Conclusions and perspectives
8.1 (Harder,) Better, Faster, Stronger

The large increase in computational power over recent years, together with the maturation of the molecular modeling field in the last decades, has unlocked unprecedented possibilities: simulations have become faster, longer, more detailed and more accurate than ever before. Simulations of large protein systems at the full atomistic detail are starting to approach the time scales of real live processes, beginning to match the fastest time scales of experiment. Further details are obtained by being able to perform calculations of large molecular systems at a higher level of theory, allowing to investigate the fine-grained detail of molecular interactions. On the other side of the modeling spectrum, fast methods focussing on simple properties such as interaction fingerprints make it possible to scan million-compound databases in the matter of hours in the search for drug candidates.

Although it may seem that a variety of the technical constraints to run real life molecular simulations has been lifted by improved methods and a vast amount of computer time, many challenges remain to be overcome in various parts in the field of molecular simulation. This has been one of the motivations of the current work which aims to improve models and methods used in protein-ligand simulations. In particular, we have studied ways to enhance the accuracy and efficiency of molecular dynamics simulations. For example, the level of detail at which traditional force fields operate contain many parameters and approximations that have been calibrated and chosen years or decades ago. These approximations were needed in times computational resources were scarce, and their effect on the quality of the results were evaluated at the time. The time scales on which simulations were performed were more of a constraint on the quality of the predictions than the approximations made in developing the force fields. With increased computational power and therefore longer simulations, this balance needs to be reconsidered, and the very basis of the treatment of interatomic interactions within the force field needs to be reevaluated (Section 8.2).

In addition, the gain in computer power has increased the ambition of scientists to model more and more complex systems. With the increase in size and the amount of components included in the simulation, the complexity of phase space of the system is steadily expanding. To fully (or at least sufficiently) explore phase space of such systems requires more than only largely extended simulation times. Currently, the longest simulation of a protein described at atomistic level extends in the range of milliseconds, but may still only cover a narrow part of phase space. From this it is obvious that methods are required to (i) further accelerate and extend the exploration of phase space, and (ii) make more effective use of available computer power (Section 8.3).

Finally, a new need has emerged from recent developments in molecular simulation techniques and in computational resources. To harness the available computational power, the diversity of facilities, and the multitude of user profiles and requirements, modeling methods need to be proposed that are very accessible.
Conclusions and perspectives

and make use of computational power without the complexity of handling numerous interfaces and settings. The setup of biomolecular calculations is ever more required to be automated such that non-expert users can have access to the methods without having to compromise on the quality of the predictions (Section 8.4).

8.2 The more detail the better?

The complexity of accurately parametrizing atomic force fields is partly due to the lack of transferability of subsets of parameters. In order to accurately reproduce the relevant physico-chemical properties based upon which the force field is parametrized, it is often needed to distinguish subcases of interactions, for which e.g. secondary repulsive van der Waals parameters are defined, or alternative charges that depend on the environment. This lack of transferability and added level of complexity of parameter sets are typically related to the difference in electronic interactions in either polar or apolar environments, and the ability of the modeled (macro-)molecule to adapt to changes in the environment. It is suggested that including polarizability in atomic force fields might be the answer to this problem. Originally, electronic polarization effects have been excluded from force fields due to the associated large computational cost, and therefore this assumption deserves to be reevaluated now sufficient computational power has been unlocked.

A major challenge in developing such polarizable force fields is assigning proper values to the new parameters to be introduced, i.e. the atomic polarizabilities. In Chapters 2 and 3, we have used a QM/MM approach to evaluate appropriate (condensed-phase) values for the atomic polarizability in an accurate way, such that the obtained values for the polarizabilities are directly suited for implementation in a polarizable force field intended for biomolecular simulation. Through an approach that combines quantum detail and molecular dynamics sampling, we show that condensed-phase values for polarizability can be computed, and that these values can directly be implemented into a force field. Chapter 3 additionally evaluates at which level of detail the solvent surrounding the quantum chemically treated solute should be included in our determination of the atomic polarizabilities. This study shows that using MM point charges provides excellent results at a modest computational cost, compared to executing the same procedure using a QM/QM description of the solute and solvent. In addition, we show that solute-solvent Pauli repulsion can only partly account for the need to scale down condensed phase polarizabilities when compared to gas-phase values.

In Chapter 4, the obtained values for condensed-phase atomic polarizabilities for alcohol solutes were used in the parametrization of a polarizable force field for linear alcohols. Faced with the challenge of the limited transferability of parameters within the alcohol series and between alcohol and hydrocarbon molecules, a generic (analytical) approach to force field optimization was designed. This analytical approach could be used to accurately predict pure-liquid properties of
a vast amount of parameter sets, and to guide calibration of a minimal parameter set that reproduces the target physico-chemical properties within experimental accuracy. The optimized parameters are generic (applicable to all alcohol molecules considered without the need for special interaction parameters) and transferable (accurate predictions for pure liquid properties and free energies of hydration for butane were made using these alcohol parameters).

Chapters 2 to 4 demonstrate that polarizable force fields can be developed using the proposed QM/MM approach to compute atomic polarizabilities. It should be noted that especially parametrization of polar functional groups benefits from the QM/MM estimations. Atomic polarizabilities for apolar groups, which are influenced to a limited extent by the electrostatics of the surrounding solvent, were found to hardly deviate from their gas-phase estimates. Using QM/MM polarizabilities, and after scaling point charges down to account for the added electrostatic term, the van der Waals interactions can be optimized to yield new parameters for a polarizable force field when starting from an available, well calibrated non-polarizable force field such as GROMOS 53A5 or 53A6. Preliminary results show that this strategy could be applied in a systematic way to various model compounds, making it possible to calibrate a full set of amino acid parameters in near future. The work presented in Chapters 2 to 4 shows that a relatively simple implementation of electronic polarization (with isotropic atomic polarizabilities, linear response and with no interactions between neighboring induced dipole moments) is sufficient to yield accurate results in terms of dielectric permittivity, heat of vaporization, density and free energy of hydration of representative model compounds, which are key physico-chemical properties that should be well described by a polarizable (biomolecular) force field. Combining the methods to derive condensed-phase polarizabilities described in Chapter 2 and the analytical approach to sample parameter space described in Chapter 4, it is now possible to recalibrate parameters for new (biomolecular) building blocks in an efficient, transparent and transferable way. In future this will allow to generate a fully polarizable force field suited for protein simulations.

Parallel to computing atomic polarizabilities and optimizing parameters for a polarizable (alcohol) force field, the nature of the van der Waals interactions were studied in more detail in Chapter 5. Starting from the idea that the attractive term of the Lennard-Jones potential is directly related to the induced dipole moment interactions and therefore related to atomic polarizabilities, we evaluated the benefit of obtaining van der Waals parameters directly from our QM/MM derived atomic polarizabilities using the Slater-Kirkwood equation. Being able to analytically derive this parameter would have a two-fold benefit: (1) it would allow for simple and accurate determination of part of the force field parameters, thereby reducing the force field optimization effort greatly, and (2) the method would become suited for automatic parametrization of novel molecules, making it possible to obtain in a simple way accurate and transferable parameters for new functional groups. Although several insights into the nature of the force field parameter space for small alkanes were obtained, the new set of van der Waals parameters was not able to improve on available force field
parameters that typically violate the Slater-Kirkwood relation when starting
from the QM/MM determined polarizabilities. Using the common force field
implementation of van der Waals interactions (i.e., Lennard-Jones potential, no
polarizability, and united-atom definition), the previously reported parameters
operate to yield the most accurate description of relevant physical properties
when compared to experiment. Combining the novel derived attractive van der
Waals parameters with different implementations of the Lennard-Jones potential
would deserve further investigation before reaching a final conclusion about the
value of directly determining van der Waals parameters from (condensed-phase)
polarizabilities. It should also be considered that although the polarizability
is not directly the key to a new set of attractive van der Waals parameters,
the magnitude of the polarizability is still expected to be correlated with this
parameter, and could therefore still be used in (automatically) reevaluating force
field parameters, such as undertaken by the Automated Topology Builder (ATB)
developers team.

8.3 Faster sampling for more accurate predictions

Being able to improve the accuracy of force field parameters will allow to
improve the representation of interactions between e.g. new molecules and their
protein environment. Calibrating these parameters in a generic way makes it
accessible to large-scale automated affinity prediction using approaches based on
molecular dynamics (MD) simulations. The efforts described in Chapters 2 to 5
directly contribute to increasing the accuracy of the interaction description, to
the development of approaches for parametrization, or to the automatization of
force field parametrization. The possible application of automatically generated
parameters is illustrated in Chapters 6 and 7, where potentially large-scale
free energy calculations are used to predict affinities between novel (drug-like)
compounds and complex protein structures.

The complexity in terms of sampling that comes with large and/or flexi-
ble molecular systems has to be addressed by more efficient means than only
substantially increasing simulation length. For now, even the most powerful
computers are still far from able to explore all relevant parts of phase space of
such complex systems, which are often well separated by (free) energy barriers.
Using the framework of the iterative Linear Interaction Energy (LIE) method
described in Chapter 6, a diverse set of complex protein-ligand conformations are
generated, thereby externalizing the sampling requirement. The MD simulation
is then only intended for relaxation and local sampling, while through running
multiple parallel simulations, various parts of phase space are explored. Beside
of making the free energy prediction more accurate in this way, the two other
main features of this approach are both related to a gain in efficiency: (1) only
short simulations are needed, since relaxation and local sampling is accessible
within limited time, and (2) the simulations starting from the various complex
conformations can be run independently and in parallel, which does not as such
reduce the total computational cost, but significantly reduces the time span
between the query and the result.

In the iterative LIE approach, the part of conformational space covered through sampling is not anymore defined by the transitions within one simulation, but depends on the quality of the methods to generate the conformations from which the independent simulations are started. In Chapter 6, methods are devised to create a subset of complex coordinate sets that are as diverse as possible for a test case, aryloxypropanolamine-like ligands binding to the flexible Cytochrome P450 2D6 protein. It is shown that the affinities can be predicted in an accurate and efficient way, and that parts of phase space that do not contribute to the total binding affinity have no negative impact on the calculated binding affinity. This is an important element in the iterative LIE methodology, since there is generally no information available allowing to discern a priori the conformations located in relevant parts of phase space from those in parts of phase space that are unlikely. In the applied iterative LIE approach, this is a posteriori resolved by a Boltzmann weighting scheme.\(^{75}\)

It is possible to fully automate the approach described in Chapter 6 (provided high quality ligand topologies can be automatically generated). The automated work flow takes user requests in the form of a database of ligands, sets up and runs MD simulations, uses the precalibrated LIE model to calculate weights to be attributed to the free energy of binding from the parallel simulations of every ligand, and yields for every ligand in the database an affinity prediction. This serves as the intended framework within the eTOX project, in which toxicological endpoints are predicted, such as the affinity of novel drug compounds for off-target proteins, based on a diverse set of (fast) automated methods. While using the automated setup presented in Chapter 6 it was noticed that in specific cases, ligands had sufficient possibilities to change binding conformation, even during relatively short simulations. In principle, this should not be an issue, as long as such conformational changes can be considered as a minimization or relaxation step within the same local part of phase space. If this is not the case, there is a risk of different simulations showing overlap in the parts of phase space covered. This is detrimental for the calculated affinity because a basic assumption of the iterative LIE scheme is that all simulations cover distinct parts of phase space, and therefore can be individually integrated in terms of (weighted) average interaction energies.\(^{228}\) To overcome issues in possible phase space overlap between simulations, a method to detect conformational transitions of the complex during simulation was introduced in Chapter 7, that could be included in the automated setup. Using noise reduction and fitting tools, transitions are detected, and automated decisions can be connected to these transitions to include or exclude specific parts of the simulation from the affinity calculation. Implementing this method to detect conformational transitions also carries the benefit of allowing to reduce the total simulation time if the system is quickly relaxed and the local space is explored within a short amount of time. Making the length of the simulation depending on local sampling and not assigning a fixed simulation length would represent a major speedup, and a more efficient use of computer resources. Next to showing that a 60% decrease of the total computation time could be achieved, it is demonstrated that the method
Conclusions and perspectives

contributes to improving the predictive quality of the models, even though the extent of the improvement depends on the nature of the system (and mainly its propensity to cross energy barriers during short simulations).

8.4 Programming for the masses

The methodological developments described in Chapters 2, 5, 6 and 7 do not only contribute to improve the quality of in silico models and to enhance the accuracy of predictions, but also aim on making molecular simulation, binding affinity prediction and force field calibration more broadly available. The QM/MM work flow to compute atomic polarizabilities described in Chapters 2 and 3 has been implemented into the ATB pipeline, making it possible to compute atomic polarizabilities for a molecule with only an ATB molecule ID or a coordinate file. For now the QM/MM pipeline within the ATB is only available for developers due to software licensing restrictions and the computational cost of running these calculations. It should be noted however, that the whole setup can be performed without any user intervention, from the QM energy minimization, to solvation, MD simulation, QM determination of solute electron densities, the calculation of molecular electrostatic potentials, the fitting of induced atomic dipole moments, the derivation of atomic polarizabilities, and even a graphical representation of the results. This means that minimal user expertise is needed to obtain the relevant results. In a very similar way, the automatization efforts of binding free energy predictions as presented in Chapters 6 and 7 allows for non-expert users to get accurate and relatively fast predictions. Proposing the tools through a web-interface such as eTOXsys even allows for usage by scientists in different fields than computational chemistry. An experimentalist for instance could use the affinity prediction tool to prioritize experiments in the lab and cut costs by running only experiments with enhanced odds of success.

From a scientific perspective, when working on the automatization, it is crucial to include routines intended at detecting unreliable results (that are either unphysical, or with a low predictive value). This becomes even more important when targeting an audience with limited modeling expertise, when the machinery is distributed as a black box, or when the tools are used in (semi-) high throughput. Improving predictive methods therefore does not only involve optimizing the machinery, predicting interactions more accurately or improving sampling, but also the development and design of efficient methods to assess the reliability of the simulations and of the output. Further improvements on the pipeline presented in Chapter 6, aside from extending the range of systems and models, is the feedback on the applicability of the model for the compounds for which affinities are predicted. Using a given (local) model, any prediction can be made, but if the model is not applicable, the value of the prediction is limited. The information on how well the model is expected to perform is therefore just as important as the result itself. The development of strategies to define the applicability domain of the LIE models are currently ongoing, which are based on detecting overlap in protein-ligand interaction fingerprints.
in the simulations of the compounds.\textsuperscript{256} Other criteria for applicability domain assessment may include molecular similarities\textsuperscript{245} and distributions of calibrated data and interaction energies (\textit{cf.} Section 6.3.2). Automated identification of the applicability domain in an automated approach also allows to evaluate the cost (in terms of accuracy) of using a global model, and when needed directs the choice towards a combination of tailored local models.

While the setup to compute polarizabilities might only become available for large audiences in a distant future, the one-step perturbation strategy described in Chapter 5 to reparametrize force fields are currently being implemented into the ATB, and should become a part of the publicly available tools once optimized. This would represent a major improvement on the already successful tools provided in the current version of the Automated Topology Builder. The Linear Interaction Energy models presented in Chapters 6 and 7 are being integrated into the \textit{eTOXsys} framework, which is intended to be broadly used to predict possible toxicities in early stages of drug development. The tools and models are made available to all partners within the \textit{eTOX} consortium, including various universities and a majority of European “big pharma” companies.