Possible values:
#     ignore    : ignore error and continue
#     next_job  : end this job and start new job
#     fail      : exit workflow with failure (exit code 1)

minimum_time_remaining=40
# In minutes
# A new job if the time left until the end of the walltime is smaller than the timelimit
# This is checked before each ligand is screened
# Thus the timelimit should be larger than the maximum time which is needed to screen a ligand (
     for one entire docking scenario)

dispersion_time_min=10
# One positive integer, resembling the time in seconds
dispersion_time_max=40
# One positive integer, resembling the time in seconds
# The dispersion time is used when jobs try to access the central task list.
# Each job has to wait a random amount of time in the dispersion interval.
# The effect of this is that when two jobs arrive at the same time at the central task list, the
     random waiting time will disperse their access on the central task list in time


************************************************************************************************
*******************************   Virtual Screening Options    *********************************
************************************************************************************************

docking_type_names=smina_vinardo_flexible_receptor1:smina_vinardo_flexible_receptor2:
     vina_flexible_receptor1:vina_flexible_receptor2
# Values have to be separated by colons ":" and without spaces, e.g: docking_type_names=vina-
     rigid:smina-rigid
# Used for instance for the folder names in which the output files are stored
# Should not be changed during runtime, and be the same for all joblines

docking_type_programs=smina_flexible:smina_flexible:vina:vina
# Possible values: qvina02, vina, smina_rigid, smina_flexible, adfr
# Values have to be separated by colons ":" and without spaces, e.g: docking_type_programs=vina:
     smina
# The same programs can be used multiple times
# Should not be changed during runtime, and be the same for all joblines

docking_type_replicas=5:5:5:5
# Series of integers separated by colons ":"
# The number of values has to equal the number of docking programs specified in the variable "
     docking_programs"
# The values are in the same order as the docking programs specified in the variable "
     docking_type_programs
# e.g.: docking_type_replicas=1:1
# possible range: 1-99999 per field/docking program
# The docking scenario is comprised of all the docking types and their replicas
# Should not be changed during runtime, and be the same for all joblines

docking_type_inputfolders=../input-files/smina_vinardo_flexible_receptor1:../input-files/
     smina_vinardo_flexible_receptor2:../input-files/vina_flexible_receptor1:../input-files/
     vina_flexible_receptor2
# Relative path wrt to the tools folders
# In each inputfolder must be the file config.txt which is used by the docking program to specify
     its options
# If other input files are required by the docking type, usually specified in the config.txt file
    , they have to be in the same folder
# Should not be changed during runtime, and be the same for all joblines


***********************************************************************************************
**********************************   Terminating Variables    *********************************
***********************************************************************************************

stop_after_current_docking=no

stop_after_collection=no

stop_after_job=no
```

Appendix C

Zusammenfassung

(German Summary)

Die Simulationswissenschaft, auch wissenschaftliches Rechnen genannt, hat das Potential viele der aktuellen Probleme der pharmazeutischen Forschung zu lösen. Eine der zentralen physikalischen Größen in diesem Gebiet ist die Bindungsenergie zwischen Molekülen. Eines der Hauptziele in der computergestützter Wirkstoffentwicklung (CADD) ist es, diese Größe mit solcher Genauigkeit, Verlässlichkeit und Effizienz vorherzusagen, dass dassexperimentelle Bestimmungen nicht mehr notwendig sind. Jedoch sind derartige Methoden derzeit noch nicht verfügbar.

Ein Schwerpunkt der vorliegenden Arbeit war die Entwicklung einer neuen Methode (QSTAR) zur Vorhersage von freien Energien. Diese Methode ist fähig die quantenmechanische Natur von Atomkernen explizit zu berücksichtigen, was in bisherigen Simulationen für freie Bindungsenergien in biomolekularen Simulationen vernachlässig wurde. Es kann angenommen werden, dass die quantenmechanische Delokalisation der Atomkerne in solchen Systemen eine erhebliche Rolle spielen kann, vor allem aufgrund der Wasserstoffatome, welche zu den Atomarten mit den stärksten Quantendelokalisationen gehören. Um diese nuklearen Quanteneffekte zu berücksichtigen, wurde der Feynman'sche Pfadintegral Formalismus verwendet und synergetisch mit einem neuen alchemischen Transformationsschema kombiniert. QSTAR stellt auch den ersten direkt anwendbaren Einfach-Topologieansatz für Elektronenstrukturmodelle (ESMs) zur Verfügung. Des Weiteren wurde ein erweitertes alchemisches Schema für relative freie Bindungsenergien entwickelt, um das van der Waals Endpunktproblem zu umgehen. QSTAR und die alchemischen Schemen wurden in HyperQ implementiert, einer neuen skalierbaren Software, welche in der Lage ist, Simulationen der freien Energie automatisch durchzuführen.

Die meisten ESMs werden mit zunehmender Größe des Systems schnell prohibitiv teuer. Dies ist eine Einschränkung, welche mit QM/MM Methoden umgangen werden kann. Um QSTAR mit ESMs on auf biomolekulare Systeme anwenden zu können, wurde ein erweitertes QM/MM Schema entwickelt. Dieses Schema ist eine Methode für diffusive Systeme, welche auf Rückhaltepotentialen beruht. Diese erlaubt, QM Regionen von angepasster Form zu definieren, und ist rechnerisch sehr effizient. Die Methode wurde in einem neuen Klienten für i-PI implementiert, und sie ermöglicht es, Simulationen der freien Energie von biomolekularen Systemen mit einer sehr hohen Genauigkeit durchzuführen.

Einer der vielversprechendsten Ansätze um neue Hit-Kandidaten im CADD zu identifizieren sind strukturbasierte virtuelle Screenings (SBVS) welche Methoden zur Schätzung von freien Bindungsenergien nutzen. In dieser Arbeit wird argumentiert, dass je umfangreicher das virtuelle Screening ist desto besser die Ergebnisse im Sinne der Bindungsaffinität sowie der Erfolgsrate werden. Ein neues Workflowsystem, Virtual Flow, wurde entwickelt, welches erlaubt Aufgaben im Zusammenhang von SBVSs mit fast perfektem Skalierungsverhalten ohne praktisch relevante Grenzen bezüglich der Anzahl der Knoten/CPUs auf Computerclustern durchzuführen. Zwei Versionen wurden bisher implementiert, VFLP und VFVS, welche spezialisiert sind auf die Aufbereitung von Moleküldatenbanken sowie die virtuelle Screeningprozedur selbst.

Eine neues Wirkstoffentwicklungsprojekt wurde begonnen, um die neuen Methoden und Software unter realistischen Bedingungen zu testen und anzuwenden. Das Ziel dieses Projektes ist es, mögliche Inhibitoren für drei Regionen an der Oberfläche des Proteins EBP1 zu identifizieren, welche sich aller Erkenntnis nach auf Protein-Protein-Interfaces befinden und eine große Herausforderung darstellen. Drei zweistufige virtuelle Screenings wurden ausgeführt mit jeweils über 100 Millionen Molekülen. Nachfolgende Bindungsexperimente deuten auf eine relativ hohe Hitrate von über 30 % hin, und Fluoreszenzmikrosopie von mindestens einem Molekül weist auf gewünschte Effekte in Krebszellen hin. Weitere Projekte, in welchen die neuen Methoden und Software für die computergestützte Hit/Lead-Identifizierung angewendet wurden, beinhalteten die Zielmoleküle SHP2 (Krebs), eIF4E (Krebs), MED15 (Krebs), UL50/UL53 (Herpes), KEAP1 (Krebs), und die Peptidoglykane (Antibiotika).

Appendix D

Declaration of Authorship

(Selbstständigkeitserklärung)

Hiermit versichere ich, dass alle Hilfsmittel und Hilfen angeben worden sind, und auf deren Grundlage die Arbeit selbständig verfasst zu haben. Die Arbeit wurde in dieser oder ähnlicher Form in nicht schon einmal in einem früheren Promotionsverfahren eingereicht.

Berlin, Mai 2018

Christoph Gorgulla