Efficient Biomolecular Computations Towards Applications in Drug Discovery

Negin Forouzesh

Dissertation submitted to the Faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science and Applications

Alexey V. Onufriev, Chair
David R. Bevan
Holger Gohlke
Liqing Zhang
Sharath Raghvendra

June 17, 2020
Blacksburg, Virginia

Keywords: Free Energy Calculation, Molecular Simulation, Drug Discovery.
Copyright 2020, Negin Forouzesh
Atomistic modeling and simulation methods facilitate biomedical research from many respects, including structure-based drug design. The ability of these methods to address biologically relevant problems is largely determined by the accuracy of the treatment of complex solvation effects in target biomolecules surrounded by water. The implicit solvent model – which treats solvent as a continuum with the dielectric and non-polar properties of water – offers a good balance between accuracy and speed. Simple and efficient, generalized Born (GB) model has become a widely used implicit solvent responsible for the estimation of key electrostatic interactions. The main goal of this research is to improve the accuracy of protein-ligand binding calculations in the implicit solvent framework. To address the problem (1) GBNSR6, an accurate yet efficient flavor of GB, has been thoroughly explored in the context of protein-ligand binding, (2) a global multidimensional optimization pipeline is developed to find the optimal dielectric boundary made of atomic and water probe radii specifically for protein-ligand binding calculations using GBNSR6. The pipeline includes (3) two novel post-processing steps for optimum robustness analysis and optimization landscape visualization. In the final step of this research, (4) accuracy gain the optimal dielectric boundary can bring in practice is explored on binding benchmarks, including the SARS-CoV-2 spike receptor-binding domain and the human ACE2 receptor.
Drug discovery is one of the most challenging tasks in biological sciences as it takes about 10-15 years and $1.5-2 billion on average to discover a new drug. Therefore, efforts to speed up this process or lower its costs are highly valuable. Computer-aided drug design (CADD) plays a crucial role in the early stage of drug discovery. In CADD, computational approaches are used in order to discover, develop, and analyze drugs and similar biologically active molecules, such as proteins. Proteins are an important class of biological macromolecules that perform their functionality mainly through interactions with other molecules, for example, binding to small molecules so-called ligands. Thorough understanding of protein-ligand interactions is central to comprehending biology at the molecular level. In this study, we introduce and analyze a computational model used for protein-ligand binding free energy calculations. A global multidimensional optimization pipeline is developed to find the optimal parameters of the model, particularly those parameters involved in the dielectric boundary. In order to examine the robustness of the optimal model to unavoidable perturbations and uncertainties, virtually inevitable in any complex system being optimized, a novel robustness metric is introduced. Finally, the robust optimal model is tested on protein-ligand benchmarks, including a complex related to the novel coronavirus. Results demonstrate relatively higher accuracy in terms of binding free energy calculations compared to reference models.
Dedication

This thesis is dedicated to my beloved husband and parents who supported me unconditionally during this journey.
Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor Prof. Alexey V. Onufriev for his ceaseless support, thoughtful guidance, and wise mentorship. His vast patience and constant encouragement made it possible for me to earn this Ph.D. degree. The only way I could imagine to give back this support is to become such a wonderful advisor for my future students. Besides him, I would like to thank the rest of my advisory committee: Prof. David R. Bevan, Prof. Holger Gohlke, Dr. Liqing Zhang, and Dr. Sharath Raghvendra for their encouragement, flexibility, and insightful comments.

I am grateful to all of those with whom I have had the pleasure to work during the Ph.D. program, particularly my labmates at the Laboratory for Theoretical and Computational Biophysics: Dr. Igor Tolokh, Dr. Saeed Izadi, Dr. Abhishek Mukhopadhyay, and Dan Folescu. In addition, I thank Prof. Layne T. Watson for his expert advice and technical consultations.

Finally, I could not have made it to this point without the help and love from my family, friends and teachers throughout the course of my education. For all the support that I have received, I am eternally grateful.
# Contents

List of Figures | ix
---|---
List of Tables | xvi

1 Introduction | 1

2 Implicit Solvent Model | 6
   2.1 Introduction | 6
   2.2 Methodology | 8
   2.3 Analysis | 16
      2.3.1 Practical Grid Spacing | 16
      2.3.2 Grid-Based vs. MSMS-Based Molecular Surface Implementations of GBNSR6 | 20
      2.3.3 Considerations for Further Speedup | 23

3 Multidimensional Global Optimization | 25
   3.1 Introduction | 25
   3.2 Materials and Methods | 28
      3.2.1 The Electrostatic Component of Binding Free Energy | 28
      3.2.2 Implicit Solvent Model | 30
3.2.3 Objective Function .................................................. 31
3.2.4 Sampling Around the Minimum Points ............................ 32
3.2.5 Data Sets for Training and Test ..................................... 33
3.2.6 VTDIRECT95: Global Multidimensional Optimization Method .... 34
3.2.7 Numerical Estimate of $\langle E \rangle$ .................................. 37
3.3 Bounds on Physically Meaningful Values of Atomic Radii ......... 37
3.3.1 Radial Distribution Function ....................................... 37
3.3.2 Radii Constraints ...................................................... 39

4 Robustness Analysis and Visualization ................................. 41
  4.1 Introduction ............................................................ 41
  4.2 Proposed Metric of Optimum Robustness ......................... 44
    4.2.1 Application to Optimization of Atomic Radii ............. 51
    4.2.2 Visualizing the Optimization Landscape .................. 53
    4.2.3 Exploring the Objective Function Landscape ............. 56
  4.3 Optimized Parameters of the Dielectric Boundary Show Promise . 60

5 Testing and Application to Protein-Protein Binding .................. 63
  5.1 Disclaimer ............................................................ 63
  5.2 Prelude ............................................................... 64
  5.3 Introduction ........................................................ 65
5.4 Methods and Materials ........................................ 68
  5.4.1 Binding Free Energy Decomposition ...................... 68
  5.4.2 MMPB/SA Free Energy Methodology ...................... 70
  5.4.3 Solvation Free Energy ..................................... 72
  5.4.4 Gas-Phase Energy ......................................... 73
  5.4.5 Configurational Entropy .................................. 73
  5.4.6 Structure Preparation and MD Simulations ............... 73
5.5 Results and Discussion ....................................... 74
  5.5.1 MM/GBSA on Ras-Raf ..................................... 74
  5.5.2 MM/GBSA on SARS-CoV-2 S RBD and ACE2 ............. 75

6 Conclusion .......................................................... 79

Bibliography .......................................................... 83
List of Figures

1.1 The pharmaceutical drug development process. It takes about 15 years on average to discover a new drug out of thousands of initial candidates. The application domain of this research is in the first stage: drug discovery.

1.2 Water models. Left: Implicit solvent model. Water is represented by its bulk dielectric permittivity $\epsilon$. Right: Explicit solvent model. A solute molecule is surrounded by individual water molecules.

1.3 Solvent excluded surface (SES) [143] exemplified for a "molecule" of two atoms in a box of water. SES is shown as the red boundary, defined as the locus of the contact points (connected by circle arcs at contact discontinuities) of water probe (red circle) when it is rolled over the molecule (gray circles).

2.1 The thermodynamic cycle used for computing the binding free energy. Water environment is shown in blue, and vacuum is in white.

2.2 Finite-difference discretization of an abstract molecule. Blue spherical surface shows the ideal SES that represents the dielectric boundary, and green lines are the approximating square surface elements. The background black mesh depicts the uniform Cartesian grid.

2.3 Spherical surface element $\delta S_k$ (blue) and the corresponding square surface element (green). For a point charge at $O$, the electric field flux through both elements is the same.
2.4 Effective Born radii calculation for atom $i$ via the "field-view" method, Eq. 2.6 and Eq. 2.7. The black circle shows the atom for which the effective Born radius is calculated, and the yellow circle shows an atom or the probe corresponding to the surface element.

2.5 Error in computing $\Delta \Delta G_{pol}$ due to finite grid resolution $h$ relative to $h = 0$ Å extrapolation. See Methods for more details.

2.6 Correlation between $\Delta \Delta G_{pol}$ computed by GBNSR6 and the numerical PB reference MIBPB ($h=0.5$ Å). Red line $x = y$ indicates a perfect match.

2.7 GBNSR6 cpu time of computing $\Delta G_{pol}$ for different grid resolutions.

2.8 Comparison of the elapsed time for computing $\Delta G_{pol}$ using the original MSMS-based GBNSR6 and the current grid-based implementation.

2.9 Comparison of the molecular surfaces computed by MSMS and grid-based methods for 1kxq complex. (a) the 3D structure of 1kxq complex, including protein (green) and ligand (gray). The internal cavities are displayed in orange. (b) dotted circles in the figure show internal cavities that are captured using the grid-based surface but are ignored by MSMS with the default setting. The black-square dots inside the dotted circles show atoms in the vicinity of the internal cavities. Shown is a 6 Å wide slice of the molecular surface. For visualization purposes grid spacing of $h = 0.3$ Å was used.

2.10 Comparison of the inverse of effective Born radii computed by MSMS and grid-based methods for 1kxq.
2.11 Time analysis of the grid-based GBNSR6 execution \((h = 0.5 \text{ Å})\). Here \(t_{\text{grid}}\) is the time for setting up the grid, \(t_{\text{rad}}\) is the time for estimating the effective Born radii, and \(t_{\text{pair}}\) is the time for computing the sum of all pairwise charge-charge interactions in Eq. 2.1.

3.1 Latin hypercube sampling (LHS). This example shows how LHS generates random sample points in a 2-dimensional space so that there exists one and only one sample point in each row and column.

3.2 Function evaluations performed by DIRECT after 0, 1, 5 and 10 iterations. The objective function values are illustrated via the contours and the corresponding color bar on the rightmost panel. Comparing the first and second graphs on the top shows how DIRECT divides a 2-dimensional box after one iteration. On the bottom right figure, DIRECT finds the global minimum at \((0.9,0.3)\) after 10 iterations. It also explores a large domain and evaluates the function near the local minimum at \((0.4,0.3)\).
3.3 Solvent excluded surface (SES) exemplified for a "molecule" of six atoms. SES is shown as the purple boundary, defined as the locus of the contact points (connected by circle arcs at contact discontinuities) of water probe (white circle) when it is rolled over the molecule (gray circles). An example of radial distribution function of atom-(water oxygen) obtained for atom type \(i\) from molecular dynamics simulations of various molecules containing that atom type is shown in \(g_{i-ow}(R)\) plot to the left of the schematic. Each color in \(g_{i-ow}(R)\) plot represents a separate instance of atom type \(i\); the bounds \((R_{\text{min}}, R_{\text{max}})\) are computed as one standard deviation about the mean (shown as the double-headed red arrow) of the RDF first peak, inferred from the combined data of the first peaks for all the instances of the atom type \(i\). These bounds are used to constrain \(\rho_i + \rho_w\) for simultaneous optimization of \(\rho_i\), atomic radii of atom \(i\), and \(\rho_w\), water probe radius.

4.1 Depth vs. width trade-off in choosing the preferred optimum. The shallow and wide local minimum point on the right may be more robust to perturbation compared to the narrow but deeper global minimum solution on the left. When the difference \(\Delta\) between the objective function at the minimum points is close to the problem-specific range in which two optima are practically indistinguishable, one may intuitively conclude that the wider local minimum solution should be preferred for a complex system with inevitable uncertainties in replication of its parameters. The goal of this work is to formulate and test a quantitative criterion for such preference.
4.2 Robustness analysis of two examples. Left panel shows two equally wide wells, which are similarly robust to small perturbations of the parameters. The right panel shows a totally different behavior of the objective function, where the wide local minimum is more robust to perturbations than the narrow global minimum.

4.3 Deviation from the optimal solution \( (X^*) \) given a new data set. Left: changes in objective function value at \( X^* = 0 \) (\( \delta \)) is proportional to \( b^2 \). Right: estimation of the standard deviation of \( b \) when several test sets are given.

4.4 Projection of OPT1 (global min) and OPT4 (local min) onto different radii coordinates. Left panel shows the behavior of OPT1 objective function projected onto \( \rho_W, \rho_C, \rho_H, \rho_N \) and \( \rho_O \) within the sampling box and in the physical bound proposed in Eq. 3.3. Right panel shows similar graphs for OPT4. Radii (x coordinates) have different ranges in order to keep the objective function values (y coordinates) in a same range, which is \( 2kT \) form OPT1 value.

4.5 Procedure for creating a distance plot, exemplified. OPT1 (global minimum) and OPT2 (local minimum) are selected in this demonstration. The large sampling box, shown in solid black, covers the space between the smaller sampling boxes (dashed red rectangles) around OPT1 and OPT2. These two smaller boxes are found by applying the sampling algorithm explained in the "Materials and Methods". For each sample point \( x \) in the large box, 5-dimensional Euclidean distances \( d_1 \) and \( d_2 \), from OPT1 and OPT2 (shown as stars) to \( x \) are calculated, and the corresponding objective function value is illustrated on the distance plot, shown in Fig. 4.7.
4.6 Boxplots of deviations in the objective function values around OPT1 (global minimum) and OPT2, OPT3, OPT4 and OPT5 (local minima) within the sampling box. Outliers are shown in circles. .............................. 56

4.7 Distance plots. Shown are only those sample points whose objective function values are within the range of $kT$ from OPT1. The 1000 and 5000 sample-point scenarios are shown in orange and blue, respectively. ....... 57

4.8 $kT$-connectivity graph. Vertices represent the global minimum point (OPT1) in the center, and local minimum points around it. An edge between OPT1 and another vertex indicates that it is possible to move between the two minima without exceeding a pre-defined threshold of the objective function, in our case $kT \sim 0.6 \text{ kcal/mol.}$ ................................. 58

4.9 Re-balancing of the training and test sets, with TIP3P explicit solvent model as the reference. Left: the current partitioning method partitions the whole data set of 15 small protein-ligand complexes into the training and test sub-sets with a similar distribution of $\Delta\Delta G_{pol}$. Training: RMSE=3.94 $\text{kcal/mol}$ and $r^2=0.76$. Test: RMSE=6.62 $\text{kcal/mol}$ and $r^2=0.37$. These results are obtained using the existing global optimum radii (OPT1). Right: New partitioning puts the single outlier (1b11) in both the training and test sets. Training: RMSE=4.39 $\text{kcal/mol}$ and $r^2=0.68$. Test: RMSE=4.98 $\text{kcal/mol}$ and $r^2=0.57$. These results are obtained using a new global optimum radii (not shown here)found by VTDIRECT95. ...................... 62

5.1 Binding scheme of the SARS-CoV-2 spike protein to the ACE2 human receptor. .......................................................... 67
5.2 Binding a ligand (shown in yellow) to a protein receptor (shown in purple) in a box of solvent (shown in blue) releases binding free energy of $\Delta G_{\text{bind}}$. A negative sign of $\Delta G_{\text{bind}}$ indicates that spontaneous binding occurs, the magnitude of $|\Delta G_{\text{bind}}|$ characterizes the binding strength (affinity).

5.3 The thermodynamic cycle used here to estimate binding free energy of a protein-ligand complex in the solvent.

5.4 MMPB/SA flowchart. The initial structure of the complex is solvated using a water model. An MD simulation is run from which a relatively large number snapshots are extracted. The average binding free energy of the snapshots is assigned as the binding free energy of the system.

5.5 Backbone RMSD of SARS-CoV-2 S RBD and ACE2 complex, relative to its experimental crystal structure, along the 50 ns production trajectory.

5.6 Truncation of SARS-CoV-2 S RBD used in the entropy estimate. The spike protein is in cyan, and the ACE2 receptor is in green. Left: original complex. Right: truncated complex. A pair of atoms on the binding interface that are 8.8 Å apart is shown in a solid red segment to illustrate the length scale.
## List of Tables

2.1 Dataset with wide range of structure sizes .................................................. 9

2.2 Deviation of $\Delta \Delta G_{pol}$ values (kcal/mol) of the set of small protein-ligand complexes from those rotated around different axes each through $\pi/4$. The first column shows the results when complexes are rotated $\pi/4$ around $x$ axis, followed by $\pi/4$ rotations around $y$ and $z$ axes, respectively. The last column shows the results when $\pi/4$ rotations around $x, y$ and $z$ axes are applied together. .............................................................. 20

3.1 The list of 11 molecules used in this work to compute the solute atom to solvent (TIP3P) oxygen radial distribution function. ............................................. 38

4.1 The lowest five optimum parameter vectors found by VTDIRECT95. Radii are in Å and objective function values of the training set, $E_{train}$, are in kcal/mol. 51

4.2 Robustness analysis of the lowest five optimum parameter vectors found by VTDIRECT95. $\langle E_{train}^{1000} \rangle$ and $\langle E_{train}^{5000} \rangle$ show the result of ranking using Gaussian distribution as the weighting function, while the last column, $\langle E_{train}^{5000} \rangle'$, uses $P'(X)$ defined in Eq. 4.4, all are in kcal/mol. ................................. 53

4.3 The accuracy (RMSE to the explicit solvent reference, Eq. 3.2) of calculating $\Delta \Delta G_{pol}$ values using the proposed optimal radii (OPT1) and two other popular sets of atomic radii. Radii are in Å and RMSE value of the training and test sets, $E_{train}$ and $E_{test}$, are in kcal/mol. ................................. 61
5.1 Two sets of atomic radii in Å used in this study. ................. 72

5.2 MMGB/SA results on RAS-RAF. All the components are in \( \text{kcal/mol} \). The experimental value is from isothermal titration calorimetry [149]. ......... 75

5.3 Binding entropy estimation of the truncated SARS-CoV-2 S RBD and ACE2 complex. The values are in \( \text{kcal/mol} \). ................................. 77

5.4 MMGB/SA results on the SARS-CoV-2 S RBD and ACE2 complex. All the components are in \( \text{kcal/mol} \). Experimental value derived from a fit to surface plasmon resonance sensogram [186]. ......................... 78
Chapter 1

Introduction

Drug discovery is one of the most challenging tasks in biological sciences as it takes about 10-15 years and $1.5-2$ billion on average to discover a new drug \cite{62}, see Fig. 1.1. Efforts to speed up this process or lower its costs are highly valuable particularly in dealing with fast-growing pandemics such as COVID-19. Computer-aided drug design (CADD) plays a crucial role in the early stage of drug discovery. In structure-based drug discovery, in silico, accuracy and computational efficiency of the binding free energy prediction of small molecules to biomolecular targets are of paramount importance for high throughput screening of potential drug candidates \cite{58, 86, 156, 192}. However, fast and accurate computational prediction of binding free energies continues to be challenging \cite{29, 109, 113, 144, 146, 180, 185, 189, 198},

Alchemical methods \cite{1, 24}, simulate changes in the free energy in a pathway that sometimes reflect non-physical properties or literally "alchemy". The required sample points along the pathway are generated via Monte Carlo or molecular dynamics (MD) simulations. Such simulations are computationally expensive specifically when it comes to large-scale virtual screening of potential drugs. Some of the popular methods in this class are thermodynamic integration (TI) and free energy perturbations (FEP) \cite{110}. Remarkably more efficient, end-point free energy methods ignore the whole pathway and estimate free energy on an ensemble of uncorrelated snapshots collected from an equilibrated MD simulation, \textit{i.e.}, a sample of snapshots extracted from the trajectory in final
Chapter 1. Introduction

Figure 1.1: The pharmaceutical drug development process. It takes about 15 years on average to discover a new drug out of thousands of initial candidates. The application domain of this research is in the first stage: drug discovery.

states. Molecular mechanics Poisson-Boltzmann surface area (MMPBSA) and molecular mechanics generalized Born surface area (MMGBSA) [51, 177, 178] are of the most popular methods in this category. They are mainly applicable in docking projects where a quick yet relatively accurate estimate of binding affinities is required [75]. Leading docking software, for instance, AutoDock Vina [172] and DOCK [4], rank the feasible poses of a ligand in a binding pocket based on a scoring function in which binding affinity plays an important role. End-point free energies can improve the accuracy of these scoring functions on-the-fly. Recently, MMGBSA was employed to improve the accuracy of AutoDock Vina and Dock in the Drug Design Data Resource (D3R) Grand Challenge 4 (GC4) [151].

The outcomes of binding free energy calculations depend strongly on the molecular modeling technique, particularly, on how well the solvent effects are approximated [114, 144]. There are two major categories of solvent models used in this field [138]: explicit and implicit. Within the explicit solvent framework, the mechanistic detail and the energetic effect of every single water molecule is explicitly considered, which in turn results in
considerable computational cost. The implicit solvent model \([9, 26, 55, 101, 162]\), which treats the solvent as a continuum dielectric with polar as well as non-polar properties of water, may often offer a good balance between accuracy and speed, see Fig. 1.2. Within this framework, the generalized Born model \([31, 38, 49, 73, 123, 129, 134, 165, 173, 175]\) is widely used due to its relative simplicity and efficiency \([135, 137]\). GB is fast in applicability domains but faces fundamental accuracy limitations that bound further adoption. Some of these limitations are fundamentally unavoidable consequences of the many approximations made by this conceptually simple models. However, further analysis of the GB parameters and optimization of those parameters in a physically meaningful framework brings the GB model to the next level of accuracy.

Figure 1.2: Water models. Left: Implicit solvent model. Water is represented by its bulk dielectric permittivity \(\varepsilon\). Right: Explicit solvent model. A solute molecule is surrounded by individual water molecules.

A key step in the GB model is the determination of the solute/solvent dielectric boundary (DB), a region of space over which the dielectric constant \(\varepsilon\) shifts from the value characteristic of the solute interior (e.g., \(\varepsilon = 1\) or 4) to that of the solvent, (e.g., 80 for water). Outcomes of implicit solvent calculations have proven to be extremely sensitive to details of the DB \([171]\). The dielectric boundary is determined by the radii of the atom types comprising the protein as well as the size of the water probe \([136]\). In the past, optimizations on DB for solvation free energies of small molecules were performed — the optimal DB minimized the deviation of the computed target from an accepted reference, either
experimental or estimated via explicit solvent [126, 127, 163, 168, 169, 188]. One potential technical issue with previously derived optimal radii is that the true global optimum may not have been found – even for small molecules, the corresponding optimization problem is highly demanding, textbook numerical approaches are unlikely to find the global optimum in a rugged, multidimensional landscape. While this issue may not be critical in practice if a "good enough" local optimum is found, it still leaves the question open of how well one can do in principle. Finding a true global optimum can point to limitations of the underlying physical theory, and thus prompt further development. For practical calculations, a much more important limitation of optimal radii based on small molecule hydration energies is that it is highly likely that parameters defining the DB that are optimal for small molecule calculations are not optimal for estimates of protein-ligand binding free energies [64, 65, 109], which is of paramount interest.

Figure 1.3: Solvent excluded surface (SES) [143] exemplified for a "molecule" of two atoms in a box of water. SES is shown as the red boundary, defined as the locus of the contact points (connected by circle arcs at contact discontinuities) of water probe (red circle) when it is rolled over the molecule (gray circles).

The goal of this research is to increase the accuracy of protein ligand-binding calculations using a flavor of generalized Born model, called GBNSR6, that has shown to be the most accurate representative of the class of GB [83]; yet inadequately accurate compared to the standard explicit/experimental references. This improvement fills the gap between the implicit and explicit solvent models, and paves the way for further employment of
GB models in computational pipelines where speed and accuracy are both required. Advantages of the optimal GB model is demonstrated by testing it on biomedically relevant problems, via prediction of protein-ligand and protein-protein binding free energies with applications in drug design. The key result of this work is a novel computational pipeline generally applicable to any multidimensional constrained optimizations, specifically studied for the dielectric boundary optimization in this paper. The core of this pipeline is GBNSR6 employed for binding free energy (ΔΔG) calculations. This method is introduced and analyzed in Chapter 2. The global optimization method, called VTDIRECT95, and the corresponding mathematical modeling of the dielectric optimization of GBNSR6 are studied in Chapter 3. A novel robustness metric and a new visualization method for studying the optimization outcomes are introduced in Chapter 4. Finally, the outcome of this pipeline, that is GBNSR6 coupled with the most robust optimal radii, is tested on practical protein-protein complexes in Chapter 5.
Chapter 2

Implicit Solvent Model

This chapter appeared as the reference [46].
Here we introduce a grid-based surface generalized Born model, called GBNSR6. This model is analyzed from accuracy and speed points of view. In the next chapters, GBNSR6 will be used for binding free energy calculations.

2.1 Introduction

Solvation free energy calculation plays a critical role in molecular design and computational drug discovery [86, 113, 156]. It is one of the major elements in calculating the binding affinity of biomolecular complexes, which is the quantity of interest in many practical applications [96, 102]. As one of the most sensitive measures of the balance in the solute-solvent interactions, solvation and binding free energies calculations [155] have been commonly used to evaluate the accuracy of molecular mechanics force fields [115, 116, 157].

The large number of discrete water molecules typically required in free energy estimates in the explicit solvent makes these calculations very demanding computationally, even for small systems [2, 50, 81, 116]. Implicit solvent model [12, 26, 55, 76, 101, 104, 105, 152, 162, 176] is a popular alternative approach that approximates the solute response to bulk solvent via a continuum medium with the average dielectric properties of water,
to achieve significant computational efficiency. Within the implicit solvation framework, the generalized Born (GB) models [17, 21, 28, 32, 37, 39, 42, 49, 54, 61, 66, 67, 73, 78, 82, 84, 95, 98, 123, 124, 130, 134, 135, 142, 148, 153, 161, 164, 165, 170, 174, 175, 184], including grid-based [16, 99, 152, 187] flavors, offer a reasonable compromise between accuracy and speed. Recently, we have developed a flavor of the GB model, GBNSR6 [3], which showed considerable promise in solvation energy calculations [2, 81]. The key ingredient of GBNSR6 model is the calculation of the effective Born radii via $r^{-6}$ (R6) integral [59], shown to bring appreciable accuracy gains [106, 117, 170]. In GBNSR6 the integration is performed over the solvent excluded (molecular) surface (SES) [143], which is the same type of surface commonly employed by the PB. For this purpose, an accurate and computationally facile representation of the molecular surface is necessary. Many algorithms have been proposed to represent the molecular surface. Numerical algorithms generally tessellate the surface or fill the interior volume with geometric objects to compute the SES [14, 103, 118, 160, 190]. In the generally fast analytical class, one well-known implementation is MSMS [150], which provides the triangulated form of the reduced surface of the molecule. MSMS has been shown to be highly efficient in computing the triangulated surface, however it is not an open source software, making it problematic to incorporate MSMS-based surface calculation into distributable software.

Here, we describe in detail, and test in the context of protein-ligand binding, a Cartesian grid-based molecular surface numerical algorithm GBNSR6 for calculating the effective Born radii (and the solvation free energy). The algorithm adapts to the R6 integration a "field-view" surface integration method previously implemented in free and open source AmberTools package [19] for finite difference PB calculations [15]. The rest of the paper is organized as follows. In Methods we introduce two datasets used
for the benchmarking, details of implicit solvent calculations, including the reference PB method, and the calculation of the binding free energy. We then describe the details of the grid-based molecular surface, and the field-view adaptation to the R6 surface integration. The analysis of the speed, accuracy and sensitivity of GBNSR6 to grid parameters are discussed in Results and Discussion, where we also present a comparison with the previous MSMS-based GBNSR6 implementation [2]. A summary of our findings is presented in Conclusion.

2.2 Methodology

Datasets. The proposed model is tested on two datasets: first, a set of 15 small protein-ligand complexes (described in [81]), which are used to test the accuracy. Second, a set of molecules of widely different sizes (see Tab. 2.1), which are used to test the speed. All of these structures were protonated via H++ webserver (http://biophysics.cs.vt.edu/) with the default setting [8].

The GB model. The polar component of the solvation free energy $\Delta G_{pol}$ is calculated using GBNSR6 [81]. The analytical linearized Poisson-Boltzmann (ALPB) model [159] is used to approximate $\Delta G_{pol}$, by the following formula:

$$\Delta G_{pol} \approx -\frac{1}{2} \left( \frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{out}} \right) \frac{1}{1 + \beta \alpha} \sum_{ij} q_i q_j \left( \frac{A}{f_{ij}^{GB}} + \frac{\alpha \beta}{A} \right)$$ \hspace{1cm} (2.1)

where $\epsilon_{in}$ and $\epsilon_{out}$ are the dielectric constants of the solute and the solvent respectively, $\beta = \epsilon_{in} / \epsilon_{out}$ and $\alpha = 0.571412$. $A$ is the electrostatic size of the molecule, which can be computed analytically, $q_i$ is the partial charge of atom $i$. The most common functional
Table 2.1: Dataset with wide range of structure sizes

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>#atoms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1brv</td>
<td>268</td>
</tr>
<tr>
<td>1byy</td>
<td>357</td>
</tr>
<tr>
<td>1eds</td>
<td>456</td>
</tr>
<tr>
<td>1eit</td>
<td>548</td>
</tr>
<tr>
<td>1ccm</td>
<td>648</td>
</tr>
<tr>
<td>1fdm</td>
<td>745</td>
</tr>
<tr>
<td>1brz</td>
<td>855</td>
</tr>
<tr>
<td>1ioj</td>
<td>948</td>
</tr>
<tr>
<td>1aps</td>
<td>1549</td>
</tr>
<tr>
<td>1ao8</td>
<td>2546</td>
</tr>
<tr>
<td>4eug</td>
<td>3553</td>
</tr>
<tr>
<td>7a3h</td>
<td>4578</td>
</tr>
<tr>
<td>1c0p</td>
<td>5566</td>
</tr>
<tr>
<td>1g6s</td>
<td>6487</td>
</tr>
<tr>
<td>1e6q</td>
<td>7819</td>
</tr>
<tr>
<td>1qks</td>
<td>8556</td>
</tr>
<tr>
<td>1kx5 (histones only)</td>
<td>15740</td>
</tr>
<tr>
<td>1kx5</td>
<td>25099</td>
</tr>
</tbody>
</table>

The form of \( f_{ij}^{GB} = \left[ r_{ij}^2 + R_i R_j \exp\left(-r_{ij}^2/4R_i R_j\right) \right]^{1/2} \) is employed where \( R_i \) is the so-called effective Born radius of atom \( i \), and \( r_{ij} \) is the distance between atoms \( i \) and \( j \). The dielectric constants are set to \( \epsilon_{in} = 1 \) and \( \epsilon_{out} = 80 \). In this work, the effective Born radii \( R_i \) are calculated by the following "R6" equation [3]:

\[
R_i^{-3} = \left( -\frac{1}{4\pi} \oint_{\partial V} \frac{r - r_i}{\| r - r_i \|^6} \cdot dS \right) \tag{2.2}
\]

where \( \partial V \) represents the molecular surface, \( dS \) is the infinitesimal surface element vector, \( r_i \) is the position of atom \( i \), and \( r \) indicates the position of the infinitesimal surface element. The ionic strength is set to 0.145 M to reflect physiological conditions. GBNSR6 includes a single additional (to the PB) parameter: a constant additive offset \( B \) to each inverse effective Born radius. The value of the offset is fixed for a given probe radius; we use
$B = 0.028 \text{ Å}^{-1}$ as proposed earlier [117] for the probe radius of 1.4 Å used in all of the calculations reported here. The above probe radius and the Bondi set of intrinsic radii [13] are used to determine the surface of the molecule. These intrinsic radii are used for all of the calculations.

Numerical solutions of the Poisson equation (referred to as the "PB" here) are commonly used [41] to evaluate the accuracy of the GB. The traditional rationale for this testing strategy is that the GB model shares the same basic physics of two-dielectric continuum electrostatics, and a number of further approximations [135], with the PB formalism [5, 10, 11, 76, 77, 100, 104, 105], and can be considered an approximation to the latter. However, we acknowledge the existence of an alternative view, in which the PB and GB are two different approximations, with neither one being an approximation to the other [27]. While a detailed discussion of the issue is out of scope here, we argue that testing of a GB model against the PB is highly beneficial for assessment of whether the model is a good candidate for further testing and refinement. These tests are particularly useful when the two models employ the same type of dielectric boundary, which is the case for GBNSR6. Moreover, the "R6" flavor of the GB was originally derived [59] as a direct approximation to the Poisson equation for a special case (sphere), which further supports the numerical PB as the "first line" reference model.

Numerical PB reference. The MIBPB package [20] is used as the reference Poisson-Boltzmann method for computing the polar component of the binding free energy. MIBPB belongs to the class of highly accurate PB solvers with the second order convergence [195]. Polar binding free energies estimated via MIBPB are not sensitive to the grid size up to reasonably large grid size [122]. To ensure close agreement with the GBNSR6 parameter setting, the grid resolution in MIBPB is set to $h = 0.5 \text{ Å}$, ionic strength to 0.145 M, and the re-
remaining parameters are set to their default values. For the calculations presented here we used the web version of the package available at http://weilab.math.msu.edu/MIBPB/.

For a small subset of the test structures the solver returned 0.0 \( \text{kcal/mol} \) solvation free energy. In these cases, a very slight change of the grid resolution from \( h = 0.5 \text{ Å} \) resulted in a meaningful output, which was used.

**The electrostatic component of binding free energy.** Our main quantity of interest, and the target for accuracy estimate against the reference, is the electrostatic component of the binding free energy. The electrostatic component of the binding free energy is computed by the following thermodynamic cycle (depicted in Fig. 2.1):

1. Transmission of the individual protein and ligand from the solvent into vacuum 
\( (−ΔG_{\text{pol}}^{\text{protein}}, −ΔG_{\text{pol}}^{\text{ligand}}) \).

2. Formation of the protein-ligand complex in vacuum (\( ΔE_{\text{Coulombic}} = E_{\text{Coulombic}}^{\text{complex}} − E_{\text{Coulombic}}^{\text{protein}} − E_{\text{Coulombic}}^{\text{ligand}} \)).

3. Insertion of the complex into the solvent (\( ΔG_{\text{pol}}^{\text{complex}} \)).

![Diagram of the thermodynamic cycle](image)

Figure 2.1: The thermodynamic cycle used for computing the binding free energy. Water environment is shown in blue, and vacuum is in white.

Altogether, the electrostatic component of the binding free energy is computed as
\[ \Delta \Delta G_{\text{pol}} = \Delta G_{\text{complex pol}} - \Delta G_{\text{protein pol}} - \Delta G_{\text{ligand pol}} + \Delta E_{\text{Coulombic}}. \] (2.3)

While a perfectly accurate \( \Delta G_{\text{pol}} \) for each complex component would obviously result in a perfectly accurate \( \Delta \Delta G_{\text{pol}} \), the relationship between the accuracy of these quantities in general is not as straightforward \([65]\). This is yet another reason to consider \( \Delta \Delta G_{\text{pol}} \) directly as our accuracy metric.

**Grid-based molecular surface.** In this section, we compute the square surface approximation of an ideal SES (Fig. 2.2), based on the "field-view" method \([15]\). "Field view" computes the contact surface and re-entrant spherical triangles of the SES exactly, and uses an approximate spherical representation of the saddle surfaces. The saddle surfaces are sampled by a number of probe spheres, where the approximation depends on the arc length between two neighboring probe sites. Here we use arc resolution = 0.1 Å. We find that changes in arc resolution, in range of 0.1-1.0 Å, have minor effect on GBNSR6 efficiency and accuracy in terms of computing \( \Delta \Delta G_{\text{pol}} \).

The resulting SES surface is approximated by a uniform Cartesian grid discretized to a desired resolution \( h \); the grid constructed here is similar to the molecular surfaces commonly used to numerically solve Poisson’s equation based on finite-difference approaches \([63, 122, 182]\). The scheme of the discretization is illustrated in Fig. 2.2. To acquire the square surface elements, the grid centers of the cells intersected by the spheres are connected.

The algorithm \([15]\) takes the atomic positions and radii as input, and produces the square element approximations to the corresponding SES. This method is inspired by the conservation of the electric flux which indicates that every square surface element traverses the
2.2. METHODOLOGY

Figure 2.2: Finite-difference discretization of an abstract molecule. Blue spherical surface shows the ideal SES that represents the dielectric boundary, and green lines are the approximating square surface elements. The background black mesh depicts the uniform Cartesian grid.

same solid angle as the corresponding spherical surface element, and consequently the same flux passes through these two types of surface elements. This key relationship is illustrated in Fig. 2.3.

Figure 2.3: Spherical surface element $\delta S_k$ (blue) and the corresponding square surface element (green). For a point charge at $O$, the electric field flux through both elements is the same.

In general, "field-view" method can be used to compute the surface integral of any vector field $\mathbf{A}$ on the surface by

$$\int_{\partial V} \mathbf{A} \cdot d\mathbf{S} \cong h^2 \left( \sum_{M_k} \frac{A_k |x_k| R_k^2 (1 + \beta_{kj}) \cos \gamma_k}{|r_k - r_j|^3} + \text{similar terms in } y \text{ and } z \text{ directions} \right),$$

(2.4)
where $A_k$ is the field magnitude on the spherical surface element introduced in the field-view method ($\delta S_k$) in the $x$ direction, $dS$ is the infinitesimal surface element, $h$ is the grid spacing, $M_k$ is the number of square surface elements of the area $h^2$ in the $x$ direction, $x_k$ is the projection of the vector connecting the center of the $k^{th}$ square surface element to the center of the probe/atomic sphere on the $x$ coordinate, $R_k$ is the radii of the spheres that intersect the $k^{th}$ grid edge in the $x$ direction, $\gamma_k$ is the angle between the vector field $A$ and the normal direction of $\delta S_k (N)$, $r_k$ is the position of the center of the square surface element $k$ in the $x$ direction, $r_j$ is the position of the center of the sphere $j$. An illustration of the geometric meaning of the above parameters for the specific case of $A$ relevant to us is shown in Fig. 2.4.

In Eq. 2.4 $\beta_{kj}$ is

$$
\beta_{kj} = \left( \frac{3}{8} - \frac{5x_k^2}{8|r_k - r_j|^2} \right) \frac{h^2}{|r_k - r_j|^2},
$$

which is the result of applying the Taylor series expansion by the third order terms to the electric flux through the $k^{th}$ square surface element in the $x$ direction at the center of the square surface element. Note that $|r_k - r_j|$ is the distance between the center of the square surface element and the center of the sphere/probe. More detailed information specific to the "field-view" method can be found in [15].

In the effective Born radii calculation (Eq. 2.2) the vector field $A$ becomes $A = \frac{r-r_i}{|r-r_i|^2}$, and so the integration over the dielectric boundary surface is performed by using the "field-view" method, Eq. 2.4:

$$R_i^{-3} = -\frac{1}{4\pi} \int_{\partial V} \frac{r - r_i}{|r - r_i|^6} \cdot dS \approx -\frac{1}{4\pi} h^2 \left( \sum_{M_k} \frac{P_{kij}}{|r_k - r_i|^3} + \text{similar terms in } y \text{ and } z \text{ directions} \right),
$$

(2.6)
where \( r_i \) is the position of the center of atom \( i \), and \( P_{kij} \) is described as:

\[
P_{kij} = \frac{|x_k| R_k^2 (1 + \beta_{kj}) \cos \gamma_{ki}}{|r_k - r_j|^3}
\]  

(2.7)

where \( \gamma_{ki} \) is the angle between the vector field \( r_k - r_i \) and the normal direction of \( \delta S_k \). An illustration of the parameters from the above equations is provided in Fig. 2.4.

Figure 2.4: Effective Born radii calculation for atom \( i \) via the "field-view" method, Eq. 2.6 and Eq. 2.7. The black circle shows the atom for which the effective Born radius is calculated, and the yellow circle shows an atom or the probe corresponding to the surface element.

**MSMS-based molecular surface.** The MSMS package [150] is used to build a triangulation of the SES molecular surface. The resolution of the molecular surface triangulation produced by MSMS is controlled by a density parameter. We use a high value (6 vertex/Å²) which is approximately equivalent to \( h = 0.5 \) Å grid spacing in terms of the number of vertices for representing the surface. Eq. 2.2 is then numerically approxi-
mated using the triangles that form the molecular surface [2, 3]. The Bondi set of intrinsic radii and the solvent probe radius equal to 1.4 Å are used to determine the surface of the molecule. MSMS-based GBNSR6 is available via http://people.cs.vt.edu/~onufriev/software.php.

Approximating $h = 0$ grid. The grid size dependence of $\Delta \Delta G_{\text{pol}}$ computed by GBNSR6, is investigated relative to $\Delta \Delta G_{\text{pol}}(h = 0 \text{ Å})$, which is approximated as follows. We assume a quadratic dependence of the grid-size error in $\Delta \Delta G_{\text{pol}}$: $\Delta \Delta G_{\text{pol}}(h) - \Delta \Delta G_{\text{pol}}(0) = ah^2$. For each complex, the unknown $\Delta \Delta G_{\text{pol}}(0)$ and $a$ are obtained by solving the system of two equations for $h = 0.3 \text{ Å}$ and $h = 0.2 \text{ Å}$.

Usage. The input structure to be processed by GBNSR6 must be in standard Amber topology and coordinate files. User can control the input parameters such as grid spacing, dielectric constants and solvent probe radius. The output of the program can be either displayed on the terminal or stored in a file. Grid-based GBNSR6 is available through Amber 2016 from http://ambermd.org/. User manuals, tutorials, and updates can also be found on the website.

2.3 Analysis

2.3.1 Practical Grid Spacing

An important consideration in using grid-based molecular surface models is to set the grid spacing so that a reasonable balance between the accuracy and efficiency is achieved. Besides, sensitivity of the model to grid parameters must be investigated to ensure robust-
ness and consistency. Specifically, while numerical integration performed via Eq. 2.6 and Eq. 2.7 becomes exact in the limit of grid size \( h \to 0 \), we need to know how this formula behaves for \( r^{-6} \) kernel for practically reasonable finite values of \( h \). Note that the integral in Eq. 2.6 yields the effective Born radii, and accuracy of the latter is key to the accuracy of the GB model [133]. Since the main quantity of practical interest in this study is \( \Delta \Delta G_{pol} \), we investigate its dependence on the grid resolution directly.

**Sensitivity to grid resolution.** To investigate the accuracy of the grid-based GBNSR6, the \( \Delta \Delta G_{pol} \)s for the set of small complexes are computed. Having the zero spacing grid as the reference, the relative errors for other grid spacings are calculated (see Methods). Based on the results, we propose to set the largest acceptable error in \( \Delta \Delta G_{pol} \) to \( k_B T = 0.59 \) kcal/mol; Fig. 2.5 illustrates that \( h = 0.5 \) Å is the maximum grid spacing which results in the acceptable error of less than \( k_B T \). In other words, this result implies that the relatively coarse grid resolution of \( h = 0.5 \) Å still leads to acceptable accuracy of GBNSR6 in estimates of protein-ligand binding energies.

**Accuracy relative to numerical PB.** As is standard in the field, the accuracy of GBNSR6 is evaluated relative to a numerical PB reference. For this purpose, we chose MIBPB [20], see "Methods" for the rationale. The computed values of \( \Delta \Delta G_{pol} \) for the protein-ligand complexes obtained from the grid-based GBNSR6 are compared with the corresponding MIBPB values in Fig. 2.6. The high correlation \((r^2 = 0.97)\), a relatively small deviation (RMSE=1.43 kcal/mol) and the virtual absence of a systematic bias indicate that GBNSR6 is a very reasonable approximation to the PB in computing \( \Delta \Delta G_{pol} \). At the default \((h = 0.5 \) Å \) grid resolution, the RMS errors relative to the PB reference are 4.74 kcal/mol for complexes, 3.76 kcal/mol for proteins and 2.16 kcal/mol for ligands. Individual solvation
free energies at various grid resolutions are available in the Supporting Information of [46].

Figure 2.6: Correlation between $\Delta \Delta G_{pol}$ computed by GBNSR6 and the numerical PB reference MIBPB ($h=0.5 \, \text{Å}$). Red line $x=y$ indicates a perfect match.

Computational cost. We measured the elapsed time for computing the $\Delta G_{pol}$ for a set of structures listed in Tab.2.1. This test was done on a commodity PC with the following
configuration: intel (R) Core (TM) i7-2600 CPU 3.40 GHz and 16 GB RAM. The elapsed time for different grid spacings, from the finest grid spacing $h = 0.2$ Å to the coarsest $h = 1.0$ Å is shown in Fig. 2.7. Note that for the two finest grid resolutions, there was not enough memory for going beyond the 7819 atoms. One can observe that by setting the grid spacing to $h = 0.5$ Å, the computational cost is relatively low while the error in $\Delta \Delta G_{pol}$ is still less than $k_B T$. A discussion of the computational cost of GBNSR6 relative to some other methods can be found in [81]; a detailed comparison of several other GB and PB methods is available in [40].

![Figure 2.7: GBNSR6 cpu time of computing $\Delta G_{pol}$ for different grid resolutions.](image)

**Sensitivity to grid orientation.** One potential source of error in grid-based surface framework is associated with the position/rotation of the molecule to be mapped onto the grid. Different spatial orientations of the molecule may result in different projections onto the grid and consequently different representations of the molecular surface. Given the uniform orthogonal Cartesian grid, any rotation of $k\pi/2, (k \in \mathbb{Z})$ of the molecule relative to
Table 2.2: Deviation of $\Delta \Delta G_{\text{pol}}$ values (kcal/mol) of the set of small protein-ligand complexes from those rotated around different axes each through $\pi/4$. The first column shows the results when complexes are rotated $\pi/4$ around $x$ axis, followed by $\pi/4$ rotations around $y$ and $z$ axes, respectively. The last column shows the results when $\pi/4$ rotations around $x$, $y$ and $z$ axes are applied together.

<table>
<thead>
<tr>
<th></th>
<th>$x$</th>
<th>$y$</th>
<th>$z$</th>
<th>xyz</th>
</tr>
</thead>
<tbody>
<tr>
<td>average</td>
<td>-0.08</td>
<td>-0.21</td>
<td>-0.22</td>
<td>-0.14</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.28</td>
<td>0.30</td>
<td>0.40</td>
<td>0.34</td>
</tr>
</tbody>
</table>

its original position in space results in the same final projection (see Fig. 2.2) and, hence, the same computed values of $\Delta \Delta G_{\text{pol}}$ within the numerical precision. However, any rotation angle in-between causes some finite-difference discretization deviation, leading to a deviation of $\Delta \Delta G_{\text{pol}}$ from the original $\Delta \Delta G_{\text{pol}}$, which is an artifact – true solvation energy is rotation independent. One may expect the artifact to reach its maximum value for the "half way" angles $\sim (2k + 1)\pi/4$. Here we examine rotation-dependency of our model by comparing the $\Delta \Delta G_{\text{pol}}$ of the small protein-ligand complexes dataset, when the complexes are rotated through $\pi/4$ around different axes (Tab. 2.2). The rotation was performed by cpptraj in Amber [18].

### 2.3.2 Grid-Based vs. MSMS-Based Molecular Surface Implementations of GBNSR6

In this section, we investigate two molecular surfaces in terms of efficiency and accuracy of computing $\Delta G_{\text{pol}}$ and $\Delta \Delta G_{\text{pol}}$ respectively, using the original implementation of GBNSR6 [2] based on MSMS [150] triangulated molecular surface, and the present grid-based molecular surface [15]. At comparable resolution, the elapsed times for computing the $\Delta G_{\text{pol}}$ on the dataset of 17 structures (Tab. 2.1) by these two methods are shown in Fig. 2.8. The computation based on the proposed grid-based surface with $h = 0.5$ Å is
2.3. Analysis

appreciably faster than that based on MSMS surface with roughly the same number of vertices \((d = 6 \text{ vertex/Å}^2)\). For instance, for the structure with 15740 number of atoms, it takes about 90 seconds for the grid-based GBNSR6 to compute \(\Delta G_{pol}\), while using the MSMS based surface it takes about 310 seconds.

The deviation between grid and MSMS based methods in computing \(\Delta \Delta G_{pol}\) for the 15 small complexes, is small: on average of 0.16 kcal/mol, with RMSE = 0.71 kcal/mol. However, for some cases the difference between the results is considerable, see below. It should also be noted that occasionally, MSMS fails to produce a reasonable molecular surface for some structures, see examples in the Supporting Information. When the problem calls for an automatic processing of a large number of structures, speed and robustness of the method become important factors.

Figure 2.8: Comparison of the elapsed time for computing \(\Delta G_{pol}\) using the original MSMS-based GBNSR6 and the current grid-based implementation.

Our implementation of "R6" based on MSMS did not incorporate the option to handle the inner cavities – by default these are not computed by MSMS. Existence of interior cavities can potentially affect the calculations of effective Born radii in GBNSR6, which eventually may influence the solvation free energy computations. For example, consider the 1kxq
Figure 2.9: Comparison of the molecular surfaces computed by MSMS and grid-based methods for 1kxq complex. (a) the 3D structure of 1kxq complex, including protein (green) and ligand (gray). The internal cavities are displayed in orange. (b) dotted circles in the figure show internal cavities that are captured using the grid-based surface but are ignored by MSMS with the default setting. The black-square dots inside the dotted circles show atoms in the vicinity of the internal cavities. Shown is a 6 Å wide slice of the molecular surface. For visualization purposes grid spacing of $h = 0.3$ Å was used.

molecule shown in Fig. 2.9: while MSMS with the adequate setting does not compute the details of the internal molecular surface, the grid-based molecular surface properly computes the internal cavities. As a result, the effective Born radii computed with these two methods are clearly different, as shown in Fig. 2.10. Compared to MSMS, the grid-based molecular surface gives smaller effective Born radii for buried atoms when these are in the vicinity of the interior cavity surfaces. Since the cavities are assumed by the model to be filled with high dielectric of the exterior, this difference leads to significantly different values of $\Delta \Delta G_{pol}$ from these two surfaces. For the 1kxq structure (with 9399 atoms) the $\Delta \Delta G_{pol}$ value from GBNSR6 using grid-based surface is 167.74 kcal/mol, compared to 293.82 kcal/mol when the MSMS surface is used. Whether internal cavities should be treated as filled with high dielectric is a question outside the scope of this work [136, 139].
2.3. Analysis

2.3.3 Considerations for Further Speedup

While the proposed grid-based model is shown to be relatively efficient, bringing the efficiency to at least the level of fast analytical GB models is the ultimate goal. In the proposed grid-based GB model, the total time \( t_{\text{tot}} \) of estimating \( \Delta G_{\text{pol}} \) by the GB is \( t_{\text{tot}} = t_{\text{grid}} + t_{\text{rad}} + t_{\text{pair}} \), where the consecutive three terms are: (1) the cost of setting up the grid \( t_{\text{grid}} \), (2) estimating the effective Born radii \( t_{\text{rad}} \), and (3) computing the sum of all pairwise charge-charge interactions \( t_{\text{pair}} \) in Eq. 2.1. The last cost component, \( t_{\text{pair}} \sim N^2 \) \( (N = \text{number of atoms}) \), is the same for GBNSR6 and fast analytical pairwise GB models. To have a better insight into possible optimization strategies, details of the time analysis for grid resolution of \( h = 0.5 \) Å are shown in Fig. 2.11, where one can observe that \( t_{\text{pair}} \) is always much smaller than \( t_{\text{grid}} \) and \( t_{\text{rad}} \). For small structures \( (#\text{atoms} < 5000) \), setting up the grid is the bottleneck of the whole computation. However, the larger the proteins the more dominating the effective Born radii calculation time. Thus, to speed up the entire calculation, the first goal should be to decrease the \( t_{\text{grid}} \) for small structures and to
decrease $t_{rad}$ for large $\xi$

Figure 2.11: Time analysis of the grid-based GBNSR6 execution ($h = 0.5 \, \text{Å}$). Here $t_{grid}$ is the time for setting up the grid, $t_{rad}$ is the time for estimating the effective Born radii, and $t_{pair}$ is the time for computing the sum of all pairwise charge-charge interactions in Eq. 2.1.
Chapter 3

Multidimensional Global Optimization

This chapter appeared as the reference [47].

In this chapter, we provide a detailed description of the employed optimization method and the corresponding mathematical modeling of the dielectric boundary optimization in our choice of implicit solvent model.

3.1 Introduction

Many cellular processes such as signal transduction, gene expression, and protein synthesis are controlled by the binding of biomolecules. In structure-based drug discovery, in silico, accuracy and computational efficiency of the binding free energy prediction of small molecules to biomolecular targets are of paramount importance for high throughput screening of potential drug candidates [58, 86, 156, 192]. However, fast and accurate computational prediction of binding free energies continues to be challenging [29, 109, 113, 144, 146, 180, 185, 189, 198], and its outcomes depend strongly on the molecular modeling technique, particularly, on how well the solvent effects are approximated [114, 144]. There are two major categories of solvent models used in this field [138]: explicit and implicit. Within the explicit solvent framework, the mechanistic detail and the energetic effect of every single water molecule are explicitly considered, which in turn results in considerable computational cost. The implicit solvent model [9, 26, 55, 101, 162], which
treats the solvent as a continuum dielectric with polar as well as non-polar properties of water, may often offer a good balance between accuracy and speed. Within this framework, the generalized Born model \[31, 38, 49, 73, 123, 129, 134, 165, 173, 175\] is widely used due to its relative simplicity and efficiency \[135, 137\].

A key step in implicit solvent modeling is the determination of the solute/solvent dielectric boundary (DB), a region of space over which the dielectric constant \(\epsilon(r)\) shifts from the value characteristic of the solute interior (e.g., \(\epsilon = 1\) or 4) to that of the solvent, (e.g., 80 for water). Outcomes of implicit solvent calculations have proven to be extremely sensitive to the details of DB \[136, 171\]. The dielectric boundary is determined by the radii of the atom types comprising the protein as well as the size of the water probe \[97, 136\]. Treating the radii as free parameters, optimization of the dielectric boundary, considering only the minimum of four most abundant atom types in proteins (O, H, N, and C) along with the radius of the water probe, would require finding a minimum of the relevant objective function in a 5-dimensional parameter space. In the past, such optimizations for solvation free energies of small molecules were performed — the optimal DB minimized the deviation of the computed target from an accepted reference, either experimental or estimated via explicit solvent \[126, 127, 163, 168, 169, 188\]. One potential technical issue with previously derived optimal radii is that the true global optimum may not have been found — even for small molecules, the corresponding optimization problem is highly demanding, textbook numerical approaches are unlikely to find the global optimum in a rugged, multidimensional landscape. While this issue may not be critical in practice if a "good enough" local optimum is found, it still leaves the question open of how well one can do in principle. Finding a true global optimum can point to limitations of the underlying physical theory, and thus prompt further development. For practical calculations, a much more important limitation of optimal radii based on small molecule hydration
3.1. INTRODUCTION

energies is that it is highly likely that parameters defining the DB that are optimal for small molecule calculations are not optimal for estimates of protein-ligand binding free energies [64, 65, 109], which is of paramount interest.

To the best of the authors’ knowledge, global DB optimization targeting protein-ligand binding has not yet been performed, likely because of the sheer challenge of the corresponding optimization problem. The objective function landscape corresponding to the protein-ligand binding profile is very likely rugged, with numerous local minima. Finding the global minimum of such a non-convex function with many local minima is a very hard problem [36, 121]. Descent methods quickly terminate at a local minimum point. Evolutionary algorithms do not explore the entire feasible space, may not even converge to a local minimum point, and are generally inefficient in terms of the number of function evaluations. Statistical methods are likewise inefficient in higher dimensions $d$. Brute force search on a grid with $S$ points in each of $d$ dimensions has complexity $S^d$, which is intractable in practice even for modest $S = 10^2$ and $d = 5$ used in this feasibility study, for the computationally expensive function evaluations of interest here. Truly global methods such as Lipschitzian optimization are efficient, but require knowledge of the Lipschitz constant that is often unavailable. Recent advances in deterministic methods for global optimization [85] have led to an algorithm (DIRECT) that is remarkably frugal in terms of the number of function evaluations, practical for $d < 100$, does not require knowledge of a Lipschitz constant, and is theoretically guaranteed to find a global minimum point. The sophisticated search strategy of DIRECT has been generalized to a massively parallel version, implemented in the package VTDIRECT95 [72] used here.

As if finding a global optimum point was not hard enough, the problem of finding a practically useful optimum is even harder: the optimum must also be robust to virtually inevitable perturbations in either the replication of the optimal parameters or in the ob-
jective function. The latter source of uncertainty is relevant here, as the objective function defined on a necessarily limited set used in the training is guaranteed to be somewhat different from that corresponding to the test set chosen by somebody else in a specific application of the optimal parameters. One approach is to design a robustness metric that can be employed as a post-processing step, decoupled from the objective function, and in principle applicable to the outcome of any optimization [92].

This work has several novel aspects: first, the atomic radii are optimized specifically for protein-ligand binding free energy calculations. Second, a Statistical Physics inspired method is developed to select the best robust solution. The basic idea is that not only the value of the minimum of the objective function, but also the width of the "well" around the point should be taken into account. In order to have a better insight into the energy landscape, it is essential to explore the objective function around candidate solutions. Here we propose a connectivity graph-based approach to the problem. Moreover, to the best of our knowledge, the global optimization technique VTDIRECT95 is new to the field of structural biology.

### 3.2 Materials and Methods

#### 3.2.1 The Electrostatic Component of Binding Free Energy

The polar component of binding free energy, $\Delta \Delta G_{pol}$, is calculated via the thermodynamic cycle illustrated in Fig. 2.1 of Chapter 2. In general, the estimation of protein-ligand binding free energy is extremely computationally demanding. In order to make possible tens of thousands of such computations required for the DB optimization, single-point energy estimates are used here. The strategy of relying on single-point calculations in the opti-
mization is consistent with the use of single snapshot, and fixed structures to obtain the explicit solvent reference $\Delta \Delta G_{pol}$ values [81] employed here. The use of single snapshots for the optimization is a limitation, but a necessary one: attempting to estimate $\Delta \Delta G_{pol}$ for each trial point in the 5-dimensional atomic radii space based on thousands of snapshots, as is common in standard MMGBSA protocols [74], would have been prohibitively expensive in the context of the type multidimensional optimization we have pursued.

We choose $\Delta \Delta G_{pol}$, as opposed to the total $\Delta \Delta G$, as the main reference for several reasons. First, the main objective is to find parameters for the optimal DB, which explains the focus on electrostatics. Second, many practical continuum solvent models are based on the approximation that the polar and non-polar components of the total free energy are decoupled from each other; while this approximation has its limitations [22, 23, 35, 194], it is widely used [138]. Here, we decouple the polar and non-polar contributions by using as the reference $\Delta \Delta G_{pol}$ values computed in explicit solvent (TIP3P), and not considering the non-polar contribution in finding the optimal parameters of the dielectric boundary. Another reason why we do not consider the total binding free energy for optimizing the DB within this proof-of-concept work is because the total includes the entropy component– practical computational estimates of the latter involve potentially large uncertainties. Fundamentally, the DB is related to the shape of the molecule, while the entropy characterizes fluctuations about this shape, which is another argument for why it makes sense to consider optimizing parameters of the two separately, at least as the first approximation.
3.2.2 Implicit Solvent Model

The employed generalized Born (GB) model, i.e., gbnsr6, is described in Chapter 2. The dielectric (solute/solvent) boundary enters into the model via the effective Born radii calculated by the "$R^6$" equation [2, 3, 60, 117]:

$$R_i^{-3} = \left( -\frac{1}{4\pi} \oint_{\partial V} \frac{r - r_i}{|r - r_i|^6} \cdot dS \right),$$

where $\partial V$ represents the chosen representation of the dielectric boundary of the molecule, $dS$ is the infinitesimal surface element vector, $r_i$ is the position of atom $i$, and $r$ represents the position of the infinitesimal surface element. Uniform offset to the inverse effective radii is set to the default (which we also found optimal in the context of this work) value that is 0.028 Å$^{-1}$. Note that the dielectric boundary (DB) is not an experimentally measurable entity, a number of different approaches exist [101, 136] for representing it within the implicit solvent model. Solvent excluded surface (SES), also known as molecular surface (MS), is a widely used option to represent the DB in continuum electrostatic calculations [10, 20, 43, 55, 125, 145, 181], and we employ it here. While it was often argued [128, 167] that the DB based on SES is physically more realistic than computationally more facile alternatives such as VDW-based surface, opposite arguments and case studies exist [139]. What is certain is that outcomes of continuum solvent calculations are very sensitive to details of the DB [136, 171], including how internal cavities are treated. While the definition and representation of internal cavities within SES is relatively simple and robust, more sophisticated approaches exist, for example those based on multiple interacting surfaces [25] or smooth Gaussian dielectric boundary [68].

Within the SES-based representation of the DB, we use a grid based molecular surface implementation of "$R^6$", called GBNSR6 [46], for calculating the integral in Eq. 3.1. The grid
3.2. MATERIALS AND METHODS

resolution is set to 0.5 by default. A detailed analysis of GBNSR6 and its input parameters can be found in ref. [46]. Briefly, GBNSR6 approximates the ideal molecular surface with orthogonal grid patches. This approximation is based on the "field-view" method [15] inspired by the conservation of the flux through different surfaces. GBNSR6 has recently been shown to be the most accurate among several other GB flavors in predicting the electrostatic binding free energies, where the results from the Poisson-Boltzmann (PB) model were chosen as the reference [83]. Notice that, while the PB [10, 11, 56, 100, 104, 125, 145, 195] is generally more accurate than the GB, using the PB model directly in a global multidimensional optimization pipeline for calculating $\Delta \Delta G_{pol}$ is extremely computationally demanding. Specifically, the use of a high accuracy PB solver [195] in our optimization pipeline would have been prohibitively expensive; GBNSR6 approximates the PB reasonably well, at a small fraction of the cost.

3.2.3 Objective Function

Considering the five radii ($\rho_w, \rho_C, \rho_H, \rho_N, \rho_O$) as free parameters, the dielectric boundary optimization turns into a multidimensional constrained optimization with respect to minimization of error in calculating $\Delta \Delta G_{pol}$. The root-mean-square error (RMSE) objective function to be minimized is

$$E_C(p) = \sqrt{\frac{1}{N} \sum_{c_i \in C} \left( \Delta \Delta G_{pol}^{GBNSR6}(c_i, p) - \Delta \Delta G_{pol}^{\text{TIP3P}}(c_i) \right)^2},$$  \hspace{1cm} (3.2)$$

where $\Delta \Delta G_{pol}^{GBNSR6}(c_i, p)$ is the electrostatic binding free energy calculated by GBNSR6 for complex $(c_i)$ given point $p$ in the 5-dimensional parameter space of $(\rho_w, \rho_C, \rho_H, \rho_N, \rho_O)$. $\Delta \Delta G_{pol}^{\text{TIP3P}}(c_i)$ is the reference electrostatic binding free energy calculated with TIP3P for complex $(c_i)$, and $C$ is a given data set of $N$ complexes. (In our case, $C$ is a data set
of $N = 15$ small protein-ligand complexes.) The optimization is performed under the constraints on the probe and atomic radii listed in Eq. 3.3. ParmEd editor in AmberTools is used for replacing the five radii, that is an old point $p$ with a new one, in complex $c_i$, at each iteration of the optimization. For previously developed radii not optimized in this work, the equation above is also used to compute the RMSE for comparison, without any optimization. The above objective function is deliberately cast in a form that retains the units (dimensionality) of the physical target quantity, energy here.

### 3.2.4 Sampling Around the Minimum Points

To have better insight into the behavior of the objective function, the robustness analysis was performed on one or five thousand sample points in the close vicinity of the best minimum points. Latin hypercube sampling (LHS) [107], a common algorithm for high dimensions [90], was selected from the QNSTOP package [7]. Briefly, LHS partitions the multidimensional space into grid cells and generates random sample points so that there exists one and only one sample point per row and column. A 2-dimensional example to demonstrate the idea is shown in Fig. 3.1.

LHS is easily generalized to high dimensions where many well-known methods, such as naive Monte Carlo, fail to explore the space comprehensively. To find the size of the sampling box, the global minimum point was examined as follows: we fixed four of the five variables around this point alternatively, and changed the fifth one so that the deviation from the minimum reached $1.2 \text{ kcal/mol} (= 2kT)$. This strategy guarantees quite a wide region to gain meaningful samples, while avoiding potential overlaps between global and local minima. Expectedly, this strategy produces an asymmetric rectangular sampling box, as the electrostatic characteristics of the atomic types are different:
Figure 3.1: Latin hypercube sampling (LHS). This example shows how LHS generates random sample points in a 2-dimensional space so that there exists one and only one sample point in each row and column.

lower bounds= \((\rho_W - 0.6, \rho_C - 0.5, \rho_H - 0.1, \rho_N - 1.0, \rho_O - 0.05)\),

upper bounds= \((\rho_W + 0.2, \rho_C + 0.5, \rho_H + 0.1, \rho_N + 0.3, \rho_O + 0.05)\).

3.2.5 Data Sets for Training and Test

The entire data set consists of 15 protein-ligand complexes for which \(\Delta\Delta G_{pol}\) estimates in explicit solvent (TIP3P) are available, and described in detail in [81]. This data set was used previously in similar contexts [81, 83, 88]. Small in size (1635-1995 atoms) and diverse with respect to values of \(\Delta\Delta G_{pol}\) (0.71-25.01 kcal/mol), these complexes are good candidates to resemble those in drug discovery. The complexes, ligands, and proteins are neutral, individually. This choice is deliberate, as it avoids various uncertainties and complications due to the use of Ewald summation and periodic boundary conditions in explicit solvent simulations used in a previous study [81] to estimate the electrostatic binding free energies employed here as the reference. Also, the structures were restrained [81] to mitigate uncertainties due to conformational variability. Unless otherwise specified,
the data set is partitioned into two subsets of eight (1pbk, 1fkf, 1bhf, 1fkh, 2hah, 2fke, 1zp8, 1f40) and seven (1b11, 1fb7, 1fkb, 1fkg, 1fkj, 1fkl, 3kfp) complexes in order to train and test the proposed computational protocol, respectively. This partitioning guarantees similar distribution of $\Delta\Delta G_{\text{pol}}$ values between the two subsets.

### 3.2.6 VTDIRECT95: Global Multidimensional Optimization Method

The deterministic DIRECT (Dividing Rectangles) global minimization algorithm \cite{ DIRECT85} is a powerful optimization method for a moderate number of dimensions. DIRECT guarantees \cite{ DIRECT85} global convergence if the objective function is Lipschitz continuous, without requiring a gradient or knowledge of the Lipschitz constant. With wide application in many practical nonlinear optimization problems, DIRECT has proven to be a straightforward and efficient optimization method. In a nutshell, DIRECT iteratively divides the search space into boxes, identifies the potentially optimal boxes (those most likely to contain a global minimum point), and subdivides them into smaller boxes. An illustration of this algorithm for a 2-dimensional global search is given in Fig. 3.2.

VTDIRECT95 \cite{ VTDIRECT95} is a Fortran 95 package containing a serial and a massively parallel implementation of DIRECT, scaling to several thousand processors, due to the usage of distributed control parallelism instead of a common master-slave paradigm, and dimension 100. Sophisticated dynamic data structures and memory management strategies make VTDIRECT95 efficient and robust \cite{ VTDIRECT95, VTDIRECT952, VTDIRECT953}. VTDIRECT95 is used for optimizing the atomic radii and the probe radius in a feasible range, to be determined in "Results and Discussion", so that the binding free energies calculated by GBNSR6 have the best agreement with those calculated by the reference explicit solvent model TIP3P \cite{TIP3P}. As with any mathematical software, VTDIRECT95 has a few input parameters whose understand-
3.2. MATERIALS AND METHODS

Figure 3.2: Function evaluations performed by DIRECT after 0, 1, 5 and 10 iterations. The objective function values are illustrated via the contours and the corresponding color bar on the rightmost panel. Comparing the first and second graphs on the top shows how DIRECT divides a 2-dimensional box after one iteration. On the bottom right figure, DIRECT finds the global minimum at (0.9,0.3) after 10 iterations. It also explores a large domain and evaluates the function near the local minimum at (0.4, 0.3).

Tuning and tuning will improve performance. However, extensive tuning of these is not necessary, and the time spent tuning usually outweighs the time from a single run with reasonable (derived from domain knowledge) and default values.

VTDIRECT95 was employed for the 5-dimensional global optimization with respect to the objective function shown in Eq. 3.2, its argument being the vector of parameters: \((\rho_w, \rho_C, \rho_H, \rho_N, \rho_O)\). We tune three parameters to improve efficiency of the global optimization with VTDIRECT95:

- \textit{eval\_limit} = 40000: This condition terminates the optimization after 40,000 number of objective function evaluation. Each round of minimization took 1.5 days using 64 CPUs (AMD Opteron (TM) Processor 6276) in parallel to run 40,000 objective
function evaluations. There was no decrease, within 5 decimal point accuracy, in objective function value beyond 38,000 iterations.

- $\text{eps}_{f\min} = 0.0001$: This parameter stops subdividing any box further unless the expected change in the objective function in that box is greater than 0.0001. This prevents wasted compute time exploring the box where the objective function is not expected to change much. On the other hand, this is a rough estimate over the expected changes in each box. To avoid losing the global minimum, and after several trials, the best setting for this parameter turned out to be 0.0001.

- $\text{min}_{\text{sep}} = 0.5$: In computing multiple ($k$) lowest minima corresponding to the global and local minimum points, without limiting the distance between them, VT-DIRECT simply returns the $k$ best values, all likely next to each other. We define two minimum point ($p_1$ and $p_2$) in the radii space to be meaningfully different if their corresponding atomic radii are 0.2 far apart, on average, per dimension (that is per atom type). This constraint leads to a minimum 0.5 distance between two such points in a 5-dimensional space, i.e., $d(p_1, p_2) = \sqrt{(p_1^1 - p_2^1)^2 + (p_1^2 - p_2^2)^2 + (p_1^3 - p_2^3)^2 + (p_1^4 - p_2^4)^2 + (p_1^5 - p_2^5)^2} = \sqrt{(5 \times (0.2)^2)} \approx 0.5$. $\text{min}_{\text{sep}}$ is the corresponding parameter in VT-DIRECT95 that controls the minimum distance allowed between any two optimal points. Note that this parameter is taken into account after the optimization, and it does not affect the global search itself, only which minima are reported.

In summary, we choose a combination of $\text{eval}_{\text{lim}}$ and $\text{eps}_{f\min}$ for an efficient exploration of the parameter space, and minimizing computational time wasted on those boxes that are not likely to contain the global minimum. After the search, by setting $\text{min}_{\text{sep}} = 0.5$ we select those best minima that are "meaningfully" far apart. The remaining param-
3.3. Bounds on Physically Meaningful Values of Atomic Radii

3.3.1 Radial Distribution Function

A set of 11 small molecules was selected from a larger set of 504 small drug-like molecules [116], see Tab. 3.1. The choice of these 11 structures was guided by a prior work [115], where 10 ns long simulation trajectories were generated for all 504 molecules using implicit [134] water Langevin dynamics at 298 K. To minimize possible uncertainties [120] due to inadequate conformational sampling of flexible molecules, these 11 structures were among the ones with the lowest time averaged RMSD with respect to the original conformation. For the "solute atom"-"water oxygen" radial distribution functions (RDF) estimates, we performed explicit water simulations on these 11 shortlisted structures using Amber12 [19] simulation package; molecule coordinate and topology files were obtained from elsewhere [116] and molecule parameters were assigned using the GAFF force field [183].
The molecules were solvated in a pre-equilibrated cubic box with the TIP3P model water with at least 12 Å distance from the solute to the nearest box edge. The solute-solvent system was prepared first by a shallow steepest descent minimization followed by a second order conjugate gradient minimization while restraining solute atoms in the Cartesian space using a harmonic potential of 200 \( kcal/mol^2 \). Subsequently, equilibration and production runs were performed using the Langevin dynamics with a collision frequency of 1 \( ps^{-1} \) and integration time step of 2 \( fs \) while the bonds were constrained by the SHAKE algorithm [112]. Positional restraints of 200 \( kcal/mol^2 \) were employed on solute atoms throughout, and electrostatic interactions were approximated via the Particle Mesh Ewald (PME) method, with 9 Å direct sum cutoff. Minimized solute-solvent system was equilibrated in two steps; first, the system was heated to 298 K for 1 ns using an NVT ensemble followed by a 298 K, 1 bar NPT ensemble simulation for another 1 ns. The RDFs were computed from the later 18 ns of a total of 20 ns long trajectory from 298 K, 1 bar NPT simulations using the radial function in cpptraj [147] feature of AmberTools between each solute atom and water oxygen. Positional restraints in the production runs were used to obtain a "clean" estimate of the bounds for the atom + water probe distances. Running the simulation without such restraints would likely lead to a larger amount of noise in the RDF, coming from conformational variability. This approach is consistent with our choice of a subset of the most rigid molecules from the small molecule data set listed in Tab. 3.1.

Table 3.1: The list of 11 molecules used in this work to compute the solute atom to solvent (TIP3P) oxygen radial distribution function.

<table>
<thead>
<tr>
<th>111_trichloroethane</th>
<th>1234_tetrachlorobenzene</th>
</tr>
</thead>
<tbody>
<tr>
<td>2_bromo_2_methylpropane</td>
<td>diethyl_sulfide</td>
</tr>
<tr>
<td>methyl_methanesulfonate</td>
<td>tetrafluoromethane</td>
</tr>
<tr>
<td>112_trichloro_122_trifluoroethane</td>
<td>1_methylcyclohexene</td>
</tr>
<tr>
<td>4_fluorophenol</td>
<td>iodobenzene</td>
</tr>
<tr>
<td>morpholine</td>
<td></td>
</tr>
</tbody>
</table>
3.3. BOUNDS ON PHYSICALLY MEANINGFUL VALUES OF ATOMIC RADII

3.3.2 Radii Constraints

To enforce physical realism and reduce over-fitting we use atom-oxygen radial distribution function (RDF) as the key constraint in constructing the dielectric boundary, see "Materials and Methods". Note that unlike the DB, which is a theoretical concept, RDF is an experimental observable. Specifically, the probe radius \( \rho_w \) and the intrinsic atomic radii \( \rho_i \) are optimized simultaneously, under the physically justified constraint that \( \rho_i + \rho_w \) is bounded within one standard deviation of the first peak of the RDF, see Fig. 3.3. The first-peak region is defined as the region bounded by the minima before and after the first peak in an RDF. Combining all the "first-peak" RDF data for a particular atom-type \( i \), the optimization range is then defined as the mean \( \pm \) standard deviation over that data. In the left panel of Fig. 3.3 we show an example of RDFs obtained from molecular dynamics simulation trajectories of different molecules; after combining the first-peak regions and computing the standard deviation, the optimization region is defined by \( (R_{\text{min}}, R_{\text{max}}) \).

The RDFs are computed using molecular dynamics simulations in TIP3P [87] explicit solvent. As the result, the following upper bounds and lower bounds are obtained:

\[
0.2 \leq \rho_w \leq 1.6 , \\
2.2 \leq \rho_w + \rho_C \leq 3.8 , \\
1.4 \leq \rho_w + \rho_H \leq 3.0 , \\
2.2 \leq \rho_w + \rho_N \leq 3.8 , \\
2.2 \leq \rho_w + \rho_O \leq 3.8 .
\] (3.3)

The bounds for the water radius \( \rho_w \) were obtained as follows: the upper bound for the water probe radius was chosen (with a buffer of 0.2 Å above) as the standard water probe radius of 1.4 Å, the lower bound was chosen as (with a 0.2 Å buffer lower than) the
Figure 3.3: Solvent excluded surface (SES) exemplified for a "molecule" of six atoms. SES is shown as the purple boundary, defined as the locus of the contact points (connected by circle arcs at contact discontinuities) of water probe (white circle) when it is rolled over the molecule (gray circles). An example of radial distribution function of atom-(water oxygen) obtained for atom type \(i\) from molecular dynamics simulations of various molecules containing that atom type is shown in \(g_{i-ow}(R)\) plot to the left of the schematic. Each color in \(g_{i-ow}(R)\) plot represents a separate instance of atom type \(i\); the bounds \((R_{\text{min}}, R_{\text{max}})\) are computed as one standard deviation about the mean (shown as the double-headed red arrow) of the RDF first peak, inferred from the combined data of the first peaks for all the instances of the atom type \(i\). These bounds are used to constrain \(\rho_i + \rho_w\) for simultaneous optimization of \(\rho_i\), atomic radii of atom \(i\), and \(\rho_w\), water probe radius.

standard water radius 1.4 Å minus the standard water oxygen-hydrogen bond length of approximately 1 Å. There are only a few complexes containing sulfur (S) atoms in the protein-ligand data set; to avoid any potential over-fitting, the S radius is set to 1.8 Å (Bondi) as the default. For a fair comparison, the same radius is considered for S in PARSE [163] and ZAP-9 [126].
Chapter 4

Robustness Analysis and Visualization

This chapter appeared as the references [47, 48].

In this chapter a novel robustness metric is introduced that helps distinguish between nearly degenerate local minima found by any optimization method, such as VTDIRECT95 described in Chapter 3 via a post-processing step. A graph-based "kT-connectivity" approach to explore and visualize the multidimensional energy landscape is proposed: local minima that can be reached from the global minimum without exceeding a given energy threshold (kT) are considered connected. The proposed robustness and visualization methods give better insight into the complex objective function landscape and are applicable in other optimization problems.

4.1 Introduction

Global multidimensional optimization in real-world applications can be very difficult as the objective function landscape is often rugged with numerous local minimum points. It is not uncommon to identify several local optimum points that are hard to distinguish in practice – differences between the corresponding objective function values may be too small to make the differences meaningful in the context of the underlying problem. What makes the problem even harder is that the evaluation of such an objective function at each search point can be computationally expensive. This computational expense makes
it difficult to explore the objective function landscape at a fine grain resolution, often needed to identify the global optimum, see [71] and references there in.

As if finding a global optimum point was not hard enough, the problem of finding a practically useful optimum is even harder: a good solution must also be robust to unavoidable perturbations and uncertainties, virtually inevitable in any complex system being optimized. As an illustration, choosing the global optimum point, corresponding to the bottom of a deep but narrow well, may lead to undesirably high deviation from the optimum when the same solution is replicated with even a small uncertainty, or applied to a different data set, not used in the original optimization, Fig. 4.1. A way to address this issue is by identifying robust optimal parameters, which when subject to small perturbations around the optimum result in acceptable variations in the objective function [108]. The robustness must be carefully balanced with the need to achieve the most optimal value of the objective function. Robust optimization is a field of optimization theory that aims to find optima under uncertainty. Assuming that modeling uncertainty with a deterministic function is possible, the initial objective function is modified into a more complex function which not only seeks to optimize the objective function but also minimizes the uncertainty. There are different approaches to implement this idea. For example, in [91] statistical decision theory is applied for quantifying uncertainties with applications in multidimensional integration. In [196, 197], Zhou et al. introduced new algorithms to find competitive diverse alternatives to global optima. All of these approaches imply that enough detailed information about the uncertainty is available, so that it could be incorporated into the objective function directly as a deterministic component. However, note that incorporating the uncertainty directly into the objective function may limit the choice of optimization methods available for the specific problem. For hard multidimensional optimization problems this choice is already limited to begin with.
4.1. INTRODUCTION

Figure 4.1: Depth vs. width trade-off in choosing the preferred optimum. The shallow and wide local minimum point on the right may be more robust to perturbation compared to the narrow but deeper global minimum solution on the left. When the difference $\Delta$ between the objective function at the minimum points is close to the problem-specific range in which two optima are practically indistinguishable, one may intuitively conclude that the wider local minimum solution should be preferred for a complex system with inevitable uncertainties in replication of its parameters. The goal of this work is to formulate and test a quantitative criterion for such preference.

A possible strategy to tackle the problem is to design a robustness metric that can be employed as a post processing step, after the optimization has been performed and the optima identified on the original objective function of the problem at hand. Such a metric decoupled from the objective function can also be not domain specific, and in principle could be easily applied to the outcome of any optimization resulting from the most appropriate method. Intuitively, the objective function near a robust optimum should not deviate sharply from the objective function value right at the optimum. In [92] five statistical robustness metrics are introduced based on this idea: the distribution of the objective function values around a robust optimum is more skewed towards the value of that robust solution, which leaves little room for uncertainty about the expected outputs. These metrics solely investigate the robustness of the main objective function by studying the statistical characteristics of the points around the optima.
In this study, we go beyond the examination of the objective function exclusively found for a training set, and investigate the robustness of optima when their corresponding optimal solutions are utilized in new test cases, which has many practical applications. We propose a novel statistical robustness metric that addresses the depth vs. width trade-off (see Fig. 4.1), in a manner motivated by the similar concept in statistical physics: enthalpy-entropy competition in free energy function of thermodynamics. Here, we engineer the energy function to obtain a simple and general post processing robustness metric. Our robustness metric is evaluated through a case study: optimization of key parameters that define the molecular shape, in our case protein-ligand (dielectric) boundary. Finding the most robust optimal solution in this problem not only improves the accuracy of these important practical calculations, but also guarantees approximately similar results when the optimized model is tested on new cases. This robustness is essential in drug discovery where binding free energy calculations play a key role. Since this work focuses on the robustness metric and not on the case study itself, we present only the minimal "bare bones" of the problem, just enough to illustrate our proposed metric.

4.2 Proposed Metric of Optimum Robustness

Even if globally optimal parameters have been found, there is no guarantee that their use in practice will always lead to the most optimal outcome due to multiple sources of error: for example, physical manufacturing of the system with the exact optimal parameters may not be possible in practice (case I) due to inevitable errors in the process. Besides, optimal parameters are obtained based on a limited training data set, so the objective function may be different for the actual problem (test set) where the optimal parameters are used in practice (case II). Although different strategies may be employed to mitigate
over-fitting, these do not completely remove the risk of low transferability between data sets. Therefore we argue that a solution that is slightly less optimal than the global optimum, but leads to less error when replicated, may be preferred over the true global optimum. In this section, we propose a general metric for studying the optimum robustness, potentially applicable to the incidents of the two sources of error. The motivation is illustrated for the manufacturing source of "noise", case I, which we believe is the most straightforward scenario. Later, a detailed application of the metric is developed for case II which is directly relevant to our problem of dielectric optimization.

**Motivation.** To illustrate, consider a 1-dimensional optimization scenario shown in Fig. 4.2. In the first example (left panel), the two minima correspond to the wells at $x = 0.5$ and $x = 1.5$, which are equally "wide", meaning that inevitable small deviations of the parameters from the optimal values (shown within the orange interval) lead to the similar deviations of the objective function from the minimum. In that respect, both minima are equally robust. As the objective function $E(x)$ at $x = 0.5$ is lower than that at $x = 1.5$ by a positive $\Delta$, the minimum at $x = 0.5$ is preferred. In contrast, for the function shown in the right panel, one can argue that the local minimum at $x = 1.5$ is a better choice under some circumstances, even though the value of the objective function at $x = 1.5$ is higher by $\Delta$ than the global minimum at $x = 0.5$. This is because the local minimum well at $x = 1.5$ is wide and flat, so that deviations of the parameters from this local minimum – due to, for example, manufacturing errors in replicating the precise optimal parameter values– do not lead to appreciable deviations in the objective function. However, small changes ("noise") in the parameters from the global minimum at $x = 0.5$ result in substantial deviations in $E(x)$. The above reasoning about depth vs. width is intuitive, but not easy to express in a mathematical form. The main difficulty is comparing the depth and the width on the same footing: in general, these are not even expressed in the same
Figure 4.2: Robustness analysis of two examples. Left panel shows two equally wide wells, which are similarly robust to small perturbations of the parameters. The right panel shows a totally different behavior of the objective function, where the wide local minimum is more robust to perturbations than the narrow global minimum.

physical units, e.g., energy vs. length in the case of the optimization discussed in this work. Insight into a possible solution to the problem comes from Statistical Physics [94]: free energy

$$F = -\xi \ln \sum_x e^{-(E(x) - E_g)/\xi}$$  \hspace{1cm} (4.1)

includes both the depth (energy) and the width (entropy) of a state, where $E_g$ is the global minimum of $E(x)$ and $\xi = kT$ is, in effect, the strength of the "thermal noise". The state $x$ with the lowest free energy $F$ corresponds to the most preferred thermodynamic state in the energy landscape $E(x)$ of the system when it is coupled to a thermal noise.

Unfortunately, Eq. 4.1 is derived for the specific case of systems in thermal equilibrium, and can not be assumed to be valid a priori for a general optimization problem. Moreover, it is not clear how to choose $\xi$ in Eq. 4.1 in general. For example, simply equating $E(x)$ in Eq. 4.1 with an objective function that corresponds to the cost of car production is difficult to justify. Note that, in Physics, $E(x)$ and $\xi$ have very specific properties that factor into the specific form of Eq. 4.1. Despite these conceptual difficulties, free-energy
4.2. **Proposed Metric of Optimum Robustness**

Like functions have been used in machine learning [154] and optimization [191] mainly as the objective function. However, it is worth mentioning that even if the entire energy landscape is explored with a perfect objective function, finding the most robust solution is nontrivial and necessitates further analysis. The discussed entropy idea cuts across multiple disciplines. For instance, von Neumann entropy was used as a measure of the complexity of protein binding pockets [45], networks [140] and graphs [34]. Here, our focus is on the robustness of optimal solutions with an application to a problem related to computational drug discovery.

In what follows a more general metric of robustness of optima is designed, free from the limitations mentioned above. Several observations about the structure of Eq. 4.1 give insights into the general structure of mathematical expressions that might be useful in comparing widths and depths of minima. The factor \(e^{-(E(x) - E_g)/\xi}\) in Eq. 4.1 penalizes heavily all the contributions to the sum in \(F\) that exceed the global minimum \(E_g\) of \(E(x)\) by more than \(\xi\); the value of \(\xi\) controls the penalty. In other words, only a few sample points contribute to the sum in \(F\) from a narrow well, while many more contribute from a wide well.

**Proposed robustness metric.** Inspired by the above example from Statistical Physics, we propose the following measure of optimum robustness: the expected value \(\langle E \rangle\) of the objective function taken over a representative neighborhood of the given optimum point. Specifically, \(\langle E \rangle = \int E(X)P(X)dX\) where \(P(X)\) is the probability distribution appropriate for the specific problem; \(P(X)\) characterizes the uncertainty of replicating the optimal parameters or the objective function optimum or both. Suppose \(\langle E_1 \rangle\) and \(\langle E_2 \rangle\) are the expected values of the objective function around minimum point \(X_1\) and \(X_2\), respectively; then, by the proposed criterion, if \(\langle E_1 \rangle < \langle E_2 \rangle\) then minimum point \(X_1\) is preferred over minimum point \(X_2\). Otherwise, \(X_2\) is preferred. Qualitatively speaking, \(\langle E \rangle\) is a robust-
ness metric compromising between "width" and "depth". Using Fig. 4.2 again as an illustration: on the left panel, the average of the objective function values in the left well is lower than that in the right one within their sampling boxes. In the right panel, while the narrow well contains the global minimum point, the average of its objective function values within the sampling box is higher than that of the wider well. The statistical meaning of the proposed robustness criterion can be made even more intuitive by noting that it is equivalent to the following: "choose \( X_1 \) if the probability that \( E_1 < E_2 \) is greater than 1/2." This is if the minimum is chosen by this criterion, chances are it delivers the lowest deviation from the reference, statistically speaking. The proof of the equivalence is particularly straightforward if one assumes normal distribution for \( P(X) \): 

\[
P(E_1 < E_2) = \frac{1}{2} \text{erfc}(\frac{\mu}{\sigma}),
\]

where \( \mu = \langle E_1 \rangle - \langle E_2 \rangle \), and \( \sigma^2 \) is the corresponding variance.

Below we develop an approach to estimate \( \langle E \rangle \) in practice. Motivated by the 1-dimensional statistical discussion earlier, consider an exponentially decaying weighted sample in a box \( B \) around a local minimum point \( X^* \) (in \( n \) dimensions) given by

\[
\langle E \rvert X^* \rangle = \int_B E(X)P(X)dX = \int_B A E(X)e^{-\left(1/2\right)(X-X^*)^t\Sigma^{-1}(X-X^*)}dX,
\]

(4.2)

where \( \Sigma \) is a \( n \times n \) diagonal matrix with \( \Sigma_{jj} \) being the empirical variance of \( X_j^* \), for \( j \in 1, ..., n \). The specific form of \( P(X) = Ae^{-\left(1/2\right)(X-X^*)^t\Sigma^{-1}(X-X^*)} \), where \( A \) is the normalization factor, is motivated by the common assumption of normal distributions for complex systems. However, note that, in general, no statistical distributional assumptions have to be made here, and that any reasonable decaying weight function \( P(X) \) based on the data could be used instead, as long as it satisfies the obvious normalization condition \( \int_B P(X)dX = 1 \). In what follows we verify the robustness of the proposed metric to the specific choice of \( P(X) \). Without loss of generality and for the sake of simplicity and il-
4.2. Proposed Metric of Optimum Robustness

Illustration, in what follows we consider $E(X)$ as a function of one variable $X$. In addition, for the sake of clarity and to simplify notation, we assume that the coordinate origin is shifted to $X^*$. 

**Uncertainty in reproducing the objective function.** Assume that the exact replication of optimal parameters is possible. (This is in fact the case in the dielectric optimization problem, where the exact optimal atomic radii can be generated computationally). As discussed earlier, it is unavoidable that, when a new data set (test set) is considered, the objective function will deviate from that used in the training to find the optimal parameters. To measure this deviation, consider the shape of the objective function in the close vicinity of the optimal parameters, see Fig. 4.3 left panel. Around its minimum point on the training set, the objective function is (nearly) a parabola such as $E(X) = aX^2 + c$. Deviation from this parabola results in another parabola such as $E'(X) = a'(X - b)^2 + c'$ on the test set. Note that shape conservation among all sets is a valid assumption because the training data set is supposed to be a legitimate representative of the whole set.

In general, each new test set will have its own values of $a$, $b$, and $c$. However, note that the value of the objective function at its minimum point on each parabola is not affected by changes in "$a". When several test data sets are studied, changes in "$c" lead to positive and negative deviations from the optimal objective function. It is not unreasonable to assume that this distribution is symmetric around its mean, and therefore the deviations in "$c" cancel out for a statistically significant number of test sets. Altogether, on average $E'(0) \propto b^2$. Using a 1-dimensional version of Eq. 4.2 for the illustration, $b \sim \mathcal{N}(0, \sigma^2)$. What the zero mean of the distribution implies is that the training set is well chosen, that is representative of the problem and unbiased. We assume this to be the case; the assumption can be verified explicitly in each specific case. Given this distribution, the
The average of the objective function values is

$$\langle E \rangle \cong A \int_{b \in \tilde{B}} E(b) e^{-\frac{b^2}{2\tilde{\sigma}^2}} \, db,$$  \hspace{1cm} (4.3)

where $\tilde{B}$ is the sampling box around $b = 0$, and $A$ normalizes the PDF, see "Materials and Methods". To estimate $\tilde{\sigma}$ in principle, one needs to compare $E^k(X)$ from a statistically significant number $k$ of independent test sets; each $E^k(X)$ is compared to $E(X)$ from the training set to identify the value of $b_k$, e.g., as in the example of Fig. 4.3 right panel. Then, $\tilde{\sigma}$ is computed as a standard deviation of $b_k$.

Figure 4.3: Deviation from the optimal solution $(X^*)$ given a new data set. Left: changes in objective function value at $X^* = 0$ ($\delta$) is proportional to $b^2$. Right: estimation of the standard deviation of $b$ when several test sets are given.
4.2. Proposed Metric of Optimum Robustness

4.2.1 Application to Optimization of Atomic Radii

Here we use VTDIRECT95 for global optimization of the probe and atomic radii. Results are shown in Tab. 4.1. The practically indistinguishable optima are re-ranked later using the proposed robustness metric.

Table 4.1: The lowest five optimum parameter vectors found by VTDIRECT95. Radii are in Å and objective function values of the training set, $E_{\text{train}}$, are in kcal/mol.

<table>
<thead>
<tr>
<th>OPT</th>
<th>$\rho_W$</th>
<th>$\rho_C$</th>
<th>$\rho_H$</th>
<th>$\rho_N$</th>
<th>$\rho_O$</th>
<th>$E_{\text{train}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.37</td>
<td>1.40</td>
<td>1.55</td>
<td>2.35</td>
<td>1.28</td>
<td>3.94</td>
</tr>
<tr>
<td>2</td>
<td>1.52</td>
<td>1.79</td>
<td>1.47</td>
<td>2.27</td>
<td>1.28</td>
<td>4.04</td>
</tr>
<tr>
<td>3</td>
<td>1.06</td>
<td>1.67</td>
<td>1.32</td>
<td>2.14</td>
<td>1.35</td>
<td>4.08</td>
</tr>
<tr>
<td>4</td>
<td>1.37</td>
<td>1.34</td>
<td>0.77</td>
<td>1.57</td>
<td>1.81</td>
<td>4.24</td>
</tr>
<tr>
<td>5</td>
<td>1.06</td>
<td>1.35</td>
<td>1.74</td>
<td>2.71</td>
<td>1.17</td>
<td>4.25</td>
</tr>
</tbody>
</table>

In what follows, a 5-dimensional form of Eq. 4.3 will be applied as the robustness metric for ranking the optimal solutions. The generalization of $\sigma^2$ in Eq. 4.3 is $\Sigma$ being the empirical variance of the global optimal solution $X^*$ from the test set. Here $\Sigma$ is a 5-dimensional diagonal matrix where $\text{diag}(\Sigma) = (\sigma^2_W, \sigma^2_C, \sigma^2_H, \sigma^2_N, \sigma^2_O)$, see Eq. 4.2. In other words, $\text{diag}(\Sigma)$ shows the variance of each radius resulting from the use of possible new test sets. The integration domain in Eq. 4.3 was estimated earlier in "Materials and Methods", and we use it here. The initial test set was introduced in "Materials and Methods"; here the test set is partitioned into seven test cases each made of one single protein-ligand complex. We are thus considering an instance of the general problem where one is interested in the performance of the optimal parameters on a single protein. As a result, we have a statistically meaningful distribution of $b$ values (see the right panel of Fig. 4.3).

To estimate $\Sigma$ we must make approximations. We assume that in going from the training to a test set, the whole objective function (energy) landscape shifts as a whole, with a similar pattern around each minimum, Fig. 4.3. Because of the $E(X)$ shift in going from
the training to the test sets, \( E^k(X^*) - E(X^*) = \delta_k > 0 \), where \( E^k(X) \) refers to the test case \( k, k \in \{1, ..., 7\} \). To find \( b_k \) we require that \( E(b_k) = \delta_k \), similar to how the sampling box bounds were identified, see "Materials and Methods". We repeat this process per dimension, assuming that the deviation in each radii contributes equally to the total deviation in energy. Given seven test cases, we calculate the variance of \( b \) which finally results in \( \text{diag}(\Sigma) = (0.0096, 0.0024, 0.0025, 0.0324, 0.0009) \). We apply the same \( \Sigma \) to evaluate robustness of all the optima in Tab. 4.2 – the use of the same \( \Sigma \) is justified by the assumption that the overall shape of the test set objective function is similar to that of the training set.

Objective function values, \( E_{\text{train}} \), and the corresponding ranking on 1000 and 5000 sample points, \( \langle E_{\text{train}}^{1000} \rangle \) and \( \langle E_{\text{train}}^{5000} \rangle \), for the lowest five optima, OPT1 to OPT5, are shown in Tab. 4.2. In order to study the effect of the underlying sharply decaying weighting function on the final ranking, we considered a modified \( P(X) \), \( P'(X) \), that equals \( A \) within the one standard deviation of the optimal solution, and zero otherwise. Formally,

\[
P'(X) = \begin{cases} 
A, & \text{if } \forall i \in \{1, ..., 5\} : |X_i - X^*_i| < \left((\text{diag}(\Sigma))\right)_i^{1/2} \\
0, & \text{otherwise}
\end{cases}
\]  

(4.4)

where \( A \) is the normalization factor, see "Materials and Methods". The corresponding ranking on 5000 sample points, \( \langle E_{\text{train}}^{5000} \rangle' \), is shown in Tab. 4.2.

Three conclusions can be inferred from this table: first, while all the \( E_{\text{train}} \) values are within the \( kT \) range, the proposed robustness method accentuates the difference between the optima. This is particularly clear when OPT1 and OPT4 are compared. Later, we will show how these two optima are qualitatively different in terms of their connectivity in the multidimensional landscape. Second, the ranking of the optima is conserved among 1000 and 5000 sampling scenarios which supports the convergence of the method. Third,
4.2. PROPOSED METRIC OF OPTIMUM ROBUSTNESS

both weighting functions lead to similar ranking, which demonstrates the stability of the proposed ranking method to the choice of the weighting function. As a complimentary analysis, we will now compare OPT1 and OPT4, the most and least robust optimal solutions.

Table 4.2: Robustness analysis of the lowest five optimum parameter vectors found by VTDIRECT95. $\langle E_{train}^{1000} \rangle$ and $\langle E_{train}^{5000} \rangle$ show the result of ranking using Gaussian distribution as the weighting function, while the last column, $\langle E_{train}^{5000} \rangle'$, uses $P'(X)$ defined in Eq. 4.4, all are in kcal/mol.

<table>
<thead>
<tr>
<th></th>
<th>$\langle E_{train}^{1000} \rangle$</th>
<th>$\langle E_{train}^{5000} \rangle$</th>
<th>$\langle E_{train}^{5000} \rangle'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPT 1</td>
<td>4.73</td>
<td>4.71</td>
<td>4.45</td>
</tr>
<tr>
<td>OPT 2</td>
<td>4.75</td>
<td>4.75</td>
<td>4.51</td>
</tr>
<tr>
<td>OPT 3</td>
<td>5.00</td>
<td>4.97</td>
<td>4.75</td>
</tr>
<tr>
<td>OPT 4</td>
<td>5.75</td>
<td>5.78</td>
<td>5.37</td>
</tr>
<tr>
<td>OPT 5</td>
<td>4.87</td>
<td>4.90</td>
<td>4.61</td>
</tr>
</tbody>
</table>

Objective function landscapes near optima. To demonstrate the difference between OPT1 and OPT4 revealed by our robustness metric, the behavior of the objective function around these two optima is shown in Fig. 4.4. Comparing the left and right panel, wide wells are clearly observed around OPT1, as opposed to OPT4 that has deep narrow wells around the optimum in each dimension.

4.2.2 Visualizing the Optimization Landscape

Visualization of a multidimensional landscape is problem specific as there is no single gold standard representation. We propose to reduce the complex landscape to a connectivity graph that can be constructed by a relatively limited sampling of the objective function around and between pairs of global and local minimum points. Our goal in this section is to facilitate the visualization of the 5-dimensional optimization landscape between the global and four local minimum points.
Figure 4.4: Projection of OPT1 (global min) and OPT4 (local min) onto different radii coordinates. Left panel shows the behavior of OPT1 objective function projected onto $\rho_W, \rho_C, \rho_H, \rho_N$ and $\rho_O$ within the sampling box and in the physical bound proposed in Eq. 3.3. Right panel shows similar graphs for OPT4. Radii (x coordinates) have different ranges in order to keep the objective function values (y coordinates) in a same range, which is $2kT$ form OPT1 value.
**Distance plot.** The key idea is to reduce the N-dimensional landscape to a 2-dimensional one, within a relatively narrow "corridor" between pairs of the global and a local minima, and then to visualize only those points in the corridor whose objective function values are below a pre-defined threshold. For mapping the 5-dimensional space onto a 2-dimensional visualizable plot, the Euclidean distance is calculated from the sample point to each of the two minima, see Fig. 4.5 in which the procedure is illustrated for the global minimum (OPT1) and a local minimum (OPT2). The distances between a sample point \(x\) and the two minima (OPT1 and OPT2) are calculated in a large sampling box, shown in black in Fig. 4.5. We call these two distances \(d_1\) and \(d_2\), respectively; these become the coordinates of \(x\) in the new 2D representation. The large box covers the space between the smaller sampling boxes (shown in red) bounded around OPT1 and OPT2. In Fig. 4.7 ("distance plot") only those points (with coordinates \(d_1\) and \(d_2\)) whose objective function values are within the range of \(kT\) from the objective function value at OPT1 are shown. We call these points \(kT\)-reachable. Similar plots are shown for OPT1 versus the remaining local minima OPT3, OPT4, and OPT5.

![Figure 4.5: Procedure for creating a distance plot, exemplified. OPT1 (global minimum) and OPT2 (local minimum) are selected in this demonstration. The large sampling box, shown in solid black, covers the space between the smaller sampling boxes (dashed red rectangles) around OPT1 and OPT2. These two smaller boxes are found by applying the sampling algorithm explained in the "Materials and Methods". For each sample point \(x\) in the large box, 5-dimensional Euclidean distances \(d_1\) and \(d_2\), from OPT1 and OPT2 (shown as stars) to \(x\) are calculated, and the corresponding objective function value is illustrated on the distance plot, shown in Fig. 4.7.](image)
4.2.3 Exploring the Objective Function Landscape

To have better insight into the shape of the objective function landscape, the distribution of the objective function values around each minimum is examined on one thousand sample points in the close vicinity of the minimum point. In Fig. 4.6 boxplots along with outliers are illustrated with a clear depiction of quartile ranges. It is observed that OPT1, OPT2, OPT3, and OPT5 have less deviations in the sample boxes which implies "wide" shapes; whereas OPT4 shows more deviations from the corresponding minimum value, and is relatively more "narrow". As we shall see below, this observation, based on a fairly involved analysis, is well reflected in our proposed robustness metric.

Figure 4.6: Boxplots of deviations in the objective function values around OPT1 (global minimum) and OPT2, OPT3, OPT4 and OPT5 (local minima) within the sampling box. Outliers are shown in circles.

**kT-connectivity graph.** An examination of the objective function landscape shown in Fig. 4.7 suggests that OPT1 is "connected" to OPT2, OPT3 and OPT5, but "disconnected" from OPT4, assuming $kT \sim 0.6 \text{ kcal/mol}$ as a threshold of meaningful difference in the objective function. Below we formalize this intuitive notion of connectivity of minima of a
Figure 4.7: Distance plots. Shown are only those sample points whose objective function values are within the range of $kT$ from OPT1. The 1000 and 5000 sample-point scenarios are shown in orange and blue, respectively.
multidimensional landscape. Namely, we define $kT$-connectivity graph, $G(V, E)$, where $V$ is the set of vertices and $E$ is the set of edges. $G$ is a star-shaped graph, in which $V$ represents the global (OPT1) and local min points with OPT1 in the center, see Fig. 4.8. The central vertex (OPT1, in our case) and another vertex in $G$ are connected if and only if there exists a "$kT$-path" between the two. We define $kT$-path as a continuous path between the global minimum point and another local minimum point such that all of the (sample) points along the path are $kT$-reachable, i.e., the objective function values for all the points along the path are within the range of $kT$ from the global minimum. In practice, the goal is to ascertain $kT$-connectivity with a high degree of certainty using a finite number of sample points.

![Figure 4.8: $kT$-connectivity graph. Vertices represent the global minimum point (OPT1) in the center, and local minimum points around it. An edge between OPT1 and another vertex indicates that it is possible to move between the two minima without exceeding a pre-defined threshold of the objective function, in our case $kT \sim 0.6 \text{ kcal/mol}$.](image)

**Establishing $kT$-connectivity.** This problem in general may be very difficult: for example, if $kT$-paths deviate significantly from a straight line connecting the two minima, extensive sampling of large portions of N-dimensional space may be required to establish one such path. Fortunately, in our case, $kT$-paths between OPT1 and any of OPT2, OPT3, and OPT5 appear to be obvious, see Fig. 4.7. We are relying on the fact that the Latin hypercube
4.2. Proposed Metric of Optimum Robustness

sampling (LHS) method employed here samples the 5-dimensional space quite uniformly, which means that a clear gap in $kT$-reachable points along a putative path may indicate the presence of a true barrier above $kT$ in the objective function. While in the case of just 1000 sampling points (orange dots), gaps of connectivity along the line connecting the minima are seen, increasing the sampling 5-fold (blue dots) clearly fills these gaps with $kT$-reachable points. We do not see a need to pursue a more formal proof here. However, if a formal proof of $kT$-connectivity for a given path is required, one can utilize the fact that our objective function is assumed Lipschitz-continuous, meaning that there exists a real constant $K \geq 0$ such that, for any $X_1$ and $X_2$:

$$|E(X_1) - E(X_2)| \leq K|X_1 - X_2|. \quad (4.5)$$

Consider a set of $N$-dimensional spheres $\{S_1, S_2, ..., S_n\}$, each of radius $r_K$, such that the center of each sphere lies on the $kT$-path being verified, spheres $i$ and $i + 1$ overlap, and the center of the first and last sphere coincide with the two minima for which the path is being established. In short, the set of spheres completely covers the putative path. (To be specific, one can choose $n$ such that the number of spheres needed for the minimal coverage.) Now choose $r_K$ small enough so that $2Kr_K < 0.1kT$, and choose the sampling density high enough so that each sphere contains at least one point $X_0$ for which $E(X_0)$ is within $0.9$ $kT$ of the global minimum; then, by Eq. 4.5, all points in each $S_i$ are $kT$-reachable, and since the spheres overlap, the path we have just verified is indeed a $kT$-path between the two minima. Note that the rationale for $2Kr_K < 0.1kT$ is as follows: if a $0.9kT$-reachable point $X_0$ exists within a given sphere, then the maximum distance from it to any point $X$ within this sphere is $2r_K$, and so the maximum deviation of $E(X)$ inside this sphere from $E(X_0)$ is less than $2Kr_K$ (by Lipschitz continuity), which in turn is less than $0.1kT$ by the imposed condition on $r_K$. Since $|E_g - E(X_0)| < 0.9kT$, where $E_g$ is the
global optimum, it means that $|E_g - E(X)| < 0.9kT + 0.1kT$, thus $X$, and any other point inside the sphere, is $kT$-reachable.

**Establishing $kT$-disconnectivity.** In stark contrast to OPT2, OPT3, and OPT5, the distance plot between OPT1 and OPT4 suggests that the latter are disconnected, see Fig. 4.8. While formal proof is not pursued in this work, we provide a qualitative rationale for why OPT4 is so different from the other minima in its connectivity to the global optimum. Consider a path between OPT4 and OPT1 where all of the radii except $\rho_O$ are kept at their OPT4 values, while the oxygen radius ($\rho_O = 1.81$ at OPT4) converges to its OPT1 value ($\rho_O = 1.28$). In doing so, the objective function becomes large very quickly: a $0.1$ decrease in the $\rho_O$ of OPT4 leads to more than $4kT$ deviation in the binding energy. This behavior is suggestive of the existence of a high barrier between OPT4 and OPT1. Comparing the $kT$-connectivity graph in Fig. 4.8 and Tab. 4.1 we observe that changes in $\rho_O$ play a key role in the $kT$-connectivity graph: OPT1 and OPT2 that share an identical $\rho_O$ are clearly connected, while OPT1 and OPT4, that have quite different $\rho_O$, are disconnected. This observation is also aligned with the electrostatic characteristic of oxygen which can substantially change the result of $\Delta\Delta G_{pol}$.

### 4.3 Optimized Parameters of the Dielectric Boundary Show Promise

For the most robust optimum (OPT1 in Tab. 4.1), the deviation of the corresponding electrostatic binding free energy from the reference on the training and test sets are shown in Tab. 4.3. We also tested two other commonly used radii: PARSE and ZAP-9, optimized previously against solvation energies of small molecules. These two sets of radii are cho-
sen for comparison since they have about the same number of independent atom types; to the best of our knowledge, no radii sets optimized specifically for protein-ligand binding exist. Four conclusions can be made. First, the global radii optimization methodology discussed here delivers around 1.5 kcal/mol improvement in the accuracy of the estimation of the electrostatic binding free energy on the test set compared to what can be achieved with existing radii sets with similar numbers of distinct atom types. This observation supports our key conclusion, that the proposed multidimensional global optimization procedure works as intended. Second, the remaining error is still appreciably larger than chemical accuracy of 1 kcal/mol, which means that the new radii set should be considered as a step in the right direction, but not the final solution. The fact that the global optimum is still outside the chemical accuracy is not surprising given the "bare minimum" number of atomic radii optimized, combined with the relatively simplistic two-dielectric continuum model and a small size of the training set of structures used in this proof-of-concept study. Third, the difference between the energies of training and test sets is significant – that issue will be addressed below. Finally, it is worth mentioning that OPT4 performs poorly on the test data set, RMSE = 7.92 kcal/mol. This, again, supports the use of the proposed robustness metric to eliminate the least promising optimization candidates.

Table 4.3: The accuracy (RMSE to the explicit solvent reference, Eq. 3.2) of calculating $\Delta \Delta G_{pol}$ values using the proposed optimal radii (OPT1) and two other popular sets of atomic radii. Radii are in Å and RMSE value of the training and test sets, $E_{train}$ and $E_{test}$, are in kcal/mol.

<table>
<thead>
<tr>
<th>Atomic Radii</th>
<th>$\rho_W$</th>
<th>$\rho_C$</th>
<th>$\rho_H$</th>
<th>$\rho_N$</th>
<th>$\rho_O$</th>
<th>$E_{train}$</th>
<th>$E_{test}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPT1</td>
<td>1.37</td>
<td>1.40</td>
<td>1.55</td>
<td>2.35</td>
<td>1.28</td>
<td>3.94</td>
<td>6.62</td>
</tr>
<tr>
<td>PARSE</td>
<td>1.4</td>
<td>1.7</td>
<td>1.0</td>
<td>1.5</td>
<td>1.4</td>
<td>10.80</td>
<td>8.07</td>
</tr>
<tr>
<td>ZAP-9</td>
<td>1.4</td>
<td>1.87</td>
<td>1.1</td>
<td>1.55</td>
<td>1.52</td>
<td>5.28</td>
<td>8.27</td>
</tr>
</tbody>
</table>

Re-balancing of the training and test sets. From Table 4.3 it is clear that the current training and test sets are not well balanced, in that the RMSE to the reference is almost 3
kcal/mol smaller for the training set compared to the test set, for all three radii sets. To close this gap between the training and test sets, a data-driven partitioning idea is proposed. Shown in the left panel of Fig. 4.9, the current partitioning assigns 1b11 complex to the test set. In the revised partitioning, this complex, whose $\Delta \Delta G_{pol}$ is an outlier, is assigned to both the training and test sets. The atomic radii are then re-optimized. Although the RMSE of the training set increases from 3.94 kcal/mol to 4.39 kcal/mol in this revision, a more consistent correlation with the reference explicit solvent model is observed. Moreover, the RMSE of the test set decreases from 6.62 kcal/mol to 4.98 kcal/mol that is quite close to the RMSE on the training set. The optimal atomic radii obtained by this re-balanced partitioning scheme will be explored in detail in a future study.

---

**Figure 4.9:** Re-balancing of the training and test sets, with TIP3P explicit solvent model as the reference. Left: the current partitioning method partitions the whole data set of 15 small protein-ligand complexes into the training and test subsets with a similar distribution of $\Delta \Delta G_{pol}$. Training: RMSE=3.94 kcal/mol and $r^2=0.76$. Test: RMSE=6.62 kcal/mol and $r^2=0.37$. These results are obtained using the existing global optimum radii (OPT1). Right: New partitioning puts the single outlier (1b11) in both the training and test sets. Training: RMSE=4.39 kcal/mol and $r^2=0.68$. Test: RMSE=4.98 kcal/mol and $r^2=0.57$. These results are obtained using a new global optimum radii (not shown here) found by VTDIRECT95.
Chapter 5

Testing and Application to Protein-Protein Binding

5.1 Disclaimer

The original plan was to refine the optimal radii in several ways, including using more accurate explicit water model as a reference, and exploring whether introducing more radii types might improve the accuracy further. This work is on-going. The second part of the original plan was to thoroughly test the resulting radii against a statistically significant set of protein-ligand and protein-protein complexes. I am still aiming to carry out the plan. However, in light of the COVID pandemic, I decided to apply what we have and what I learned to help address today’s most pressing challenge. I fully realize that what is presented below is at a different level of scientific rigor compared to the previous chapters, and should be considered more of a preliminary result. This chapter may be appeared as the reference [44].
5.2 Prelude

The ability to estimate protein-protein binding free energy in a computationally efficient manner is of benefit to research focused on how some viruses bind to their target proteins. Implicit solvation methodology may be particularly useful in the early stages of such research, as it can offer, quickly, valuable qualitative insights into the binding process. The methodology is known to be well suited to generating physical insights and guidance. Here we evaluate the potential of the related Molecular Mechanics Generalized Born Surface Area (MMGB/SA) approach to estimate the binding free energy between the SARS-CoV-2 spike receptor-binding domain (SARS-CoV-2 S RBD) and the human ACE2 receptor. The calculations are based on a recent flavor of the generalized Born model, GBNSR6, shown to be effective in protein-ligand binding estimates (see Chapter 2). Two options for representing the boundary of the molecule are evaluated: one based on standard bondi radii, and the other based on a newly developed set of atomic radii, optimized specifically for protein-ligand binding (see Chapter 3 and 4). We first test the entire computational pipeline on the well-studied Ras-Raf protein-protein complex, which has similar binding free energy to that of the SARS-CoV-2 S RBD and ACE2 complex. Predictions based on both radii sets are closer to experiment than a previously published estimate based on MMGB/SA, but are still about 50% off the experimental reference. Likewise, none of the two estimates for SARS-CoV-2 S RBD and ACE2 are in quantitative agreement with the available experimental $\Delta G_{\text{bind}} \approx -10.6 \text{ kcal/mol}$, the best MMGB/SA estimate is at least 5 kcal/mol away from the experiment. On the other hand, both estimates of $\Delta G_{\text{bind}}$ of the SARS-CoV-2 S RBD and ACE2 point to the expected near cancellation of the relatively large enthalpy and entropy contributions, suggesting that the approach may be trustworthy, qualitatively. Specifically, the MMGB/SA approach could be reasonably accurate for future analysis of relative binding free energies in this complex, including the effects of
mutations, relative contributions from various residues to $\Delta G_{\text{bind}}$, congeneric series of ligands, etc. The availability of a conceptually simple, fast qualitative tool for analysis of the SARS-CoV-2 S RBD and ACE2 binding may be particularly beneficial in light of the need to move forward fast.

5.3 Introduction

Emerged as a global threat to human health, the SARS-CoV-2 virus that causes COVID-19 disease has been studied widely since the start of 2020 [6]. Despite sequence and structure similarities with other viruses [30], no highly effective treatment option for the novel coronavirus is available. As of today, more than 7 million people across the globe have tested positively for the virus, and around 450,000 have died of COVID-19 [33]. This fast-growing pandemic highlights the role of computational structural biology and computer-aided drug design (CADD), which have the ability to accelerate the slow and expensive process of drug discovery [62]. In structure-based drug discovery, accuracy and speed of the binding free energy prediction of drug-like compounds (ligands) to target biomolecules plays a key role in virtual screening of drug candidates [58, 86, 192]. Despite decades of research, efficient and accurate computational prediction of binding free energies is still a challenge [113, 144, 158, 180, 198].

In theory, the binding free energy of a molecular system can be estimated directly from thermodynamic first principles [166]. However, for any realistic molecular systems, e.g., the complex of interest in this work that is made of more than 12,000 atoms [93], approximations must be made to make the estimate computationally feasible. For example, alchemical methods [1, 24], simulate changes in the free energy along a pathway that sometimes reflects non-physical properties or literally “alchemy”. The required sample
points along the pathway are generated via Monte Carlo or Molecular Dynamics (MD) simulations. Some of the popular methods in this class are thermodynamic integration (TI) and free energy perturbations (FEP) [110]. However, these simulations are still computationally expensive, specifically when it comes to high throughput virtual screening of thousands of potential drugs.

Remarkably more efficient, end-point free energy methods ignore details of the complex to unbound state pathway and estimate free energy on an ensemble of uncorrelated snapshots representing the initial (complex) and the final (unbound) states only. These snapshots can be generated by an MD simulation. Molecular mechanics Poisson-Boltzmann surface area (MMGB/SA) and molecular mechanics generalized Born surface area (MMGB/SA) [51, 177, 178] are arguably among the most popular methods in this category. They are often used in docking projects where a quick estimate of binding affinities is required [75]. Leading docking software, for instance, AutoDock Vina [172] and DOCK [4], rank the feasible poses of a ligand in a binding pocket based on a scoring function in which binding affinity plays an important role. End-point free energies can improve the accuracy of these scoring functions on-the-fly. Recently, MMGB/SA was employed to improve the accuracy of AutoDock Vina and Dock in the Drug Design Data Resource (D3R) Grand Challenge 4 (GC4) [151].

While calculations based on practical implicit solvation models such as generalized Born (GB) are arguably not as accurate as corresponding estimates based on the best available explicit solvent models, the use of implicit solvent has an undeniable appeal. And not only of computational efficiency: reasoning about physical origins of observed effects is often much more transparent in an implicit than explicit solvent [131, 138, 193]. That last advantage may be particularly valuable now, when so much about COVID-19 structure-infectivity relationship remains unknown.
Here, we employ MMGB/SA implemented in AmberTools18 [79] for binding free energy calculation of SARS-CoV-2 S RBD and ACE2 (PDB ID:6m0j), see Fig. 5.1. Through the MMGB/SA approach, the absolute binding free energy of a complex is calculated as the sum of gas-phase energy, solvation free energy, and entropic contributions averaged over several snapshots extracted from the main MD trajectory. A grid-based surface GB model is used for estimating the polar component of solvation free energy, coupled with water and atomic radii introduced earlier [47]. Human H-Ras and the Ras-binding domain of C-Raf1, so-called Ras-Raf complex [57, 58], is chosen as the reference for the initial evaluation of the MMGB/SA model. Final results are compared with those from a few available relevant studies. The main goal of the work is a quick assessment of the potential of the simple and efficient MMGB/SA method to future studies of SARS-CoV-2 to ACE2 binding.

Figure 5.1: Binding scheme of the SARS-CoV-2 spike protein to the ACE2 human receptor.
5.4 Methods and Materials

5.4.1 Binding Free Energy Decomposition

Binding free energy, \(\Delta G_{\text{bind}}\), of a molecular system is calculated as follows:

\[
\Delta G_{\text{bind}} = \Delta H - T \Delta S,
\]

(5.1)

where \(\Delta H\) is the the enthalpy change of the system, \(T\) is the absolute temperature in K, and \(\Delta S\) is the entropy change of the system. A high-level illustration of \(\Delta G_{\text{bind}}\) between bound and unbound states of a solvated complex is shown in Fig. 5.2.

![Figure 5.2](image)

Figure 5.2: Binding a ligand (shown in yellow) to a protein receptor (shown in purple) in a box of solvent (shown in blue) releases binding free energy of \(\Delta G_{\text{bind}}\). A negative sign of \(\Delta G_{\text{bind}}\) indicates that spontaneous binding occurs, the magnitude of \(|\Delta G_{\text{bind}}|\) characterizes the binding strength (affinity).

In theoretical/computational studies, a useful way of calculating \(\Delta G_{\text{bind}}\) is through a thermodynamic cycle shown in Fig. 5.3.

With this approach, \(\Delta G_{\text{bind,solv}}\) is calculated as follows:

\[
\Delta G_{\text{bind,solv}} = \Delta G_{\text{bind,vacuum}} + \Delta G_{\text{solv,complex}} - (\Delta G_{\text{solv,ligand}} + \Delta G_{\text{solv,receptor}}).
\]

(5.2)
Figure 5.3: The thermodynamic cycle used here to estimate binding free energy of a protein-ligand complex in the solvent.

The solvation free energy, $\Delta G_{solv}$, is broken into the polar and non-polar components:

$$\Delta G_{solv} = \Delta G_{pol} + \Delta G_{nonpol}. \quad (5.3)$$

The free energy in vacuum, $\Delta G_{vacuum}$, is decomposed into the gas-phase energy ($\Delta E_{MM}$) and the configurational entropy of the solute ($T \Delta S$):

$$\Delta G_{vacuum} = \Delta E_{MM} - T \Delta S. \quad (5.4)$$

Note that the $T \Delta S$ above does not exactly correspond to $T \Delta S$ in Eq. 5.1; specifically, the entropy of solvent re-arrangement [131, 138] is subsumed into $\Delta G_{solv}$, see below, which is then considered a part of $\Delta H$. Combining the free energy components defined above, we obtain: $\Delta H = \Delta E_{MM} + \Delta G_{pol} + \Delta G_{nonpol}$. Our approaches for calculating $\Delta G_{solv}$, $\Delta E_{MM}$ and $T \Delta S$ are explained in sections 5.4.3, 5.4.4 and 5.4.5, respectively.
5.4.2 MMPB/SA Free Energy Methodology

MMPB/SA is a popular end-point free energy method which estimates $\Delta G_{\text{solv}}$ by Poisson-Boltzmann implicit solvent model [89], while components of $\Delta E_{\text{MM}}$ are estimated based on a classical Molecular Mechanics force-field. (In MMGB/SA, discussed below and used here, the role of the PB is played by the faster GB). Significantly faster than the conventional Alchemical methods, MMPB/SA can be very useful, particularly in the early stages of structure-based virtual screening. As another important advantage, is that through MMPB/SA it is possible to decompose the total free energy into sub-components and measure their contributions separately [57, 58]. This feature is certainly useful when it comes to comparing several different free energy methods. Finally, MMPB/SA is applicable to a wide range of structures [119], from small host-guest systems to large protein-protein complexes with thousands of atoms [58].

Through the MMPB/SA approach, the average of $\Delta G_{\text{solv}}$ is calculated on a collection of snapshots extracted from an MD simulation. Several decisions have to be made in applying the approach in practice. First, the computational protocol must be selected between the “single-trajectory” (one trajectory of the complex), or “separate-trajectory” (three separate trajectories of the complex, receptor and ligand). In this study we choose the former protocol as it was shown [74] to be not only much faster than the alternative, but also less “noisy” due to the cancellation of inter-molecular energy contributions. This protocol applies to cases where significant structural changes upon binding are not expected. Shown in Fig. 5.4, the single-trajectory MMPB/SA starts with the initial structure of the complex in vacuum. After solvating the structure in a solvent model, an MD simulation is performed to generate the snapshots for further analysis. Then, a relatively large number (typically $N > 100$) of uncorrelated snapshots are extracted to represent the structural ensemble. Next, binding free energy of these structures are calculated in the implicit solvent
after removing the explicit solvent molecules. The average binding free energy over these snapshots is reported as the final $\Delta G_{\text{bind}}$.

![MMPB/SA flowchart](image)

Figure 5.4: MMPB/SA flowchart. The initial structure of the complex is solvated using a water model. An MD simulation is run from which a relatively large number snapshots are extracted. The average binding free energy of the snapshots is assigned as the binding free energy of the system.

With the single-trajectory protocol, the binding free energy of a protein-protein complex is formally calculated as follows:

$$\Delta G_{\text{bind}} = \langle G^{\text{complex}}(i) - G^{\text{protein}}(i) - G^{\text{protein}}(i) \rangle_i$$

where $\langle ... \rangle_i$ denotes an average over $i$ snapshots extracted from the main MD trajectory. The implementation of this protocol is available in AmberTools18 in Perl [74] and Python [111]. In this work the former is used to maintain consistency with the reference study [57] opted for tuning the MMGB/SA model.
5.4.3 Solvation Free Energy

**Polar Component.** A computationally efficient alternative to the PB, the GB implicit solvent model \[132, 137\] can be used for computing $\Delta G_{solv}$. Here, we employed a grid-based surface GB model called GBNSR6 [46], which, in a recent study [83], was shown to be the most accurate among several GB models in terms of the ability to approximate $\Delta G_{pol}$ relative to the numerical PB. See Chapter 2 for more details.

**Non-polar Component.** A common method to estimate the non-polar contribution to the solvation free energy in Eq. 5.3 is to assume that it is proportional to the solvent accessible surface area (SASA) of the molecule:

$$G_{nonpol} = \gamma \times SASA.$$  (5.6)

While there are more accurate methods to estimate the non-polar [49] contribution, here we use the simple Eq. 5.6 for the sake of simplicity and consistency with ref [57]. Also for consistency with the same work, here we use $\gamma = 0.0072 \text{ kcal/mol/Å}^2$. Atomic radii that form SASA not only play an important role in the non-polar component, but also enter the polar component through the dielectric boundary. Therefore, the right choice of atomic radii is crucial to the accuracy of binding free energy estimation. Three sets of atomic radii are used here: OPT1 [47], bondi, and mbondi2. The first two are listed in Tab. 5.1. Mbonid2 is indeed bondi whose hydrogen atoms bound to a nitrogen are expanded from 1.2 Å to 1.3 Å, see [134]. Carbon (C), hydrogen (H), oxygen (O), nitrogen (N) and sulfur (S) are the main atomic types in this study. The water probe radius is fixed to 1.4 Å.

Table 5.1: Two sets of atomic radii in Å used in this study.

<table>
<thead>
<tr>
<th></th>
<th>$\rho_C$</th>
<th>$\rho_H$</th>
<th>$\rho_N$</th>
<th>$\rho_O$</th>
<th>$\rho_S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>bondi</td>
<td>1.70</td>
<td>1.20</td>
<td>1.55</td>
<td>1.52</td>
<td>1.80</td>
</tr>
<tr>
<td>OPT1</td>
<td>1.40</td>
<td>1.55</td>
<td>2.35</td>
<td>1.28</td>
<td>1.80</td>
</tr>
</tbody>
</table>
5.4.4 Gas-Phase Energy

Gas-phase energy of the solute, $\Delta E_{MM}$, is the summation of internal energies, electrostatic energies, and van der Waals energies. $\Delta E_{MM}$ is calculated using ff99 (for the Ras-Raf complex) and ff14SB (for the SARS-CoV-2 S RBD and ACE2 complex) force fields in AMBER.

5.4.5 Configurational Entropy

Normal-mode analysis (NMA) and quasi-harmonic analysis are of the two common methods for calculating configurational entropy of the solute. Since the latter has shown poor convergence in several cases, NMA is selected for entropy calculations. The main drawback of this method is the computational cost that becomes intractable for large systems, e.g., systems with more than 8,000 atoms in MMGB/SA (Perl version) of AMBER18 are not supported for NMA. To tackle this problem, one approach is to truncate the complex so that the binding interface is preserved in its original shape [53]. Another approach is to combine NMA with an efficient sampling of the system’s energy landscape [179]. In this study, the SARS-CoV-2 S RBD and ACE2 complex is truncated for NMA calculations. See Sec. 5.5 for more details.

5.4.6 Structure Preparation and MD Simulations

Ras-Raf Complex. This complex was selected as the reference for tuning the parameters of the employed MM/GBSA model. The system was parameterized using the tleap module in AMBER18 with ff99 force field after solvation in a box of TIP3P [87] water model (12 Å buffer). The GTP molecule and the magnesium ion ($Mg^{2+}$) bound to it were eliminated for the sake of simplification.
SARS-CoV-2 S RBD and ACE2 Complex. H++ server [8] was employed to protonate the complex at pH=7.5. The server automatically generates the solvated structure in a box of OPC [80] explicit water model (10 Å buffer), with AMBER ff14SB force field. The model will be re-evaluated using ff99 force field in order to enforce full self-consistency with the reference [57].

MD Simulations. All equilibration and production simulations were executed with the GPU-enabled pmemd.cuda MD engine in AMBER18. All simulations were performed using the Langevin dynamics with an integration time step of 2 fs while the bonds involving hydrogen atoms were constrained by the SHAKE algorithm. Electrostatic interactions were approximated via the Particle Mesh Ewald (PME) method, with a non-bond cutoff set to 9 Å. The solvated complex was minimized with the following steps: 50 ps of heating (from 1 K to 300 K), 50 ps of density equilibration at 300 K with 2 kcal/mol/A^2 restraint on the complex, followed by 2 ns of constant pressure equilibration at 300 K. A production of 10 ns for the Ras-Raf complex and a production of 50 ns for the SARS-CoV-2 S RBD and ACE2 complex were carried out with recording the coordinates every 10 ps.

5.5 Results and Discussion

5.5.1 MM/GBSA on Ras-Raf

Despite possible drawbacks of using the old ff99 force field and a relatively short MD production run of 10 ns, we still selected this setting in order to maintain maximum consistency with the Gohlke et al. study [57], in order to see if the new radii set in conjunction with the GBNSR6 model might provide an improvement. A similar strategy was devised for calculating different components of ΔG_{bind}: entropy was computed with NMA on 150
snapshots, and $\Delta G_{nonpol}$ was calculated as explained in Sec. 5.4.3. Here, we study the accuracy of $\Delta G_{bind}$ calculation using a different GB model (GBNSR6) coupled with two sets of atomic radii on 500 snapshots. According to Tab. 5.2, it is observed that $\Delta G_{bind}$ calculated by GBNSR6 coupled with OPT1 radii underestimates the binding affinity whereas GBNSR6 coupled with bondi radii overestimates it. Yet, both of these results have better agreement with the experiment [149] compared to the reference MGB model in [57]. Therefore, we have decided to utilize GBNSR6 with both bondi and OPT1 radii options for the following estimate of $\Delta G_{bind}$ of the SARS-CoV-2 S RBD and ACE2 complex. As a side note, it is worth mentioning that an offset of 1.79 kcal/mol has been subtracted from the $\Delta H$ component of MGB based on the author’s recommendation in [57]. This component, which accounts for the energy due to six translational and rotational degrees of freedom, has been added automatically to the MM/PBSA model in later versions of AMBER including AMBER18 that is employed in this work.

Table 5.2: MMGB/SA results on RAS-RAF. All the components are in kcal/mol. The experimental value is from isothermal titration calorimetry [149].

<table>
<thead>
<tr>
<th>Component</th>
<th>GBNSR6 (bondi)</th>
<th>GBNSR6 (OPT1)</th>
<th>MGB (bondi) ref [57]</th>
<th>Experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{MM}$</td>
<td>-937.42</td>
<td>-937.42</td>
<td>-1308.6</td>
<td></td>
</tr>
<tr>
<td>$\Delta G_{nonpol}$</td>
<td>-9.63</td>
<td>-11.02</td>
<td>-9.5</td>
<td></td>
</tr>
<tr>
<td>$\Delta G_{pol}$</td>
<td>887.39</td>
<td>897.59</td>
<td>1275.3</td>
<td></td>
</tr>
<tr>
<td>$\Delta H$</td>
<td>-59.66</td>
<td>-50.85</td>
<td>-42.7</td>
<td></td>
</tr>
<tr>
<td>$-T\Delta S$</td>
<td>45.31</td>
<td>45.31</td>
<td>41.4</td>
<td></td>
</tr>
<tr>
<td>$\Delta G_{bind}$</td>
<td>-14.35</td>
<td>-5.54</td>
<td>-3.1</td>
<td>-9.7</td>
</tr>
</tbody>
</table>

5.5.2 MM/GBSA on SARS-CoV-2 S RBD and ACE2

Enthalpy Calculations. The RMSD of the SARS-CoV-2 S RBD and ACE2 backbone compared to its crystal structure is shown in Fig. 5.5. The trajectory is stable after 50 ns of
production, with the RMSD from the X-ray reference of around 1.75 Å. A protocol similar to that used for Ras-Raf has been implemented for $\Delta H$ calculation on 500 snapshots extracted from this MD trajectory.

Figure 5.5: Backbone RMSD of SARS-CoV-2 S RBD and ACE2 complex, relative to its experimental crystal structure, along the 50 ns production trajectory.

**Entropy Calculations.** To execute NMA of MM/GBSA (Perl version) in AMBER18, the structure must have fewer than 8,000 atoms. Accordingly, the original structure was truncated from 12,515 atoms (791 residues) to 7,286 atoms (463 residues) by removing residues, one by one, starting from the N-terminus of the spike protein, and the C-terminus of the ACE2 protein. The remaining atoms are still within 8 Å from the binding interface. A weak restraint of 0.01 kcal/mol/Å$^2$ was applied to the atoms of the truncated complex, relative to the X-ray positions, during a 50 ns production to prevent the truncated complex from falling apart.

The entropy of the truncated complex is calculated over 10 snapshots extracted from the trajectory explained above. Results are shown in Tab. 5.3. It is observed that changes
5.5. RESULTS AND DISCUSSION

Figure 5.6: Truncation of SARS-CoV-2 S RBD used in the entropy estimate. The spike protein is in cyan, and the ACE2 receptor is in green. Left: original complex. Right: truncated complex. A pair of atoms on the binding interface that are 8.8 Å apart is shown in a solid red segment to illustrate the length scale.

in vibrational degrees of freedom ($T_{\text{vib}}$) have a much larger impact on the final entropy compared to translational degrees of freedom ($T_{\text{trans}}$) and rotational degrees of freedom ($T_{\text{rot}}$). A more sophisticated truncation or restraining method could contain this component. Comparing the estimated $\Delta H$ between the truncated and original structures (results not shown) demonstrates that the difference between the two is only $3.5 \text{ kcal/mol}$, which indirectly validates the use of the truncation procedure.

Table 5.3: Binding entropy estimation of the truncated SARS-CoV-2 S RBD and ACE2 complex. The values are in $\text{kcal/mol}$.

<table>
<thead>
<tr>
<th>Contribution</th>
<th>Complex</th>
<th>Receptor</th>
<th>Ligand</th>
<th>Delta</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_{\text{trans}}$</td>
<td>17.56</td>
<td>17.29</td>
<td>16.35</td>
<td>-16.08</td>
</tr>
<tr>
<td>$T_{\text{rot}}$</td>
<td>18.89</td>
<td>18.39</td>
<td>16.76</td>
<td>-16.26</td>
</tr>
<tr>
<td>$T_{\text{vib}}$</td>
<td>5394.04</td>
<td>4003.69</td>
<td>1411.57</td>
<td>-21.22</td>
</tr>
<tr>
<td>$T_{\text{total}}$</td>
<td>5430.49</td>
<td>4039.37</td>
<td>1444.68</td>
<td>-53.56</td>
</tr>
</tbody>
</table>

**Estimated $\Delta G_{\text{bind}}$ of SARS-CoV-2 S RBD and ACE2.** Shown in Tab. 5.4, our two final estimates are presented, and compared to two other: in one study [141], an MM/GBSA computational method was used for calculating $\Delta G_{\text{bind}}$ based on a homology model of the SARS-CoV-2 S RBD and ACE2 complex. In the second study [186], $\Delta G_{\text{bind}}$ of a structure
similar to the SARS-CoV-2 S RBD and ACE2 complex has been determined experimentally. These findings are briefly discussed in Conclusion.

Table 5.4: MMGB/SA results on the SARS-CoV-2 S RBD and ACE2 complex. All the components are in kcal/mol. Experimental value derived from a fit to surface plasmon resonance sensogram [186].

<table>
<thead>
<tr>
<th></th>
<th>GBNSR6 (bondi)</th>
<th>GBNSR6 (OPT1)</th>
<th>GB$^{OBC}$ (mbondi2) ref [141]</th>
<th>Experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{MM}$</td>
<td>-611.68</td>
<td>-610.06</td>
<td>-761.06</td>
<td></td>
</tr>
<tr>
<td>$\Delta G_{\text{nonpol}}$</td>
<td>-14.31</td>
<td>-16.47</td>
<td>-12.21</td>
<td></td>
</tr>
<tr>
<td>$\Delta G_{\text{pol}}$</td>
<td>566.49</td>
<td>578.89</td>
<td>737.98</td>
<td></td>
</tr>
<tr>
<td>$\Delta H$</td>
<td>-59.50</td>
<td>-47.64</td>
<td>-35.30</td>
<td></td>
</tr>
<tr>
<td>$-T\Delta S$</td>
<td>53.56</td>
<td>53.56</td>
<td>13.56</td>
<td></td>
</tr>
<tr>
<td>$\Delta G_{\text{bind}}$</td>
<td>-5.94</td>
<td>5.92</td>
<td>-21.74</td>
<td>-10.6</td>
</tr>
</tbody>
</table>
Chapter 6

Conclusion

The main outcome of this work is a novel computational pipeline that can be employed to address highly complex and computationally demanding optimization problems where global optimization is desirable. Using the novel pipeline, we have performed, to the best of our knowledge for the first time, a global multidimensional optimization of atomic radii specifically for the purpose of computing protein-ligand binding free energies in implicit solvent. Our approach is distinctly different in several respects from the past efforts to optimize atomic radii for continuum solvent calculations. First and foremost, the introduced optimization protocol targets reference binding free energy directly, which is computationally much more demanding than using the solvation free energy of small molecules as the reference, as was done in several previous studies. The necessary computational efficiency was achieved here by the use of a highly accurate numerical generalized Born model (GBNSR6), instead of the numerical Poisson-Boltzmann model employed in the past in radii optimization efforts. Second, the highly parallel optimization approach (VTDIRECT95) used in this work is able to deliver global, rather than local optima. Global optimization of parameters of the dielectric boundary at this scale was all but impossible in the past, but is now within reach through the computational pipeline developed in this work. Third, a new general metric was introduced for robustness analysis of the multiple nearly degenerate optimum points. The metric helped us to clearly distinguish several optima otherwise indistinguishable. The exploration of the complex
multidimensional objective function landscape was facilitated by what may be a novel visualization approach.

With respect to the globally optimized atomic (and water probe) radii obtained with the new pipeline, at least two results have emerged that should be of interest to the bio-computational field. First, compared to two well-known sets of "electrostatic" atomic radii, previously developed based on hydration free energies of small molecules, the new radii result in a better agreement with the explicit solvent electrostatic free energy, used as the reference. The improvement should be viewed as a consistency check of the optimization method rather than a claim of an immediate practical value of the new radii. It is still noteworthy that the number of distinct radii, or atom types, in the proposed radii set is only five, including that of the water probe. To the extent that better agreement with the explicit solvent improves the accuracy of implicit solvation with respect to reality, the new atomic radii warrant further exploration to see if they improve outcomes of practical protein-ligand binding calculations within the GB/PB framework. At the same time, the remaining error, relative to the explicit solvent, is still appreciably above the desired chemical accuracy threshold. Given that the global optimum was found, this result points to a fundamental limitation of the common continuum solvent model at the GB/PB level.

The proposed optimization pipeline, and especially the proposed parameters (atomic radii) of the resulting "electrostatically optimal" dielectric boundary have several limitations, within the continuum solvent framework. To begin with, we expect the optimal radii to be specific to the dielectric boundary definition used here, i.e., sharp SES. Future efforts should explore to what extend the accuracy of the implicit solvent-based protein-ligand binding energies may improve if alternative definitions of the dielectric boundary are used [171]. The optimal radii are also specific to the explicit water model used here as the reference (TIP3P); a future optimization effort should consider at least two different
accurate water models as alternative accuracy targets. Another limitation of the approach is the focus on the polar component of the solvation, and the neglect of possible coupling to the non-polar part of the total binding free energy. Adding computationally feasible parts of the non-polar energy and optimizing against the resulting total may improve the outcomes. We also note that the optimization pipeline does not account for the entropy component of the binding free energy: thus if the given protein-ligand complex binding is dominated by the entropy, the optimal dielectric boundary will have little effect on the overall accuracy. However, the "electrostatically optimal" dielectric boundary proposed here may still serve as a good starting point for more sophisticated optimizations that account for the entropy component. Finally, the training and test sets of protein-ligand complexes used here are relatively small, which raises transferability concerns. This limitation is not of the optimization pipeline, but of the specific radii set proposed.

In the future it would be interesting to explore to what extend the accuracy of the implicit solvent-based protein-ligand binding energies can improve if the number of atom types with distinct radii is increased – the developed computational pipeline can easily handle global optimization even if the number of atom types is doubled. However, fundamentally, the accuracy limitations revealed by this work point to the need to develop and test, within the context of protein-ligand binding, implicit solvation models of higher accuracy than the GB/PB for the electrostatic effects. Global optimization for models comparable in efficiency to the GB, such as fast numerical PB flavors, can be handled easily by the new pipeline. In fact, it will be easy to check if the optimal radii developed here perform as well, or nearly as well within the PB. Perhaps a more interesting investigation would involve models, such as 3D-RISM, which incorporates many of the explicit solvent effects beyond the PB, and has shown promise in end-point ligand binding estimates [52]. An optimization pipeline based on VTDIRECT95 has the potential to handle such relatively
expensive optimizations, given an appropriately scaled computational resource. This is because VTDIRECT95 can efficiently utilize all of the CPUs made available to it, for sampling of the vast parameter space. That is given 100x the computational power used in this work, not only will the parallel implementation scales to 100x per single-point evaluation, but it will also scale to 100x concurrent evaluations.

Finally, we tested GBNSR6 coupled with the introduced optimal radii relative to experiment on two protein-protein complexes: Ras-Raf and SARS-CoV-2 S RBD/ACE2 complexes. Encouraged by better agreement with experiment, for Ras-Raf, compared to a previous work, we applied the approach to estimate $\Delta G_{\text{bind}}$ of the SARS-CoV-2 S RBD and ACE2 complex, critical in the mechanism of the novel coronavirus infection. A common set of atomic radii was also tested. None of the computational estimates for SARS-CoV-2 S RBD and ACE2, including two of our own, and one reported previously, are in quantitative agreement with the available experiment. Our best MMGB/SA estimate is at least 5 kcal/mol away from the experiment. Moreover, the radii set that performed best for Ras-Raf, has led to a larger discrepancy relative to the experiment. This somewhat disappointing outcome is not unexpected, given the known uncertainties inherent in these estimates, especially of the binding entropy. On the other hand, our two reported estimates of binding free energy of the SARS-CoV-2 S RBD and ACE2 complex point to the expected near cancellation of the relatively large enthalpy and entropy terms, suggesting that the qualitative results may be trustworthy. Specifically, the MMGB/SA approach could be reasonably accurate for future analysis of relative binding free energies in this system, including the effects of mutations, relative contributions from various residues to binding free energy, congeneric series of ligands, etc. We are hoping that a more thorough fine-tuning of the computational protocol against a statistically significant set of protein-ligand reference points, may ultimately improve the outcome of these efficient estimates.
Bibliography


[53] Samuel Genheden, Oliver Kuhn, Paulius Mikulskis, Daniel Hoffmann, and Ulf


[100] Bo Li, Xiaoliang Cheng, and Zhengfang Zhang. Dielectric Boundary Force in Molec-


[116] David L. Mobley, Christopher I. Bayly, Matthew D. Cooper, Michael R. Shirts, and Ken A. Dill. Small Molecule Hydration Free Energies in Explicit Solvent: An Ex-


[141] Cheng Peng, Zhengdan Zhu, Yulong Shi, Xiaoyu Wang, Kaijie Mu, Yanqing Yang, Xinben Zhang, Zhijian Xu, and W Zhu. Exploring the Binding Mechanism and Ac-


