Published in final edited form as:

J Chem Theory Comput. 2020 July 14; 16(7): 4669–4684. doi:10.1021/acs.jctc.0c00142.

Multidimensional Global Optimization and Robustness Analysis in the Context of Protein-Ligand Binding

Negin Forouzesh[†], **Abhishek Mukhopadhyay**[‡], **Layne T. Watson**[†,¶,§,‖], **Alexey V. Onufriev**[†,‡,‖]

[†]Department of Computer Science, Virginia Polytechnic Institute & State University, Blacksburg, VA 24061, USA

[‡]Department of Physics, Virginia Polytechnic Institute & State University, Blacksburg, VA 24061, USA

[¶]Department of Mathematics, Virginia Polytechnic Institute & State University, Blacksburg, VA 24061, USA

[§]Department of Aerospace and Ocean Engineering, Virginia Polytechnic Institute & State University, Blacksburg, VA 24061, USA

[‖]Center for Soft Matter and Biological Physics, Virginia Polytechnic Institute & State University, Blacksburg, VA 24061, USA

Abstract

Accuracy of protein-ligand binding free energy calculations utilizing implicit solvent models is critically affected by parameters of the underlying dielectric boundary, specifically the atomic and water probe radii. Here, a global multidimensional optimization pipeline is developed to find optimal atomic radii specifically for protein-ligand binding calculations in implicit solvent. The computational pipeline has these three key components: (1) a massively parallel implementation of a deterministic global optimization algorithm (VTDIRECT95), (2) an accurate yet reasonably fast generalized Born implicit solvent model (GBNSR6), and (3) a novel robustness metric that helps distinguish between nearly degenerate local minima via a post-processing step of the optimization. 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. As an illustration of the capabilities of the optimization pipeline, we apply it to find a global optimum in the space of just five radii: four atomic (O, H, N, and C) radii and water probe radius. The optimized radii, ρ_W = 1.37 Å, ρ_C = 1.40 Å, ρ_H = 1.55 Å, ρ_N = 2.35 Å and ρ_O = 1.28 Å, lead to a closer agreement of electrostatic binding free energies with the explicit solvent reference than two commonly used sets of radii previously optimized for small molecules. At the same time, the ability of the optimizer to find the global optimum reveals fundamental limits of the common 2-dielectric implicit solvation model: the computed electrostatic binding free energies are still almost 4 *kcal/mol* away from the

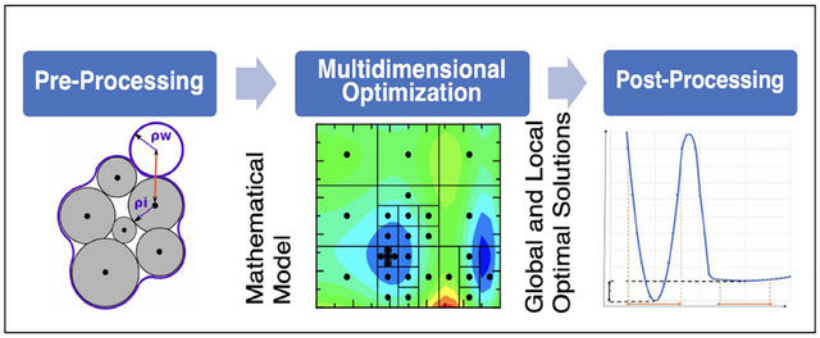
alexey@cs.vt.edu .

Supporting Information Available

It includes the pipeline made of optimization, sampling and robustness components as well as the relevant data. This material is available free of charge via the Internet at http://pubs.acs.org/.

explicit solvent reference. The proposed computational approach opens the possibility to further improve the accuracy of practical computational protocols for binding free energy calculations.

Graphical Abstract



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.[1–4] However, fast and accurate computational prediction of binding free energies continues to be challenging,[5–13] and its outcomes depend strongly on the molecular modeling technique, particularly, on how well the solvent effects are approximated.[11,14] There are two major categories of solvent models used in this field:[15] 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,[16–20] 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[21–30] is widely used due to its relative simplicity and efficiency.[31,32]

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 $\epsilon(\mathbf{r})$ 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.[33,34] The dielectric boundary is determined by the radii of the atom types comprising the protein as well as the size of the water probe.[34,35] 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.[36–41] 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,[6,42,43] 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.[44,45] 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[46] 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[47] 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 objective 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.[48]

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.

Materials and Methods

The electrostatic component of binding free energy

The total solvation free energy ΔG_{solv} of a molecule is decomposed into the polar and non-polar component:[49]

$$\Delta G_{solv} = \Delta G_{pol} + \Delta G_{nonpol}\,. \tag{1}$$

Given ΔG_{pol}, one can calculate the polar component of binding free energy, $\Delta\Delta G_{pol}$, via the following thermodynamic cycle illustrated in Fig. 1; full details can be found in.[50]

$$\Delta\Delta G_{pol} = \Delta G_{pol}^{complex} - \Delta G_{pol}^{protein} - \Delta G_{pol}^{ligand} + \Delta E_{Coulombic}\,. \tag{2}$$

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 optimization is consistent with the use of single snapshot, and fixed structures to obtain the explicit solvent reference $\Delta\Delta G_{pol}$ values[51] 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,[52] 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 in Eq. 1, where the polar and non-polar components of the total free energy are decoupled from each other; while this approximation has its limitations,[53–56] it is widely used.[15] 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 non-polar component of 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.

Implicit solvent model

The generalized Born (GB) model has become popular in implicit solvent framework due to its reasonable compromise between accuracy and speed, and the availability of its diverse flavors in leading molecular modeling packages. In this work, the polar component of the solvation energy, ΔG_{pol}, is calculated by the modification[57,58] of the generalized Born[59] model:

$$\Delta G_{pol} = \sum_{ij} \Delta G_{ij}^{pol} \approx -\frac{1}{2}\left(\frac{1}{\epsilon_{\text{in}}} - \frac{1}{\epsilon_{\text{out}}}\right)\frac{1}{1+\beta\alpha}\sum_{ij} q_i q_j \left(\frac{1}{f_{ij}^{\text{GB}}} + \frac{\alpha\beta}{A}\right), \tag{3}$$

where $\epsilon_{\text{in}} = 1$ and $\epsilon_{\text{out}} = 80$ are the dielectric constants of the solute and the solvent, respectively, $\beta = \epsilon_{\text{in}}/\epsilon_{\text{out}}$, $\alpha = 0.571412$, and A is the electrostatic size of the molecule, which is essentially the overall size of the structure that can be computed analytically. Here we employ the most widely used functional form $f_{ij}^{\text{GB}} = \left[r_{ij}^2 + R_i R_j \exp\left(-r_{ij}^2/4R_iR_j\right)\right]^{\frac{1}{2}}$, where r_{ij} is the distance between atomic charges q_i and q_j, and R_i, R_j are the so-called *effective Born radii* of atoms i and j, which represent each atom's degree of burial within the solute. The dielectric (solute/solvent) boundary enters into the model via these radii. The effective Born radii R are calculated by the "R^6" equation:[60–63]

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

where ∂V represents the chosen representation of the dielectric boundary of the molecule, $d\mathbf{S}$ is the infinitesimal surface element vector, $\mathbf{r_i}$ is the position of atom i, and $\mathbf{r}$ represents the position of the infinitesimal surface element. Note that the dielectric boundary (DB) is not an experimentally measurable entity, a number of different approaches exist[19,34] 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,[20,64–69] and we employ it here. While it was often argued[70,71] 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.[72] What is certain is that outcomes of continuum solvent calculations are very sensitive to details of the DB,[33,34] 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 [73] or smooth Gaussian dielectric boundary.[74]

Within the SES-based representation of the DB, we use a grid based molecular surface implementation of "R^6", called GBNSR6,[75] for calculating the integral in Eq. 4. The grid resolution is set to 0.5 $\AA$ by default. A detailed analysis of GBNSR6 and its input parameters can be found in Ref.[75] Briefly, GBNSR6 approximates the ideal molecular surface with orthogonal grid patches. This approximation is based on the "field-view" method[76] 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.[50] Notice that, while the PB[40,65–67,77–81] 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[77] in our optimization pipeline would have been prohibitively expensive; GBNSR6 approximates the PB reasonably well, at a small fraction of the cost.

Radial distribution function

A set of 11 small molecules was selected from a larger set of 504 small drug-like molecules,[82] see Tab. 1. The choice of these 11 structures was guided by a prior work,[83] where 10 *ns* long simulation trajectories were generated for all 504 molecules using implicit[29] water Langevin dynamics at 298 *K*. To minimize possible uncertainties[84] 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 Amber[85] v. 12 simulation package; molecule coordinate and topology files were obtained from elsewhere[82] and molecule parameters were assigned using the GAFF force field.[86] 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 $kcal/mol/Å^2$. 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.[87] Positional restraints of 200 $kcal/mol/Å^2$ 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[88] 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. 1.

Objective function

Considering the five radii (ρ_w, ρ_C, ρ_H, ρ_N, ρ_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

$$E_C(p) = \sqrt{\frac{1}{N}\sum_{c_i \in C}\left(\Delta\Delta G_{pol}^{GBNSR6}(c_i, p) - \Delta\Delta G_{pol}^{TIP3P}(c_i)\right)^2}, \quad (5)$$

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 (ρ_w, ρ_C, ρ_H, ρ_N, ρ_O). $\Delta\Delta G_{pol}^{TIP3P}(c_i)$ 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. (9). 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.

Sampling around the minimum points

To have better insight into the behavior of the objective function, the robustness analysis was performed on one thousand sample points in the close vicinity of the best minimum points. Latin hypercube sampling (LHS),[89] a common algorithm for high dimensions,[90] was selected from the QNSTOP package[91]. 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. 2.

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: fix four of the five variables around this point alternatively, and change the fifth one so that the deviation from the optimal solution reaches 1.2 *kcal/mol* (~ $2kT$). This strategy guarantees quite a wide region to gain meaningful samples, while avoiding potential overlaps between global and local solutions. Expectedly, this strategy produces an asymmetric rectangular sampling box, as the electrostatic characteristics of the atomic types are different:

$$\text{lower bounds} = (\rho_W - 0.6\ \text{Å},\ \rho_C - 0.5\ \text{Å},\ \rho_H - 0.1\ \text{Å},\ \rho_N - 1.0\ \text{Å},\ \rho_O - 0.05\ \text{Å}),$$

$$\text{upper bounds} = (\rho_W + 0.2\ \text{Å},\ \rho_C + 0.5\ \text{Å},\ \rho_H + 0.1\ \text{Å},\ \rho_N + 0.3\ \text{Å},\ \rho_O + 0.05\ \text{Å}).$$

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 Ref.[51] This data set was used previously in similar contexts.[50,51,92] 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[51] to estimate the electrostatic binding free energies employed here as the reference. Also, the structures were restrained[51] to mitigate uncertainties due to conformational variability. Unless otherwise specified, the data set is partitioned into two subsets of eight (1pbk, 1fkf, 1bkf, 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 $\Delta\Delta G_{pol}$ values between the two subsets.

VTDIRECT95: global multidimensional optimization method

The deterministic DIRECT (Dividing Rectangles) global minimization algorithm[46] is a powerful optimization method for a moderate number of dimensions. DIRECT guarantees[46] 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.

VTDIRECT95[47] 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.[93–95] 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.[96] As with any mathematical software, VTDIRECT95 has a few input parameters whose understanding 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. (5), its argument being the vector of parameters: (ρ_W, ρ_C, ρ_H, ρ_N, ρ_O). We tune three parameters to improve efficiency of the global optimization with VTDIRECT95:

- *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.

- *eps_fmin* = 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.

- *min_sep* = 0.5: In computing multiple (k) lowest minima corresponding to the global and local minimum points, without limiting the distance between them, VTDIRECT 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 $\AA$ far apart, on average, per dimension (that is per atom type). This constraint leads to a minimum 0.5 $\AA$ distance between two such points in a 5-dimensional space, i.e., $d(p_1, p_2) = \sqrt{\left(p_1^1 - p_2^1\right)^2 + \left(p_1^2 - p_2^2\right)^2 + \left(p_1^3 - p_2^3\right)^2 + \left(p_1^4 - p_2^4\right)^2 + \left(p_1^5 - p_2^5\right)^2}$.
$= \sqrt{\left(5 \times (0.2)^2\right)} \approx 0.5$

  min_sep is the corresponding parameter in VTDIRECT95 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 *eval_lim* and *eps_fmin* 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 *min_sep* = 0.5 we select those best minima that are “meaningfully” far apart. The remaining parameters of VTDIRECT95 are left as default. See Supplementary Information for a complete list of parameters used in this work.

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 the 1-dimensional optimization scenario shown in Fig. 4. 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 Δ, 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 Δ 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 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:[97] free energy

$$F = -\xi \ln \sum_x e^{-\left(E(x) - E_g\right)/\xi} \tag{6}$$

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. (6) 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 ξ in Eq. (6) in general. For example, simply equating $E(x)$ in Eq. (6) with an objective function that corresponds to the cost of car production is difficult to justify. Note that, in Physics, $E(x)$ and ξ have very specific properties that factor into the specific form of Eq. (6). Despite these conceptual difficulties, free-energy like functions have been used in machine learning[98] and optimization[99] 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,[100] networks[101] and graphs.[102] 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. (6) give insights into the general structure of mathematical expressions that might be useful in comparing widths and depths of minima. The factor $e^{-\left(E(x) - E_g\right)/\xi}$ in Eq. (6) penalizes heavily all the

contributions to the sum in F that exceed the global minimum E_g of $E(x)$ by more than ξ; the value of ξ 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, $\langle E \rangle = \int E(\mathbf{X})P(\mathbf{X})d\mathbf{X}$ where $P(\mathbf{X})$ is the probability distribution appropriate for the specific problem; $P(\mathbf{X})$ 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 $\mathbf{X_1}$ and $\mathbf{X_2}$, respectively; then, by the proposed criterion, if $\langle E_1 \rangle < \langle E_2 \rangle$ then minimum point $\mathbf{X_1}$ is preferred over minimum point $\mathbf{X_2}$. Otherwise, $\mathbf{X_2}$ is preferred. Qualitatively speaking, $\langle E \rangle$ is a robustness metric compromising between "width" and "depth". Using Fig. 4 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 $\mathbf{X_1}$ if the probability that $\langle E_1 \rangle < \langle E_2 \rangle$ is greater than 1/2." That is if the minimum is chosen by this criterion, chances are it delivers the lowest deviation from the reference, statistically speaking. The proof is particularly straightforward if one assumes normal distribution for $P(\mathbf{X})$: $P(E_1 < E_2) = \frac{1}{2}\mathrm{erfc}\left(\frac{\mu}{2\sigma}\right)$, where $\mu = \langle E_1 \rangle - \langle E_2 \rangle$, and σ^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 $\mathbf{X}^*$ (in n dimensions) given by

$$\left\langle E \mid \mathbf{X}^* \right\rangle = \int_B E(\mathbf{X})P(\mathbf{X})d\mathbf{X} = \int_B A\,E(\mathbf{X})e^{-(1/2)(\mathbf{X}-\mathbf{X}^*)^t \Sigma^{-1}(\mathbf{X}-\mathbf{X}^*)}d\mathbf{X}, \tag{7}$$

where Σ is a $n \times n$ diagonal matrix with Σ_{jj} being the empirical variance of $\mathbf{X}_j^*$, for $j \in 1, \ldots, n$. The specific form of $P(\mathbf{X}) = Ae^{-(1/2)(\mathbf{X}-\mathbf{X}^*)^t \Sigma^{-1}(\mathbf{X}-\mathbf{X}^*)}$, 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 $P(\mathbf{X})$ based on the data could be used instead, as long as it satisfies the obvious normalization condition $\int_B P(\mathbf{X})d\mathbf{X} = 1$. In what follows we verify the robustness of the proposed metric to the specific choice of $P(\mathbf{X})$. Without loss of generality and for the sake of simplicity and 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. 5 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. (7) for the illustration, $b \sim \mathcal{N}\left(0, \tilde{\sigma}^2\right)$. 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 average of the objective function values is

$$\left\langle E \right\rangle \cong A \int_{b \in \tilde{B}} E(b) e^{-\frac{b^2}{2\tilde{\sigma}^2}} db, \tag{8}$$

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. 5 right panel. Then, $\tilde{\sigma}$ is computed as a standard deviation of b_k.

Numerical estimate of $\langle E \rangle$

Here we estimate the expected value $\left\langle E \right\rangle = \int_B E(\mathbf{X}) P(\mathbf{X}) d\mathbf{X}$ of $E(\mathbf{X})$ over the box B of volume $V(B)$, where $P(\mathbf{X})$ is the probability density function (PDF) of $\mathbf{X}$ in B taken from Eq. (7). $\left\langle E \right\rangle \cong \frac{V(B)}{N} \sum_{i=1}^{N} E(\mathbf{X}_i) P(\mathbf{X}_i)$. The PDF is normalized so that $\frac{V(B)}{N} \sum_{i=1}^{N} P(\mathbf{X}_i) = 1$, for random variables $\mathbf{X}$. We use $N = 1000$ points everywhere, except for the purposes of testing convergence where $N = 5000$ is used.

Results and Discussion

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. This section introduces those components of the proposed optimization pipeline that are completely new, followed by an illustrative application to a concrete example, including an analysis. Existing methodological components, such as the GB model or VTDIRECT95 method, are described in "Materials and Methods", along with several technical details. The gist of our proposed pipeline is shown in Fig. 6.

Bounds on physically meaningful values of atomic radii

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 (ρ_w) and the intrinsic atomic radii (ρ_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. 7. 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 ± standard deviation over that data. In the left panel of Fig. 7 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_{min}, R_{max}). The RDFs are computed using molecular dynamics simulations in TIP3P[96] explicit solvent. As the result, the following upper bounds and lower bounds are obtained:

$$\begin{aligned} &0.2\ \text{Å} \le \rho_w \le 1.6\ \text{Å},\\ &2.2\ \text{Å} \le \rho_w + \rho_C \le 3.8\ \text{Å},\\ &1.4\ \text{Å} \le \rho_w + \rho_H \le 3.0\ \text{Å},\\ &2.2\ \text{Å} \le \rho_w + \rho_N \le 3.8\ \text{Å},\\ &2.2\ \text{Å} \le \rho_w + \rho_O \le 3.8\ \text{Å}. \end{aligned} \tag{9}$$

The bounds for the water radius ρ_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 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[41] and ZAP-9.[40]

Application to optimization of atomic radii

Here we use VTDIRECT95 for global optimization of the probe and atomic radii. Results are shown in Tab. 2. The practically indistinguishable optima are re-ranked later using the proposed robustness metric.

In what follows, a 5-dimensional form of Eq. (8) will be applied as the robustness metric for ranking the optimal solutions. The generalization of $\tilde{\sigma}^2$ in Eq. (8) is $\tilde{\Sigma}$ being the empirical variance of the global optimal solution $\mathbf{X}^*$ from the test set. Here $\tilde{\Sigma}$ is a 5-dimensional diagonal matrix where $\text{diag}(\tilde{\Sigma}) = \left(\sigma_W^2, \sigma_C^2, \sigma_H^2, \sigma_N^2, \sigma_O^2\right)$, see Eq. (7). In other words, $\text{diag}(\tilde{\Sigma})$

shows the variance of each radius resulting from the use of possible new test sets. The integration domain in Eq. (8) 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. 5).

To estimate $\widetilde{\Sigma}$ 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. 5. Because of the $E(\mathbf{X})$ shift in going from the training to the test sets, $E^k(\mathbf{X}^*) - E(\mathbf{X}^*) = \delta_k > 0$, where $E^k(\mathbf{X})$ refers to the test case k, $k \in \{1, \ldots, 7\}$. To find $\mathbf{b}_k$ we require that $E(\mathbf{b}_k) = \delta_k$, 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 $\mathbf{b}$ which finally results in $\mathrm{diag}(\widetilde{\Sigma}) = (0.0096,\ 0.0024,\ 0.0025,\ 0.0324,\ 0.0009)$. We apply the same $\widetilde{\Sigma}$ to evaluate robustness of all the optima in Tab. 3 – the use of the same $\widetilde{\Sigma}$ 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_{train}, and the corresponding ranking on 1000 and 5000 sample points, $\left\langle E_{train}^{1000} \right\rangle$ and $\left\langle E_{train}^{5000} \right\rangle$, for the lowest five optima, OPT1 to OPT5, are shown in Tab. 3. In order to study the effect of the underlying sharply decaying weighting function on the final ranking, we considered a modified $P(\mathbf{X})$, $P'(\mathbf{X})$, that equals A within the one standard deviation of the optimal solution, and zero otherwise. Formally,

$$P'(\mathbf{X}) = \begin{cases} A, & \text{if } \forall i \in \{1, \ldots, 5\} : \left|\mathbf{X_i} - \mathbf{X_i^*}\right| < ((diag(\widetilde{\Sigma})))_i^{1/2} \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

where A is the normalization factor, see “Materials and Methods”. The corresponding ranking on 5000 sample points, $\left\langle E_{train}^{5000} \right\rangle'$, is shown in Tab. 3.

Three conclusions can be inferred from this table: first, while all the E_{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 solutions 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, 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.

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. 8. 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.

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.

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. 9 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. 9. 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. 10 ("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.

kT-connectivity graph.—An examination of the objective function landscape shown in Fig. 10 suggests that OPT1 is "connected" to OPT2, OPT3 and OPT5, but "disconnected" from OPT4, assuming $kT \sim 0.6$ *kcal/mol* as a threshold of meaningful difference in the objective function. Below we formalize this intuitive notion of connectivity of minima of a 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. 11. 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.

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. 10. We are relying on the fact that the Latin hypercube

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:

$$|E(X_1) - E(X_2)| \leq K|X_1 - X_2|. \tag{11}$$

Consider a set of N-dimensional spheres $\{S_1, S_2, \ldots, S_n\}$, 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 coverage is minimal.) 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. (11), 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 0.1 kT 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. 11. 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 ρ_O are kept at their OPT4 values, while the oxygen radius (ρ_O = 1.81 Å at OPT4) converges to its OPT1 value (ρ_O = 1.28 Å). In doing so, the objective function becomes large very quickly: a 0.1 Å decrease in the ρ_O of OPT4 leads to more than 4 kT 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. 11 and Tab. 2 we observe that changes in ρ_O play a key role in the kT-connectivity graph: OPT1 and OPT2 that share an identical ρ_O are clearly connected, while OPT1 and OPT4, that have quite different ρ_O, are disconnected. This observation is also aligned with the electrostatic characteristic of oxygen which can substantially change the result of $\Delta\Delta G_{pol}$.

Optimized parameters of the dielectric boundary show promise

For the most robust optimum (OPT1 in Tab. 2), the deviation of the corresponding electrostatic binding free energy from the reference on the training and test sets are shown in Tab. 4. 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 chosen 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.

Re-balancing of the training and test sets.—From Table 4 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. 12, 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.

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 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.[33] 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.[103] 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. Ultimately, we hope that the optimization methodology proposed in this work will help reduce the error of the implicit solvation approach relative to the experiment in protein-ligand binding estimates.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

The authors acknowledge Advanced Research Computing at Virginia Tech for providing computational resources and technical support that have contributed to the results reported within this paper. We thank Dr. Igor Tolokh for his help in setting up the computational pipeline. This work was supported by the NIH R21 GM131228 to A.V.O.

References

(1). Shirts MR; L. MD; Brown SP In Structure Based Drug Design, 1st ed.; Merz KM, Ringe D, Reynolds CH, Eds.; Lecture Notes in Computer Science; Cambridge University Press: Cambridge, New York USA, 2010; pp 61–85.

(2). Jorgensen WL Science 2004, 303, 1813–1818. [PubMed: 15031495]

(3). Gohlke H; Kiel C; Case DA J. Mol. Biol 2003, 330, 891–913. [PubMed: 12850155]

(4). Zhou H-X; Gilson MK Chem. Rev 2009, 109, 4092–4107. [PubMed: 19588959]

(5). Mobley DL; Dill KA Structure 2009, 17, 489–498. [PubMed: 19368882]

(6). Merz KM Jr J. Chem. Theory Comput 2010, 6, 1769–1776.

(7). Rocklin GJ; Mobley DL; Dill KA J. Chem. Theory Comput 2013, 9, 3072–3083. [PubMed: 24015114]

(8). Deng Y; Roux BJ Phys. Chem. B 2009, 113, 2234–2246.

(9). Wickstrom L; He P; Gallicchio E; Levy RM J. Chem. Theory Comput 2013, 9, 3136–3150. [PubMed: 25147485]

(10). Yin J; Henriksen NM; Muddana HS; Gilson MK J. Chem. Theory Comput 2018, 14, 3621–3632. [PubMed: 29874074]

(11). Rizzi A; Murkli S; McNeill JN; Yao W; Sullivan M; Gilson MK; Chiu MW; Isaacs L; Gibb BC; Mobley DL; Chodera JDJ Comput.-Aided Mol. Des 2018, 32, 937–963.

(12). Zou J; Tian C; Simmerling CJ Comput.-Aided Mol. Des 2019, 33, 1021–1029.

(13). Wang J; Alekseenko A; Kozakov D; Miao Y Front. Mol. Biosci 2019, 6, 112. [PubMed: 31737642]

(14). Mobley DL; Gilson MK Annu. Rev. Biophys 2017, 46, 531–558. [PubMed: 28399632]

(15). Onufriev AV; Izadi S Wiley Interdiscip. Rev. Comput. Mol. Sci 2018, 8, e1347.

(16). Cramer CJ; Truhlar DG Chem.Rev 1999, 99, 2161–2200. [PubMed: 11849023]

(17). Simonson T Rep. Prog. Phys 2003, 66, 737–787.

(18). Baker N; Bashford D; Case D Implicit Solvent Electrostatics in Biomolecular Simulation. New Algorithms for Macromolecular Simulation 2006; pp 263–295.

(19). Li L; Li C; Zhang Z; Alexov EJ Chem. Theory Comput 2013, 9, 2126–2136.

(20). Gilson MK Curr.Opin.Struct.Biol 1995, 5, 216–223. [PubMed: 7648324]

(21). Hoijtink G; De Boer E; Van der Meij P; Weijland W Recl. Trav. Chim. Pays-Bas 1956, 75, 487–503.

(22). Tucker SC; Truhlar DG Chem. Phys. Lett 1989, 157, 164–170.

(23). Still WC; Tempczyk A; Hawley RC; Hendrickson TJ Am. Chem. Soc 1990, 112, 6127–6129.

(24). Dominy BN; Brooks CL J. Phys. Chem. B 1999, 103, 3765–3773.

(25). Tsui V; Case DJ Am. Chem. Soc 2000, 122, 2489–2498.

(26). Gallicchio E; Levy RM J. Comp. Chem 2004, 25, 479–499. [PubMed: 14735568]

(27). Nymeyer H; Garcia AE Proc. Natl. Acad. Sci. U.S.A 2003, 100, 13934–13949. [PubMed: 14617775]

(28). Feig M; Im W; Brooks CL J. Chem. Phys 2004, 120, 903–911. [PubMed: 15267926]

(29). Onufriev A; Bashford D; Case DA Proteins: Struct., Funct., Bioinf 2004, 55, 383–394.

(30). Nguyen H; Roe DR; Simmerling CJ Chem. Theory Comput 2013, 9, 2020.

(31). Onufriev A In Modeling Solvent Environments, 1st ed.; Feig M, Ed.; Wiley: USA, 2010; pp 127–165.

(32). Onufriev AV; Case DA Annu. Rev. Biophys 2019, 48, 275–296. [PubMed: 30857399]

(33). Tjong H; Zhou H-XJ Chem. Theory Comput 2008, 4, 507–514.

(34). Onufriev AV; Aguilar BJ Theor. Comput. Chem 2014, 13, 1440006.

(35). Lee B; Richards FM J. Mol. Biol 1971, 55, 379–IN4. [PubMed: 5551392]

(36). Swanson JM; Wagoner JA; Baker NA; McCammon JA J. Chem. Theory Comput 2007, 3, 170–183. [PubMed: 26627162]

(37). Tan C; Yang L; Luo RJ Phys. Chem. B 2006, 110, 18680–18687.

(38). Yamagishi J; Okimoto N; Morimoto G; Taiji MJ Comput. Chem 2014, 35, 2132–2139.

(39). Nina M; Beglov D; Roux BJ Phys. Chem. B 1997, 101, 5239–5248.

(40). Grant JA; Pickup BT; Nicholls AJ Comput. Chem 2001, 22, 608–640.

(41). Sitkoff D; Sharp KA; Honig BJ Phys. Chem 1994, 98, 1978–1988.

(42). Harris RC; Mackoy T; Fenley MO J. Chem. Theory Comput 2015, 11, 705–712. [PubMed: 26528091]

(43). Harris RC; Mackoy T; Fenley MO Computational and Mathematical Biophysics 2013, 1, 63–74.

(44). Easterling DR; Watson LT; Madigan ML; Castle BS; Trosset MW Comput. Optim. Appl 2014, 57, 469–492.

(45). Murty KG Operations Research: Deterministic Optimization Models; Prentice-Hall, Inc., 1994.

(46). Jones DR; Perttunen CD; Stuckman BE J. Optmiz. Theory App 1993, 79, 157–181.

(47). He J; Watson LT; Sosonkina M Comput. Optim. Appl. Trans. Math. Software 2009, 36, 17.

(48). Kwakkel JH; Eker S; Pruyt E Robustness Analysis in Decision Aiding, Optimization, and Analytics; Springer, 2016; pp 221–237.
(49). Zhou H-X; Pang X Chem. Rev 2018, 118, 1691–1741. [PubMed: 29319301]
(50). Izadi S; Harris RC; Fenley MO; Onufriev AV J. Chem. Theory Comput 2018, 14, 1656–1670. [PubMed: 29378399]
(51). Izadi S; Aguilar B; Onufriev AV J. Chem. Theory Comput 2015, 11, 4450–4459. [PubMed: 26575935]
(52). Homeyer N; Gohlke H Mol. Inf 2012, 31, 114–122.
(53). Dzubiella J; Swanson JM; McCammon JA J Chem Phys 2006, 124, 084905. [PubMed: 16512740]
(54). Cheng LT; Dzubiella J; Mccammon JA; Li BJ Chem. Phys 2007, 127, 084503.
(55). Chen Z; Baker NA; Wei GW J. Comput. Phys 2010, 229, 8231–8258. [PubMed: 20938489]
(56). Zhou S; Cheng L-T; Dzubiella J; Li B; McCammon JA J. Chem. Theory Comput 2014, 10, 1454–1467. [PubMed: 24803864]
(57). Sigalov G; Scheffel P; Onufriev AJ Chem. Phys 2005, 122, 094511.
(58). Sigalov G; Fenley A; Onufriev AJ Chem. Phys 2006, 124, 124902.
(59). Still WC; Tempczyk A; Hawley RC; Hendrickson TJ Am. Chem. Soc 1990, 112, 6127–6129.
(60). Grycuk TJ Chem. Phys 2003, 119, 4817–4826.
(61). Mongan J; Svrcek-Seiler WA; Onufriev AJ Chem. Phys 2007, 127, 11B608.
(62). Aguilar B; Shadrach R; Onufriev AV J. Chem. Theory Comput 2010, 6, 3613–3630.
(63). Aguilar B; Onufriev AV J. Chem. Theory Comput 2012, 8, 2404–2411. [PubMed: 26588972]
(64). Fogolari F; Briggs JM Chemical Physics Letters 1997, 281, 135–139.
(65). Baker NA; Sept D; Joseph S; Holst MJ; McCammon JA Proceedings of the National Academy of Sciences 2001, 98, 10037–10041.
(66). Rocchia W; Sridharan S; Nicholls A; Alexov E; Chiabrera A; Honig BJ Comput. Chem 2002, 23, 128–137.
(67). Nicholls A; Honig BJ Comput. Chem 1991, 12, 435–445.
(68). Wang J; Tan C; Chanco E; Luo R Phys. Chem. Chem. Phys 2010, 12, 1194–1202. [PubMed: 20094685]
(69). Chen D; Chen Z; Chen C; Geng W; Wei G-WJ Comput. Chem 2011, 32, 756–770.
(70). Swanson JM; Mongan J; McCammon JA J. Phys. Chem. B 2005, 109, 14769–14772. [PubMed: 16852866]
(71). Nina M; Im W; Roux B Biophys. Chem 1999, 78, 89–96. [PubMed: 17030305]
(72). Pang X; Zhou H-XX Commun. Comput. Phys 2013, 13, 1–12. [PubMed: 23293674]
(73). Cooper CD; Bardhan JP; Barba LA Comput. Phys. Commun 2014, 185, 720–729. [PubMed: 25284826]
(74). Hazra T; Ullah SA; Wang S; Alexov E; Zhao SJ Math. Biol 2019, 79, 631–672.
(75). Forouzesh N; Izadi S; Onufriev AV J. Chem. Inf. Model 2017, 57, 2505–2513. [PubMed: 28786669]
(76). Cai Q; Ye X; Wang J; Luo RJ Chem. Theory Comput 2011, 7, 3608–3619.
(77). Zhou Y; Feig M; Wei G-WJ Comput. Chem 2008, 29, 87–97.
(78). Luo R; David L; Gilson MK J. Comput. Chem 2002, 23, 1244–1253. [PubMed: 12210150]
(79). Bashford D; Karplus M Biochemistry 1990, 29, 10219–10225. [PubMed: 2271649]
(80). Li B; Cheng X; Zhang Z SIAM J. Appl. Math 2011, 71, 2093–2111. [PubMed: 24058212]
(81). Gilson MK; Honig BH Proteins 1988, 4, 7–18. [PubMed: 3186692]
(82). Mobley DL; Bayly CI; Cooper MD; Shirts MR; Dill KA J. Chem. Theory Comput 2009, 5, 350–358. [PubMed: 20150953]
(83). Mobley DL; Dill KA; Chodera JD J. Phys. Chem. B 2008, 112, 938–946. [PubMed: 18171044]
(84). Mukhopadhyay A; Aguilar BH; Tolokh IS; Onufriev AVJ Chem. Theory Comput 2014, 10, 1788–1794.

(85). Case DA; Cheatham TE; Darden T; Gohlke H; Luo R; Merz KM; Onufriev A; Simmerling C;, B.; Woods RJ J. Comput. Chem 2005, 26, 1668–1688. [PubMed: 16200636]

(86). Wang J; Wolf RM; Caldwell JW; Kollman PA; Case DA J. Comput. Chem 2004, 25, 1157–1174. [PubMed: 15116359]

(87). Miyamoto S; Kollman PA J. Comput. Chem 1992, 13, 952–962.

(88). Roe DR; Cheatham III TE J. Chem. Theory Comput 2013, 9, 3084–3095. [PubMed: 26583988]

(89). McKay MD; Beckman RJ; Conover WJ Technometrics 1979, 21, 239–245.

(90). Kucherenko S; Albrecht D; Saltelli A arXiv preprint arXiv:1505.02350 2015,

(91). Amos BD; Easterling DR; Watson LT; Thacker WI; Castle BS; Trosset MW ACM Trans. Math. Software 2020, 46, 17:1–17:20.

(92). Katkova E; Onufriev A; Aguilar B; Sulimov VJ Mol. Graph. Model 2017, 72, 70–80.

(93). He J; Watson LT; Ramakrishnan N; Shaffer CA; Verstak A; Jiang J; Bae K; Tranter WH Comput. Optim. Appl 2002, 23, 5–25.

(94). He J; Verstak A; Watson LT; Sosonkina M Int. J. High Perform. C 2009, 23, 14–28.

(95). He J; Verstak A; Sosonkina M; Watson LT Int. J. High Perform. C 2009, 23, 29–41.

(96). Jorgensen WL; Chandrasekhar J; Madura JD; Impey RW; Klein ML J. Chem. Phys 1983, 79, 926–935.

(97). Landau L; Lifshitz E Publisher: Butterworth-Heinemann 1980, 3.

(98). Shimazaki H arXiv preprint arXiv:1902.11233 2019,

(99). Zhang Y; Saxe AM; Advani MS; Lee AA Mol. Phys 2018, 116, 3214–3223.

(100). Forouzesh N; Kazemi MR; Mohades A Structure-Based Analysis of Protein Binding Pockets Using Von Neumann Entropy. International Symposium on Bioinformatics Research and Applications. 2014; pp 301–309.

(101). Passerini F; Severini S International Journal of Agent Technologies and Systems (IJATS) 2009, 1, 58–67.

(102). Du W; Li X; Li Y; Severini S Linear Algebra Appl 2010, 433, 1722–1725.

(103). Genheden S; Luchko T; Gusarov S; Kovalenko A; Ryde U J. Phys. Chem. B 2010, 114, 8505–8516. [PubMed: 20524650]

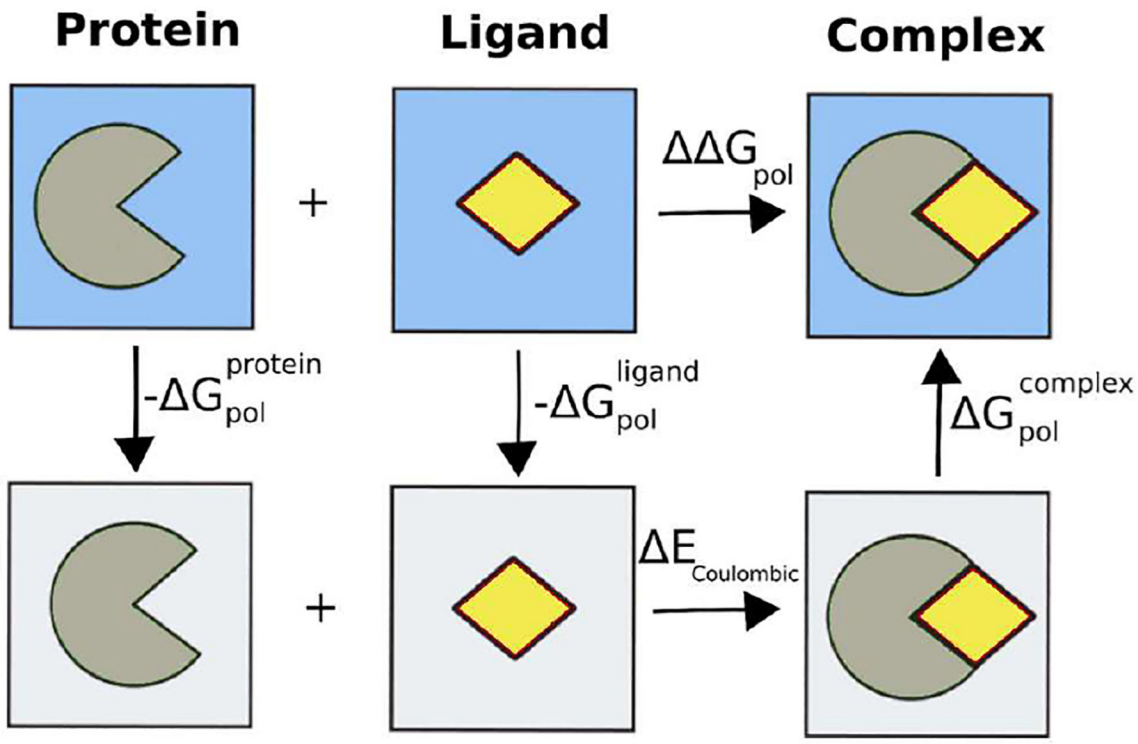


Figure 1: Thermodynamic cycle for calculating the polar component of binding free energy.
The vacuum environment is shown in white background, and the water is in blue.

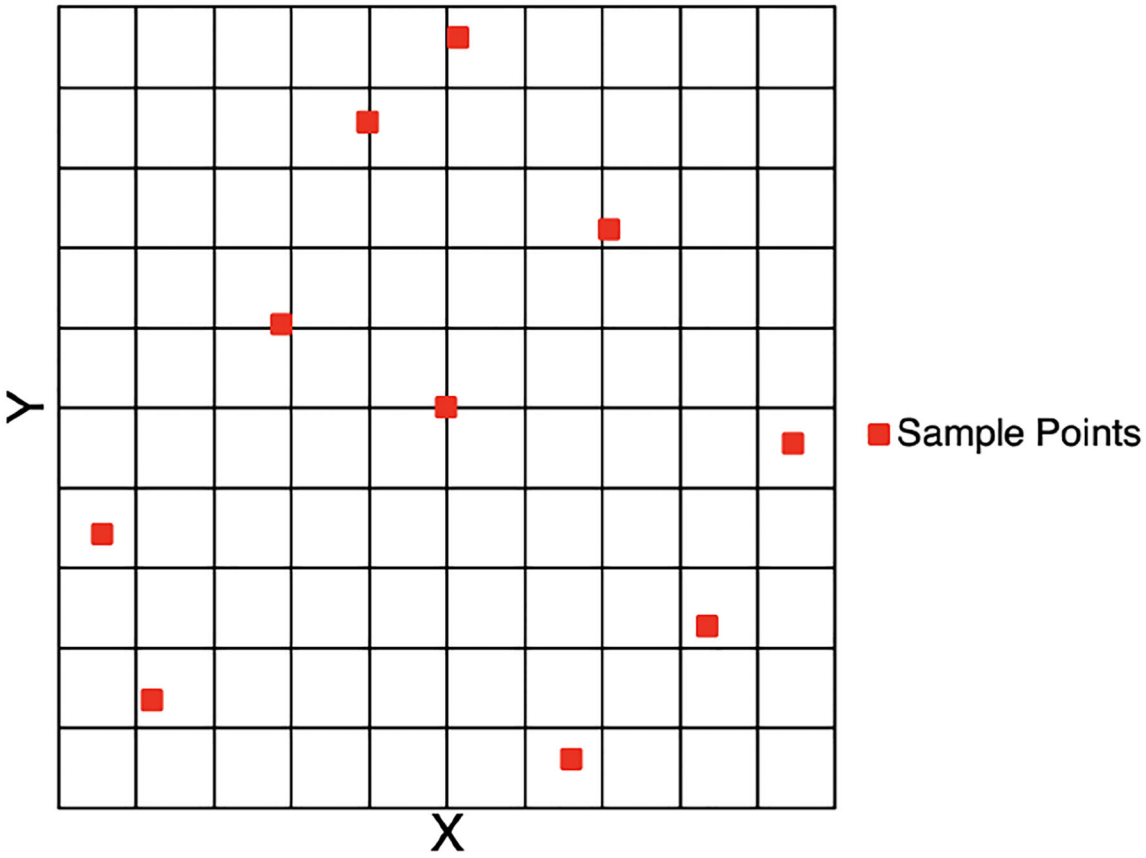


Figure 2: Latin hypercube sampling (LHS).

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.

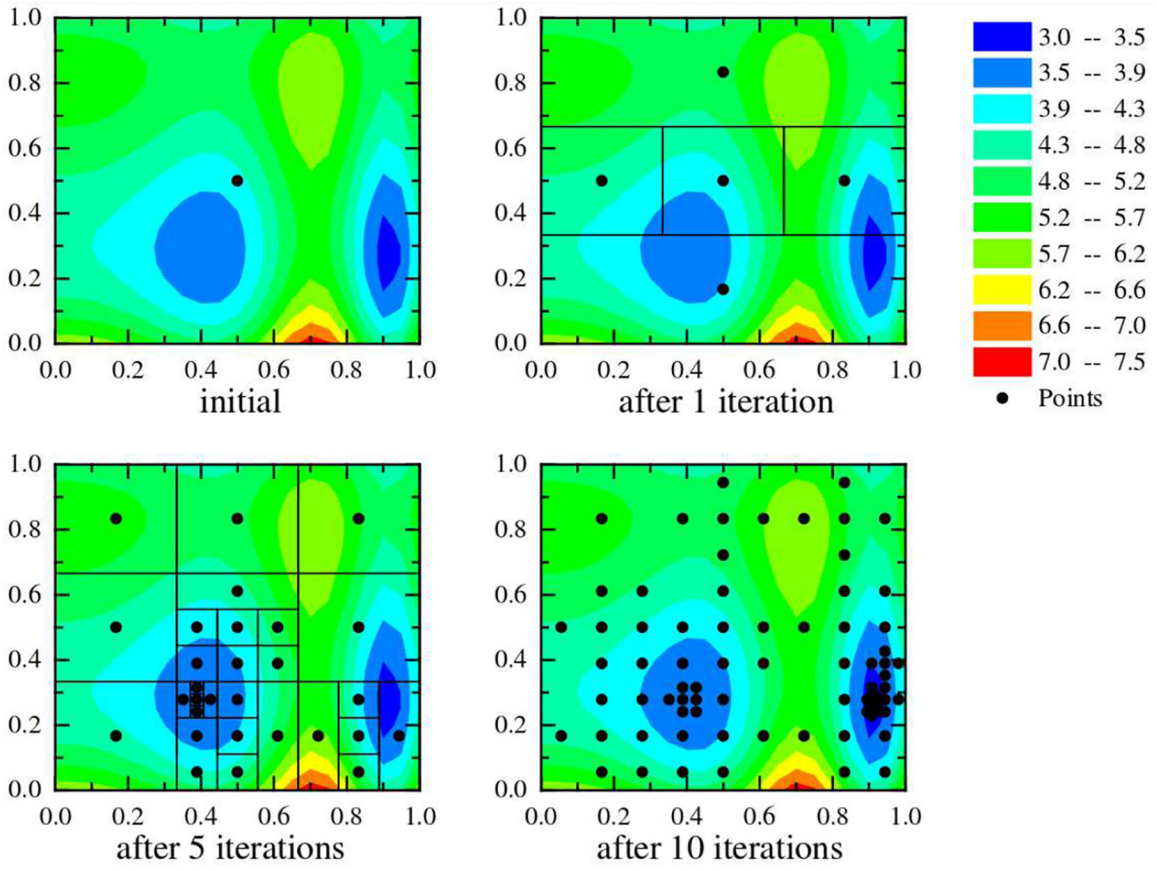


Figure 3: Function evaluations performed by DIRECT after 0, 1, 5 and 10 iterations.

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).

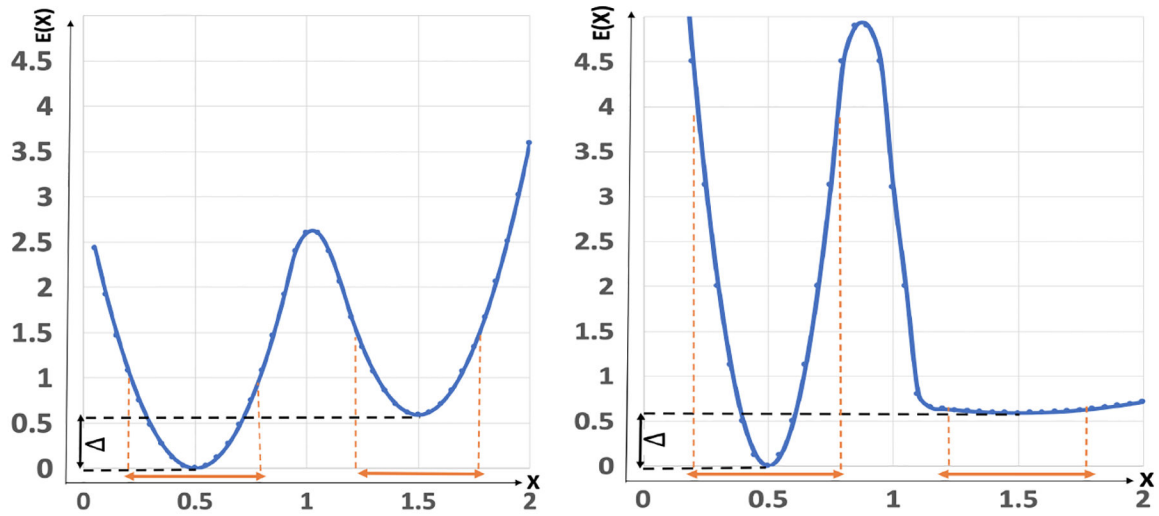


Figure 4: Robustness analysis of two examples.

Left panel shows two equally wide wells, which are similarly robust to small perturbations of the parameters. Right panel shows a totally different behavior of the objective function, where the local minimum is more robust to perturbations.

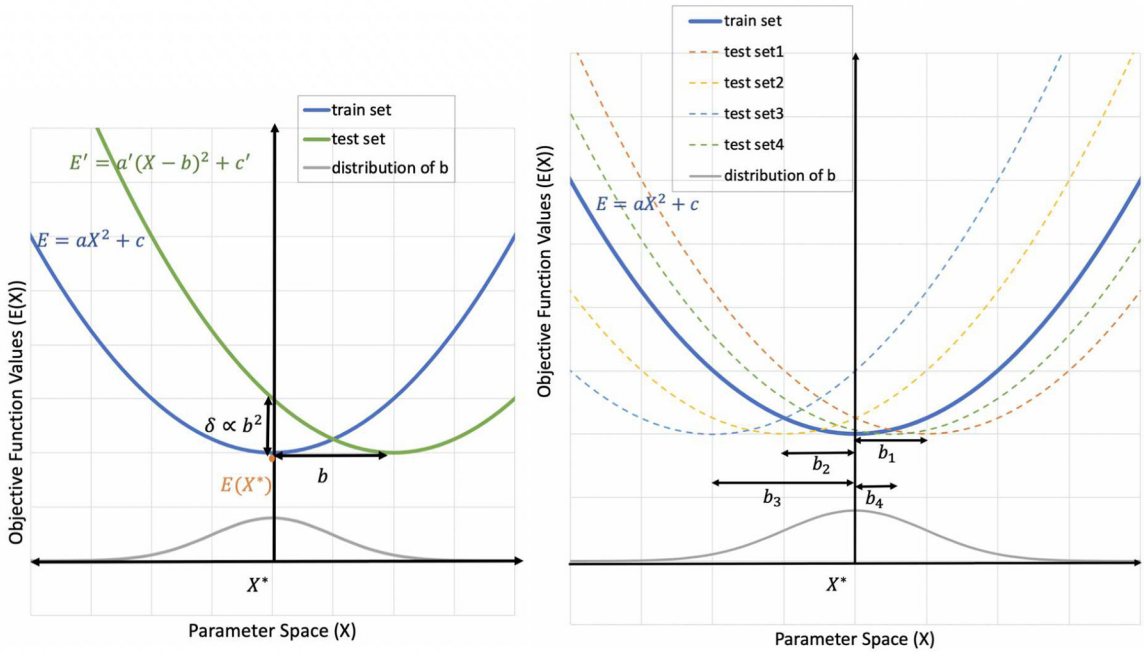


Figure 5: Deviation from the optimal solution (*X) given a new data set.**

Left: changes in objective function value at $X^* = 0$ (δ) is proportional to b^2. Right: estimation of the standard deviation of b when several test sets are given.

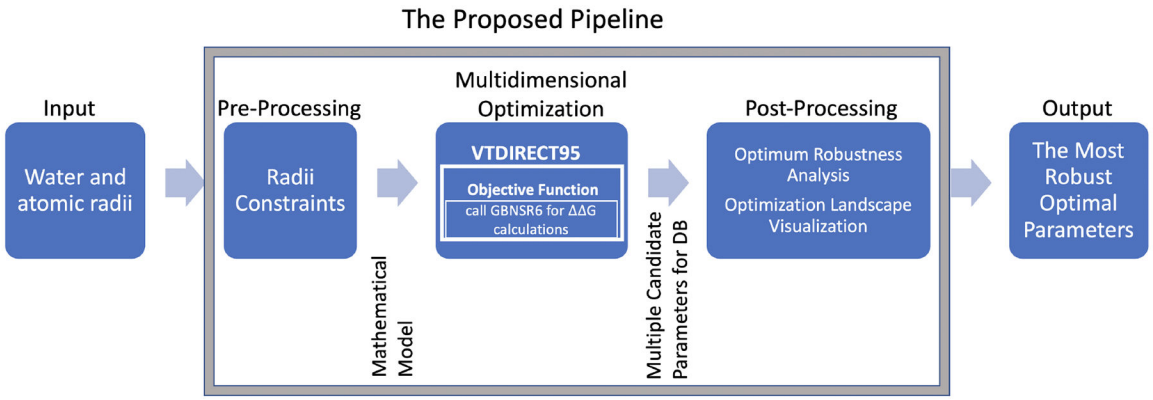


Figure 6: The proposed pipeline flowchart.

Radii constraints, optimum robustness metric and visual analysis are novel and explained in detail in "Results and Discussion". GBNSR6[75] and VTDIRECT95[47] are reviewed briefly in "Materials and Methods".

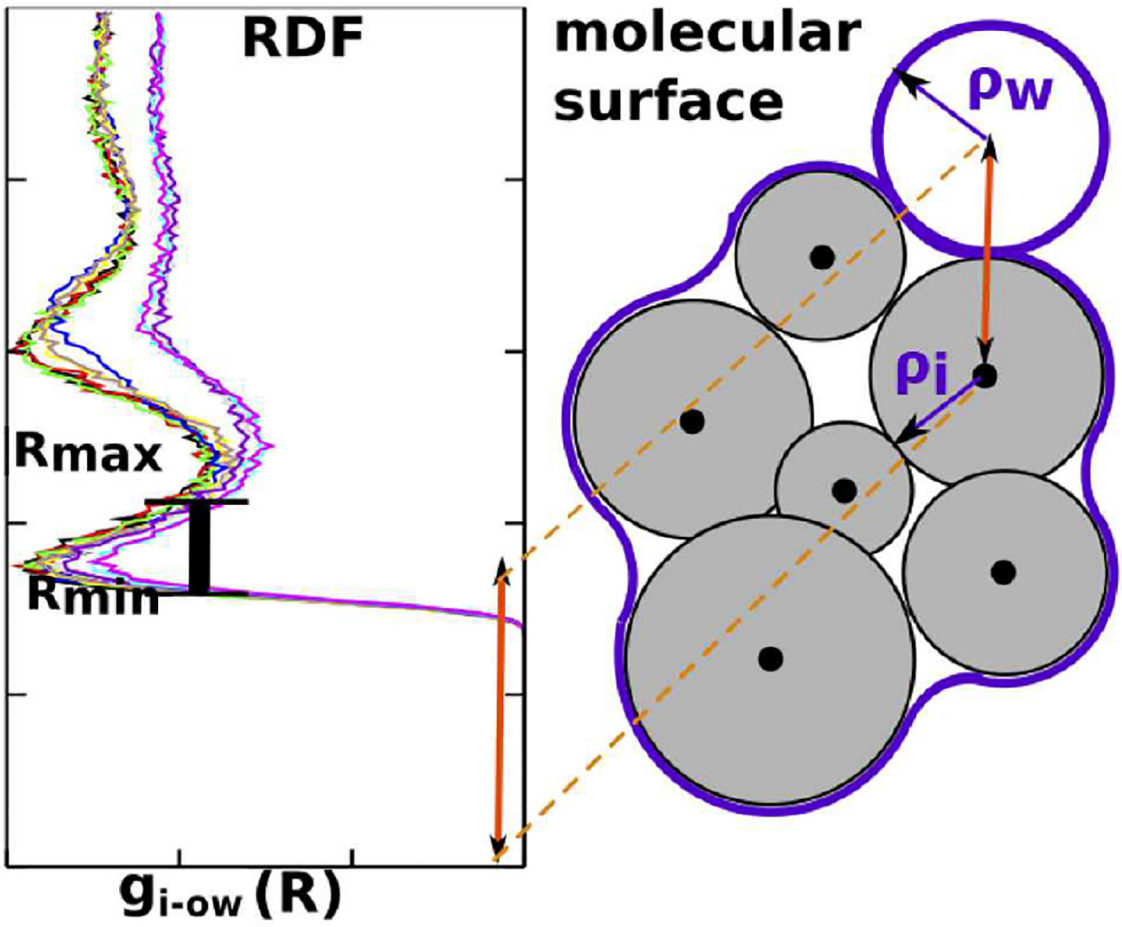


Figure 7: Solvent excluded surface (SES) exemplified for a "molecule" of six atoms.

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_{min},R_{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 ρ_i, atomic radii of atom *i*, and ρ_w, water probe radius.

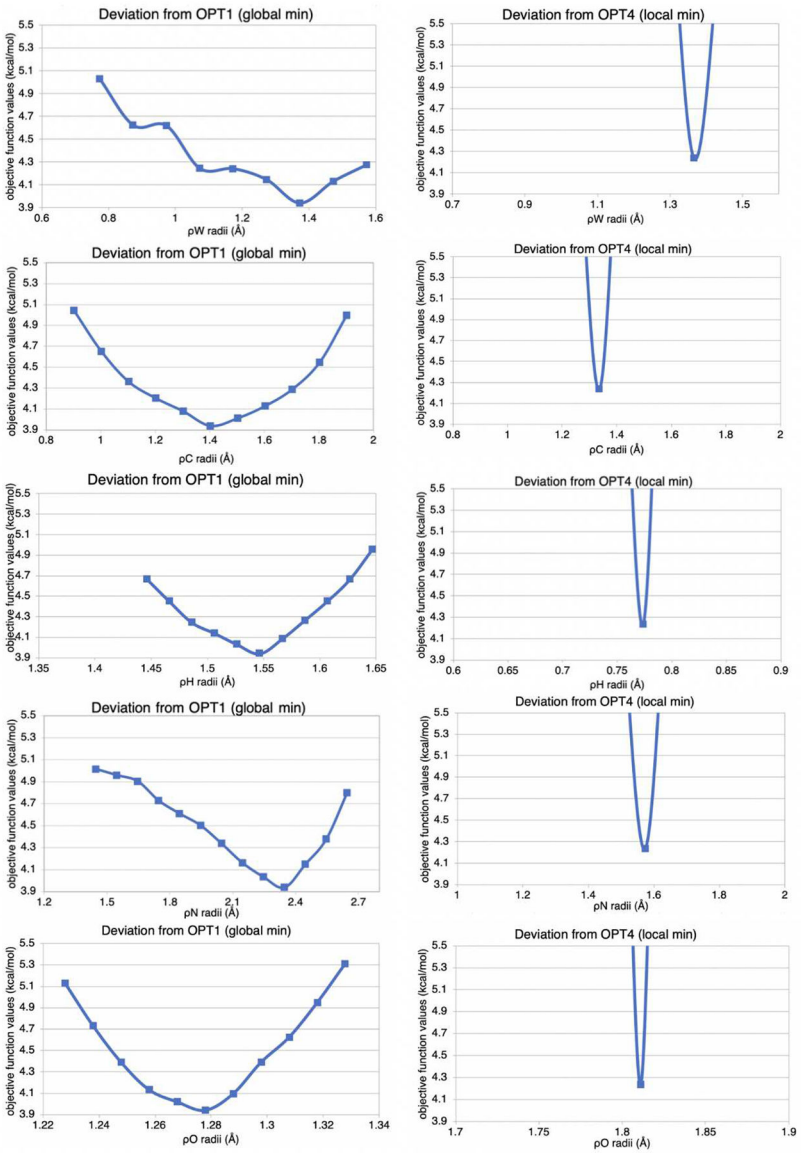


Figure 8: Projection of OPT1 (global min) and OPT4 (local min) onto different radii coordinates.

Left panel shows the behavior of OPT1 objective function projected onto ρ_W, ρ_C, ρ_H, ρ_N and ρ_O within the sampling box and in the physical bound proposed in Eq. (9). 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.

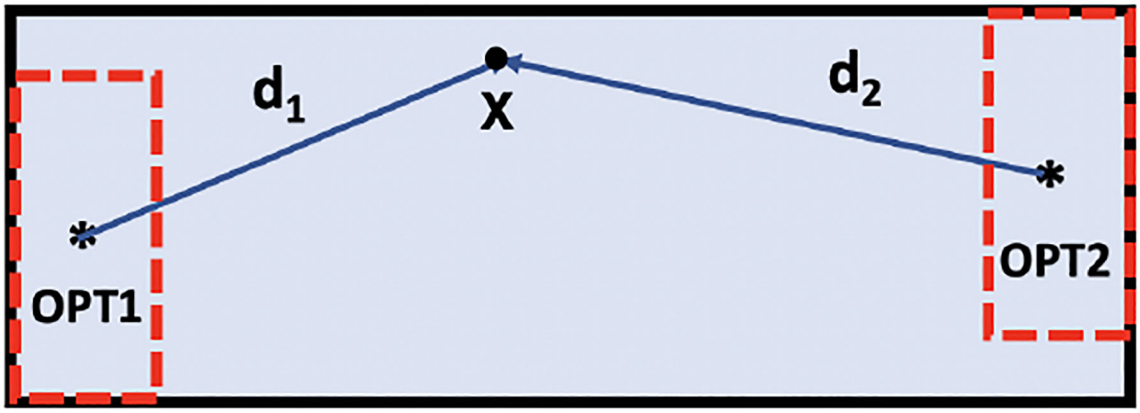


Figure 9: Procedure for creating a distance plot, exemplified.

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. 10.

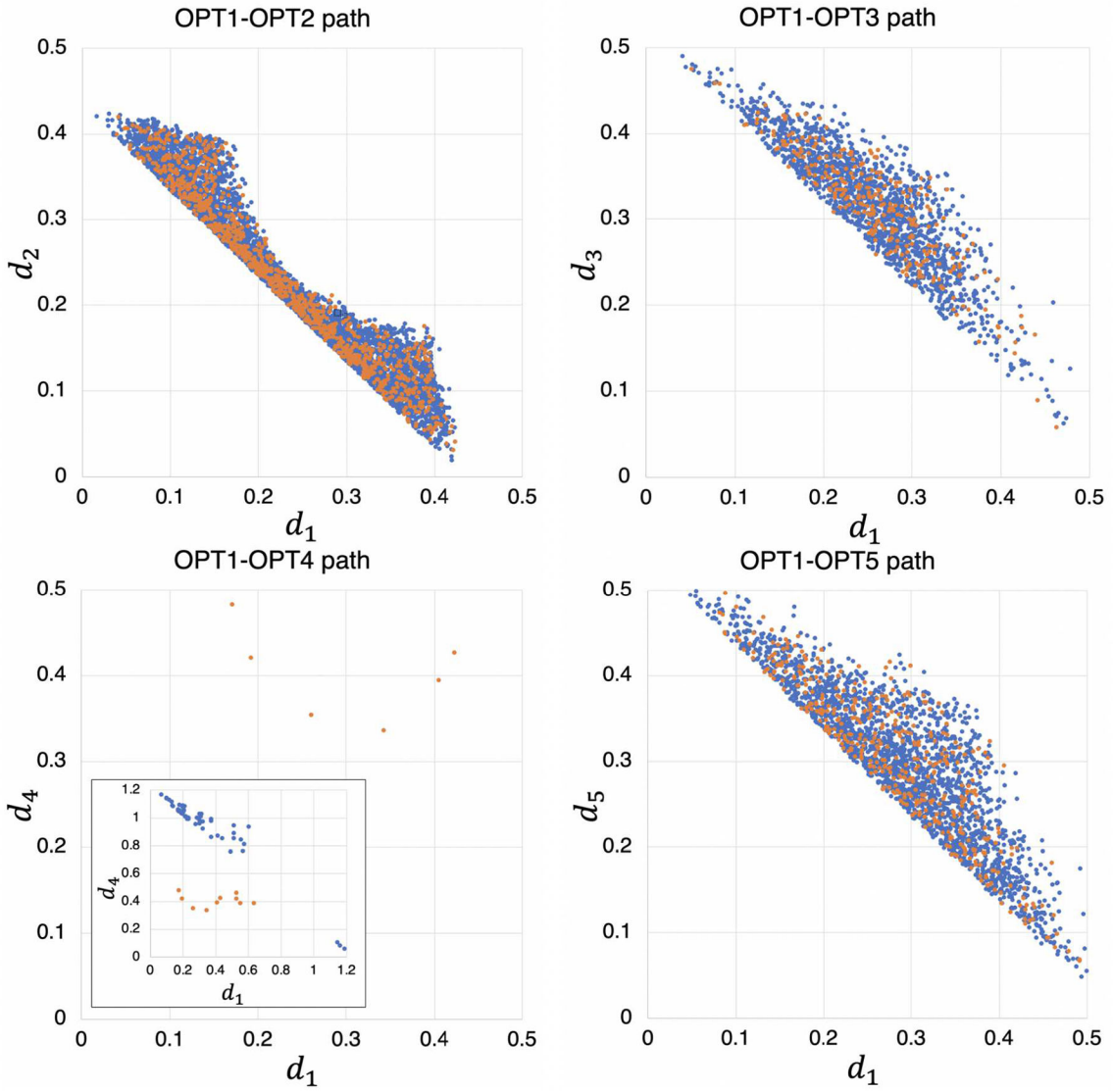


Figure 10: Distance plots.

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.

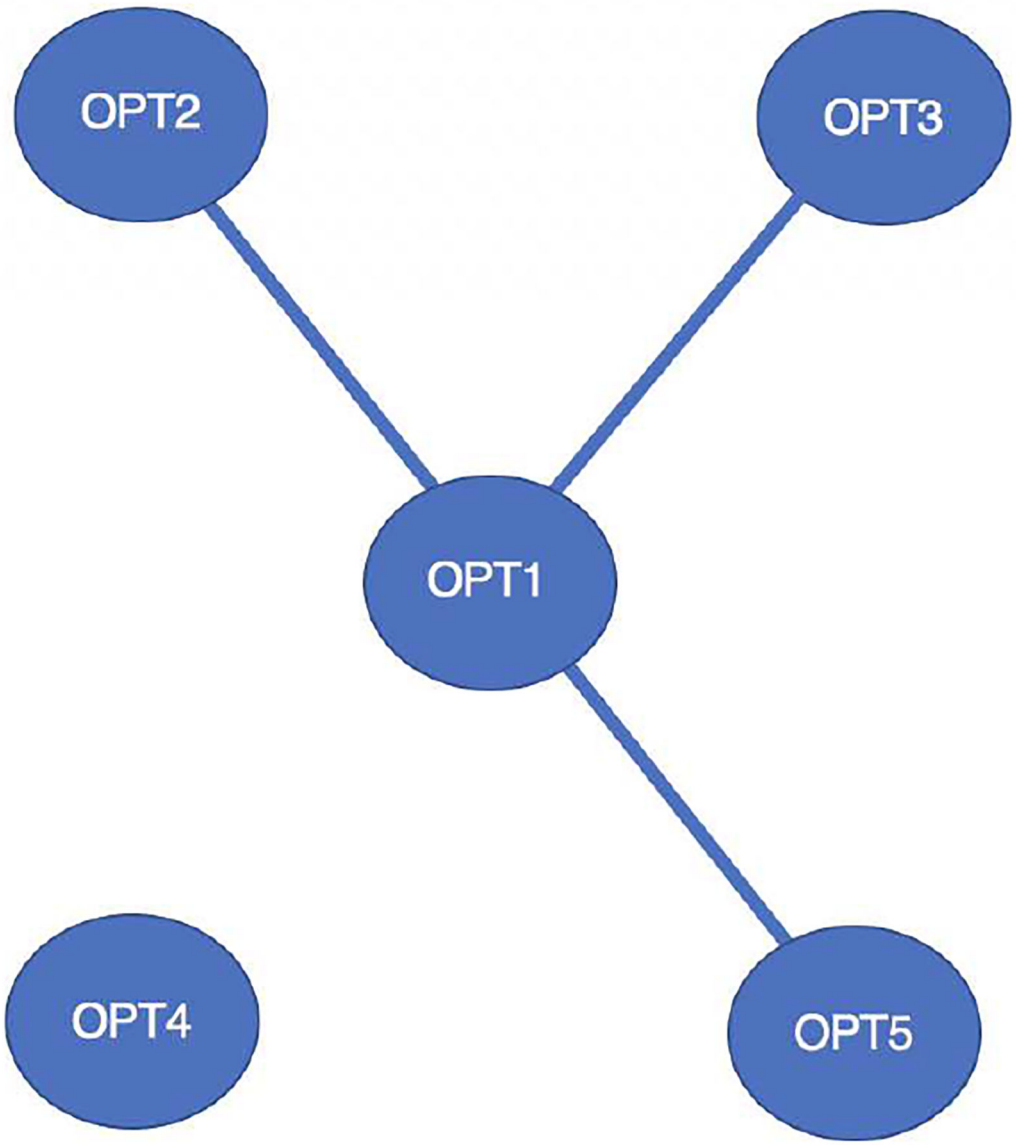


Figure 11: kT-connectivity graph.

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 mnima without exceeding a pre-defined threshold of the objective function, in our case $kT \sim 0.6$ *kcal/mol*.

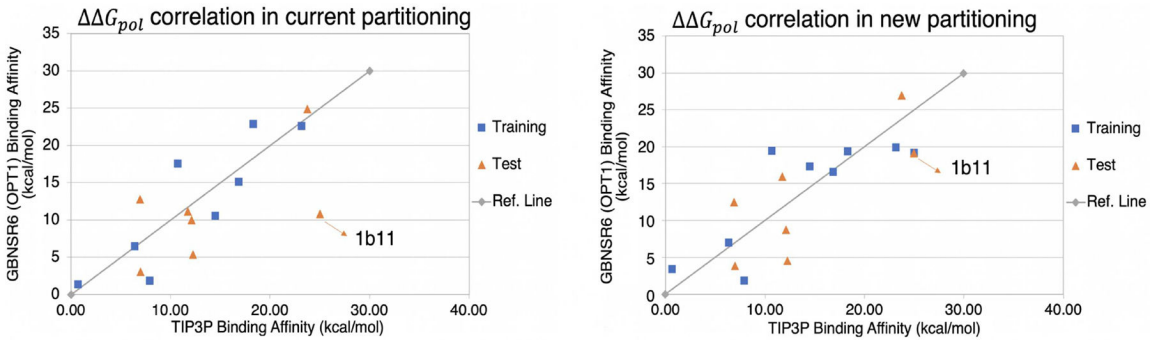


Figure 12: Re-balancing of the training and test sets, with TIP3P explicit solvent model as the reference.

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 found by VTDIRECT95.

Table 1:

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

111_trichloroethane	1234_tetrachlorobenzene
2_bromo_2_methylpropane	diethyl_sulfide
methyl_methanesulfonate	tetrafluoromethane
112_trichloro_122_trifluoroethane	1_methylcyclohexene
4_fluorophenol	iodobenzene
morpholine	

Table 2:

The lowest five optimum parameter vectors found by VTDIRECT95.

	ρ_W	ρ_C	ρ_H	ρ_N	ρ_O	E_{train}
OPT 1	1.37	1.40	1.55	2.35	1.28	3.94
OPT 2	1.52	1.79	1.47	2.27	1.28	4.04
OPT 3	1.06	1.67	1.32	2.14	1.35	4.08
OPT 4	1.37	1.34	0.77	1.57	1.81	4.24
OPT 5	1.06	1.35	1.74	2.71	1.17	4.25

Radii are in $\AA$ and objective function values of the training set, E_{train}, are in *kcal/mol*.

Table 3:

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. (10), all are in *kcal/mol*.

	$\langle E_{train}^{1000} \rangle$	$\langle E_{train}^{5000} \rangle$	$\langle E_{train}^{5000} \rangle'$
OPT 1	4.73	4.71	4.45
OPT 2	4.75	4.75	4.51
OPT 3	5.00	4.97	4.75
OPT 4	5.75	5.78	5.37
OPT 5	4.87	4.90	4.61

Table 4:

The accuracy (RMSE to the explicit solvent reference, Eq. (5)) 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*.

Atomic Radii	ρ_W	ρ_C	ρ_H	ρ_N	ρ_O	E_{train}	E_{test}
OPT1	1.37	1.40	1.55	2.35	1.28	3.94	6.62
PARSE	1.4	1.7	1.0	1.5	1.4	10.80	8.07
ZAP-9	1.4	1.87	1.1	1.55	1.52	5.28	8.27