Statistical Framework for Uncertainty Quantification in Computational Molecular Modeling

Muhibur Rasheed
Computer Science Dept.
University of Texas at Austin
Austin, TX 78712, USA.
muhibur@utexas.edu

Nathan Clement
Computer Science Dept.
University of Texas at Austin
Austin, TX 78712, USA.
nclement@utexas.edu

Abhishek Bhowmick
Computer Science Dept.
University of Texas at Austin
Austin, TX 78712, USA.
ab.abhishek.bhowmick@gmail.com

Chandrajit Bajaj
Computer Science Dept.
University of Texas at Austin
Austin, TX 78712, USA.
bajaj@cs.utexas.edu

ABSTRACT

Computational molecular modeling often involves noisy data including uncertainties in model parameters, computational approximations etc., all of which propagates to uncertainties in all computed quantities of interest (QOI). This is a fundamental problem that is often left ignored or treated without sufficient rigor. In this article, we introduce a statistical framework for modeling such uncertainties and providing certificates of accuracy for several QOI. Our framework treats sources of uncertainty as random variables with known distributions, and provides both a theoretical and an empirical technique for propagating those uncertainties to the QOI, also modeled as a random variable. Moreover, the framework also enables one to model uncertainties in a multi-step pipeline, where the outcome of one step cascades into the next. While there are many sources of uncertainty, in this article we have applied our framework to only positional uncertainties of atoms in high resolution models, and in the form of B-factors and their effect in computed molecular properties. The empirical approach requires sufficiently sampling over the joint space of the random variables. We show that using novel pseudo-random number generation techniques, it is possible to achieve the required coverage using very few samples. We have also developed intuitive visualization models to analyze uncertainties at different stages of molecular modeling. We strongly believe this framework would be immensely valuable in evaluating predicted computational models, and provide statistical guarantees on their accuracy.

BCB '16 October 02–05, 2016, Seattle, WA, USA

DOI: http://dx.doi.org/10.1145/2975167.2975182

Categories and Subject Descriptors

G.1.2 [**Mathematics of Computing**]: Approximation—*Chebyshev approximation and theory*; G.3 [**Mathematics of Computing**]: Probability and Statistics—*Random number generation*; I.6.4 [**Computing Methodologies**]: SIMULATION AND MODELING—*Model Validation and Analysis*; J.3 [**Computer Applications**]: LIFE AND MEDICAL SCIENCES—*Biology and genetics*

General Terms

Algorithms, Experimentation, Measurement, Theory

Keywords

Uncertainty Quantification; Sampling; Molecular Modeling

1. INTRODUCTION

Computational models of any biological substance are, by nature, prone to error. The source of this error can range anywhere from discrete representations of continuous data, to inadequate sampling of the parameter/search space, computational approximations, or even failures to include all relevant aspects of the biological system. In some cases, these errors are slight or insignificant, but when the errors combine—as they frequently do for computations that involve geometry and complicated (linear or non-linear) numerical systems—they can create a result that is unreliable. Modeling of protein structures and their interactions is a field of research that is especially susceptible to cascading errors, for various computed quantities of interest (QOIs). These computations include multi-step methods for protein sequence alignment and homology modeling, implicit solvation interfaces (i.e. molecular surfaces) generation [16, 6, 20, 35, 9, 8], configuration-dependent binding affinity calculations [15, 7], molecular docking and structure refinement via molecular substructure replacement and fitting [39, 5, 22, 36], etc. For each of these computations, the confidence in the reported results could necessarily be bolstered if each estimation of a QOI in the computational pipeline also included rigorous evaluation of their uncertainty. Such uncertainty bounds, along with quantitative visualization, would

be invaluable in rational design and analysis in molecular modeling tasks.

Unfortunately, the majority of current computational structure modeling and prediction protocols do not rigorously consider the effect of such uncertainties and/or report the confidence on the final model, quantity or prediction they compute. For instance, structure prediction protocols including fitting, docking, homology modeling, etc., addresses uncertainties in an indirect way by reporting several structures ranked under some metric with the hope that at least one of the predicted structural models is close to the truth. However, it is not clear how to ascertain the quality or confidence on individual models in the ranked list. Furthermore, there is no statistical guarantee that a near-accurate structure model is present in the entire ranked list. Similarly, protocols for computing specific properties of molecules like surface area, binding free energy, solvation, etc., sometimes provide theoretical guarantees on the computational approximation errors due to numerical approximations, discretization, etc., but do not address the inherent uncertainty of the input itself. While some work does attempt to bound the uncertainty on individual input models (see, for example, [32] for X-ray crystallography or [40, 24] for NMR structure prediction using probabilitistic analysis), determining how this uncertainty propagates to future stages in a pipeline is left unaddressed. An exception is recent work [31] that addresses the influence of conformational uncertainty on biomolecular solvation under elastic network dynamics on an input structural model, where the individual residue positions are independent and identically distributed Gaussian random variables. In this paper, we present a mathematical and an empirical framework, both of which take into account uncertainties present in the input to any computational step and provides an upper bound on the uncertainty of the outcome.

We define statistical uncertainty quantification as a tail bound $Pr[|f - E[f]| > t] < \epsilon$. In other words, a probabilistic certificate as a function of a parameter t that the computed value, $f(\mathbf{X})$, of a QOI, expressed as some complicated function or optimization functional involving noisy data $\mathbf{X}$, is not more than t away from the true value (with high probability). In this article, treating each component of $\mathbf{X}$ as random variables (RVs), we adopt a method of bounded differences, which is a modification of Markov and Chebyshev inequalities, used for independent RVs to derive Chernoff-like bounds of the form mentioned above. However, for most biophysical properties the uncertainties in the input need not necessarily be independent. We show that for such cases, a variation of the Chernoff-Hoeffding bound, namely the Azuma inequality [1], may be applied. Here a stochastic processes is formulated as a Doob martingale and the Azuma inequality applied to such a martingale becomes what is known as the McDiarmid's inequality [33]. In the methods section, we describe this framework in greater detail and also show how it can be readily applied to compute such certificates for functions with decaying kernels. Examples of such functions include the van der Waals interaction energy, atom-atom contact potentials based on distance cutoffs, integrals over point neighborhoods, etc.

Theoretical tail bounds under the McDiarmid model often tend to be too conservative (capturing the worst possible case). Also for some functions, deriving the bound analytically may be challenging. So, we have also developed an alternative technique where Quasi Monte Carlo sampling of the uncertainty space leads to an approximation of the distribution of the values for the QOI, whose expectation can be used to estimate the tail bounds. Our empirical analysis over a diverse set of proteins (from the Zlab benchmark [26]) showed that a fairly small number of samples often suffices to generate a robust approximation of the distribution of the QOI.

While there are many possible sources of uncertainty and many possible QOI in the realm of computational molecular modeling, and our framework could technically be applied to model all of these at the same time, we have applied it in a slightly limited scope in the experiments we report in this article. We have chosen to understand the effect of small positional uncertainties of atoms in high resolution crystal structures on a few important QOIs. The QOIs we considered are surface area (SA), volume, internal van der Waals energy or Lennard-Jones potential (LJ), coulombic energy (CE), and solvation energy under both generalized Born (GBSA) and Possion-Boltman (PBSA) models for single molecules; as well as interface area, and binding free energy calculation for pairs of bound molecules (please see [38] for details on the energy functions). We used B-factors reported in PDB files as an implicit description of positional uncertainty of an atom.

It should be noted that with the use of B-factors, we make two simplifying assumptions. First, while there may be many equally-good (and even possibly better) assignments of positional uncertainty from electron scattering data, we assume that those reported in the PDB are "correct." (For methods that improve the B-factor estimation see, among others, [12, 44, 30, 28, 29]) Second, due to the way atomic positions and uncertainties are resolved from raw data, there may be correlations between reported values; we treat each coordinate as independent random variables. Both of these assumptions are allowable as B-factors are a well-known, readily available statistic, and are included with all molecules in the PDB. Furthermore, the same methodology could (and should) be applied to additional models of uncertainty, and we believe similar results would be attained.

Our empirical study on 57 x-ray structures of bound protein complexes showed that when these uncertainties are accounted for, there is relatively low uncertanty for simple quantities such as exposed surface area (>2% error on just over 5% of the sampled configurations), but a significant probability (>10%) of having more than 5% error in total energy (PBSA) calculation. These examples not only show the uncertainty in commonly-reported statistics, but also illustrate the *propagation* of uncertainty from one simple QOI to another more complex one. This illustrates the great need for proper quantification and reporting of uncertainty in QOI.

Additionally, our QMC based protocol can also be used to generate data for quantitative visualization of uncertainties at different levels of granularity. In this article we have provided several examples.

We developed tools which implement the mathematical framework of sampling needed for statistical UQ, use existing tools to compute the QOIs [13, 15, 3, 21, 45], and compute uncertainty bounds as well the visualization directives which can be directly loaded into existing molecular, surface/volume visualization software [4, 42]. Our methods enable the end user of these tools to achieve a more quanti-

tative and visual evaluation of various molecular models for structural and property correctness—or the lack thereof.

2. METHODS

We define statistical uncertainty quantification on QOI's as a tail bound, namely, a probabilistic certificate as a function of a parameter t that the computed value $f(\mathbf{X})$ of a QOI, expressed as some complicated function or optimization functional involving noisy data $\mathbf{X}$, is not more that t away from the true value (with high probability). Or alternatively the probability of the error being greater than or equal to t is very small. Such a certificate is expressed as a Chernoff-Hoeffding [14] like bound as follows:

$$Prob(f, X, t, \epsilon) = \Pr[|f(X) - E[f]| > t] \leq \epsilon. \quad (1)$$

In this article, we adopt a method of bounded differences, which is a modification of Markov and Chebyshev inequalities, used for independent RVs to derive Chernoff-like bounds. Here, we briefly introduce a loose uncertainty bound based on Doob martingales as introduced by Azuma [1] and Hoeffding [25] and later extended by McDiarmid [33]. The McDiarmid inequality is stated as follows:

DEFINITION 2.1 (MCDIARMID BOUND). *Let (X_i) be independent RVs with discrete space A_i. Let $f : \prod_i A_i \rightarrow \mathbb{R}$, and $|f(x_1, \ldots x_k, \ldots, x_n) - f(x_1, \ldots, x'_k, \ldots, x_n)| \leq c_k$. Then, for $t > 0$, $\Pr[|f(\mathbf{X}) - \mathbf{E}[f]| > t] < 2\exp\left(-2t^2/\sum_k c_k^2\right)$.*

In the next few subsections, we discuss the derivation of the bound and extend it to include cases where the RVs are not necessarily independent. We also show that for molecular modeling tasks that involve summations over decaying kernels (e.g. electrostatic interactions), one can analytically derive such bounds. However, such theoretical upper bounds often overestimate the error. So, we also explore an alternate quasi Monte Carlo (QMC) approach [34, 27]. The QMC is useful in approximating the distribution of values for the QOI under a given statistical model of the input uncertainty, and then empirically establishing the uncertainty of individual values of the QOI, as well as providing certificates like Equation 1. Additionally, this can be applied to complex functionals (like Poisson-Boltzman (MM-PBSA) energy calculations) which are not amenable to analytical treatment. We should mention that the quality of the QMC-based certificate itself depends on the quality and size of the samples, but our experiments (see Results and Discussion) showed that typically fewer than 500 samples under a low discrepancy sampling technique provides sufficiently accurate approximations of the certificates.

2.1 Thoretical framework for statistical uncertainty quantification

To prove theoretical uncertainty bounds, one often uses Chernoff-Hoeffding style bounds. However, this is useful only if the the underlying random variables are independent and we are analyzing the sum of the random variables. In practical situations the random variables have dependencies. In such cases, we can still prove large deviation bounds using the theory of martingales, specifically Doob martingales and their extension.

2.1.1 Martingales and McDiarmid inequality

DEFINITION 2.2. *Let $(Z_i)_{i=1}^n$ and $(X_i)_{i=1}^n$ be a sequence of random variables on a space Ω. Suppose $\mathbf{E}[X_i|Z_1, \ldots, Z_{i-1}] = X_{i-1}$. Then (X_i) forms a martingale with respect to (Z_i).*

Essentially, the expected value of the i^{th} observable is the same as the observed value of the $(i-1)^{th}$, irrespective of the values of all the the others. A variation, the Doob martingale, can be constructed from any random variable in the following universal way.

CLAIM 2.3. *Let A and (Z_i) be random variables on space Ω. Then, $X_i = \mathbf{E}[A|Z_1, \ldots, Z_{i-1}]$ is a martingale with respect to Z_i. This is called the Doob martingale of A with respect to (Z_i).*

Now we present the Azuma inequality for martingales.

CLAIM 2.4 (AZUMA INEQUALITY). *Let (X_i) be martingale with respect to (Z_i). Suppose $|X_i - X_{i-1}| \leq c_i$. Then*

$$\Pr[|X_n - X_0| > t] \leq 2\exp(-t^2/2\sum_i c_i^2).$$

The weak form of the McDiarmid inequality follows directly from the Azuma inequality. Please see [38] for detailed proofs and derivations.

2.2 Analytical uncertainty bounds for biophysical quantities

Consider the summations over decaying kernels of the form shown below, when the variables are uncertain.

$$F(A, B) = \sum_{\mathbf{x_1} \in A} \sum_{\mathbf{x_2} \in B} \sum_{k=1}^{n} \frac{a_k}{\|\mathbf{x_1} - \mathbf{x_2}\|^{b_k}} \quad (2)$$

where b_k are non-negative constants, a_k are constants, and A and B are two sets of points.

This is a form that is seen in van der Waals energy calculations, computing contact properties (e.g. number of atom contacts at a binding interface, binding interface area etc.), and many similar biophysical quantity of interest. In such applications the uncertain quantities will be the positions of the atoms.

In the following, we introduce some notations and then analytically express the uncertainties of successively more complex functions.

2.2.1 Notation

A single decaying kernel in the above summation is represented as

$$f_{\mathbf{x_1}}(\mathbf{x_2}) = \sum_{k=1}^{n} \frac{a_k}{\|\mathbf{x_1} - \mathbf{x_2}\|^{b_k}} \quad (3)$$

where the kernel is centered at $\mathbf{x_1}$ and evaluated at $\mathbf{x_2}$. The following result is immediate:

LEMMA 2.5. *For a given set of a_k and b_k, $f_{\mathbf{x_1}}(\mathbf{x_2}) = f_{\mathbf{0}}(\mathbf{\Delta x})$ where $\mathbf{\Delta x} = (\mathbf{x_2} - \mathbf{x_1})$.*

When both $\mathbf{x_1}$ and $\mathbf{x_2}$ are uncertain such that every component x_{1i} of $\mathbf{x_1}$ is uniformly sampled from the interval $[l_{1i}, u_{1i}]$, and every component x_{2i} of $\mathbf{x_2}$ is uniformly sampled from the interval $[l_{2i}, u_{2i}]$- we can assume that every

component Δx_i of $\mathbf{\Delta x}$ is uniformly sampled from the interval $[l_i, u_i]$ computed based on $[l_{2i}, u_{2i}]$ and $[l_{1i}, u_{1i}]$. The error of $f_{\mathbf{x_1}}(\mathbf{x_2})$ due to the uncertainty of $\mathbf{x_1}$ and $\mathbf{x_2}$ can hence be equivalently computed as the error of $f_{\mathbf{0}}(\mathbf{\Delta x})$ due to the uncertainty of $\mathbf{\Delta x}$. In our discussion, we shall often drop the Δ when the context does not require the distinction.

2.2.2 Uncertainty of a single kernel at a single point

We begin with the simplest case when the kernel is embedded in 2D (the 1D case is trivial):

$$f_1(x, y) = \frac{a}{(x^2 + y^2)^{b/2}} \tag{4}$$

Assuming that x and y are uniformly sampled from the intervals $[l_x, u_x]$ and $[l_y, u_y]$ respectively where l_x, l_y, u_x and u_y are non-negative, we can define the maximum deviation due to the change of x as

$D_{1x} = \max_y |f_1(l_x, y) - f_1(u_x, y)|$

Note that $g_1(y) = f_1(l_x, y) - f_1(u_x, y)$ is positive for $l_x < u_x$, and $\frac{d}{dy} g(y) < 0$. Hence, $g_1(y)$ is maximized when $y = l_y$. So,

$$\begin{aligned} D_{1x} &= \max_y |f_1(l_x, l_y) - f_1(u_x, l_y)| \\ &= |a| \left(\frac{1}{\left(l_x^2 + l_y^2\right)^{b/2}} - \frac{1}{\left(u_x^2 + l_y^2\right)^{b/2}} \right) \end{aligned} \tag{5}$$

D_{1y} can also be computed the same way. Using McDiarmid's theory of bounded differences, we have the following result-

LEMMA 2.6. *For the decaying kernel f_1 in Equation 4, $\Pr[|f_1 - E[f_1]| > t] \leq 2e^{\frac{-2t^2}{D_{1x}^2 + D_{1y}^2}}$ where D_{1x} and D_{1y} are defined in Equation 5.*

The above results can be readily extended to d dimensions for the function f_2 defined below.

$$f_2(\mathbf{x}) = \frac{a_k}{\|\mathbf{x}\|^{b_k}} \tag{6}$$

Let, $f_{2i}(\mathbf{x}, y)$ represent $f_2(\mathbf{x})$ such that the value of the i^{th} component is fixed to y. So we define the maximum deviation of f_2 due to the change of one variable x_i between the range $[l_i, u_i]$ as:

$$D_{2i} = \max_{\mathbf{x}} g_{2i}(\mathbf{x}) = \max_{\mathbf{x}} |f_{2i}(\mathbf{x}, l_i) - f_{2i}(\mathbf{x}, u_i)| \tag{7}$$

Again $g_{2i}(\mathbf{x})$ is positive and $\frac{d}{dx_j} g_2(\mathbf{x}) < 0$ for all components x_j of $\mathbf{x}$. Hence, $g_2(\mathbf{x})$ is maximized when $x_j = l_j$ for all j where l_j is the lowest possible value for x_j.

$$D_{2i} = |a| \left(\frac{1}{\left(\sum_k l_k^2\right)^{b/2}} - \frac{1}{\left(u_i^2 + \sum_{k \neq i} l_k^2\right)^{b/2}} \right) \tag{8}$$

LEMMA 2.7. *For the decaying kernel f_2 defined in Equation 6, $\Pr[|f_2 - E[f_2]| > t] \leq 2e^{\frac{-2t^2}{\sum_i D_{2i}^2}}$ such that D_{2i} is defined as in Equation 8.*

Note that Lemmas 2.6 and 2.7 hold even when $a < 0$ (i.e. negative).

2.2.3 Uncertainty of multiple kernels at a single point

Now we extend the scope to consider functions which are expressed as a sum of n decaying kernels centered at the origin.

$$f_3(\mathbf{x}) = \sum_{k=1}^{n} \frac{a_k}{\|\mathbf{x}\|^{b_k}} \tag{9}$$

Let ${f_3}^k(\mathbf{x}) = \frac{a_k}{\|\mathbf{x}\|^{b_k}}$ denote the k_th decaying term in Equation 9. Now, the maximum deviation will be defined similar to Equation 7.

$$\begin{aligned} D_{3i}(\mathbf{x}) &= \max_{\mathbf{x}} g3_i(\mathbf{x}) \\ &= \max_{\mathbf{x}} |f_{3i}(\mathbf{x}, l_i) - f_{3i}(\mathbf{x}, u_i)| \\ &= \max_{\mathbf{x}} \left| \sum_k \left(f_{3_i}^k(\mathbf{x}, l_i) - f_{3_i}^k(\mathbf{x}, u_i) \right) \right| \\ &\leq \max_{\mathbf{x}} \sum_k \left| \left(f_{3_i}^k(\mathbf{x}, l_i) - f_{3_i}^k(\mathbf{x}, u_i) \right) \right| \\ &\leq n \max_k \max_{\mathbf{x}} \left| \left(f_{3_i}^k(\mathbf{x}, l_i) - f_{3_i}^k(\mathbf{x}, u_i) \right) \right| \\ &= n \max_k D_{2_i}^k \end{aligned} \tag{10}$$

where $D_{2_i}^k$ is defined the same way as D_{2i} in Equation 8 for the k^th kernel.

LEMMA 2.8. *For the sum of decaying kernel f_3 given in Equation 9,*

$$\Pr[|f_3(\mathbf{x}) - E[f_3(\mathbf{x})]| > t] \leq 2e^{\frac{-2t^2}{\sum_i D_{3_i}^2(\mathbf{x})}}$$

such that $D_{3i}(\mathbf{x})$ is defined as in Equation 10.

2.2.4 Uncertainty of a multiple kernels at multiple points

Let us define a volumetric function in d dimensions as a sum over multiple kernels defined at multiple points belonging to the set A as follows.

$$f_4(A, \mathbf{y}) = \sum_{\mathbf{x} \in A} \sum_{k=1}^{n} \frac{a_k}{\|\mathbf{x} - \mathbf{y}\|^{b_k}} \tag{11}$$

Now, f_4 can be expressed as :

$$\begin{aligned} f_4(A, \mathbf{y}) &= \sum_{\mathbf{x} \in A} f_{3_{\mathbf{x}}}(\mathbf{y}) \\ &= \sum_{\mathbf{x} \in A} f_{3_{\mathbf{0}}}(\mathbf{y} - \mathbf{x}) \\ &= \sum_{\mathbf{x} \in A} f_{3_{\mathbf{0}}}(\Delta x) \end{aligned} \tag{12}$$

Since, f_4 is a simple summation over independent points, the result in Lemma 2.9 follows immediately from Lemma 2.8.

LEMMA 2.9. *For the sum of decaying kernel $f_4(A, \mathbf{y})$ given in Equation 11,*

$$\Pr[|f_4(A, \mathbf{y}) - E[f_4(A, \mathbf{y})]| > t] \leq 2e^{\frac{-2t^2}{\sum_{\mathbf{x} \in A} \sum_i D_{3_i}^2(\mathbf{\Delta x})}}$$

such that $D_{3i}(\mathbf{\Delta x})$ is defined as in Equation 10.

2.2.5 Uncertainty of a integral over multiple kernels at multiple points

Finally, we bound the uncertainties in the integral function we mentioned at the beginning of this section in Equation 2.

LEMMA 2.10. *For the sum of decaying kernel $F(A, B)$ given in Equation 11,*

$$\Pr[|F(A,B) - E[F(A,B)]| > t] \leq 2e^{\frac{-2t^2}{\sum_{\mathbf{x_1}\in A}\sum_{\mathbf{x_2}\in A}\sum_i D_{3i}^2(\mathbf{\Delta x})}}$$

such that $D_{3i}(\mathbf{\Delta x})$ is defined as in Equation 10 and $\mathbf{\Delta x} = (\mathbf{x_2} - \mathbf{x_1})$.

2.3 Empirical uncertainty quantification using quasi-Monte Carlo methods

Under the quasi-Monte Carlo method of uncertainty propagation, it is assumed that x_i are independent random variables and their probability distribution functions (PDFs) are known. Now, a low discrepancy sampling of the product space $x_1 \times x_2 \times \ldots \times x_n$ must be generated which would define an approximate PDF for $f(\mathbf{X})$ from which we could compute all the necessary tail bounds. We explore several such product spaces and use low discrepancy sampling strategies to derive PDFs of functions $f(\mathbf{X})$ with bounded error and guaranteed convergence. Note that a simpler Quasi-Monte Carlo (QMC) sampling can be applied to find the minimum c_k for each x_k, and hence derive the loose McDiarmid bound. Hence the most crucial component of UQ under this QMC framework boils down to identification of the set of independent random variables $\mathbf{x}$, which affect the computation of the QOI, along with their approximate PDFs and corresponding sampling techniques.

2.3.1 Parameterizations and uncertainties of molecular models

Molecular structural models are typically parameterized using either a list of their XYZ coordinates, or using internal coordinates (which is a series of bond length, bond angle and dihedral angle). In the first representation the degrees of freedom or the space of configurational uncertainty is related to each coordinate value; in the latter representation, typically the dihedral angles are the only degrees of freedom since bond lengths and angles are considered constants.

X-ray crystallography experiments reconstruct a 3D electron density cloud from the diffraction pattern generated from a crystal lattice of the molecule. For high resolution reconstructed electron densities, expected locations of individual atoms can be identified. Hence, it is common to report such models using the XYZ coordinates of the atoms. However, the expected location is not necessarily perfectly determined. There is a degree of uncertainty which is often expressed as temperature factors or B-factors. Simply stated, B-factors are a measure of the error in the match and fit of specific atoms within the electron density cloud constrained by the protein's primary, secondary and tertiary structure and inter-atom biochemical/biophysical forces.

B-factors are derived from structure factors, which are based on the Fourier transform of the average density of the scattering matter. The structure factor, $F(\vec{h})$, for a given reflection vector, $\vec{h}$, is the sum of the optimized parameters for each atom type j, and position $\vec{x_j}$ and as defined by the following equation:

$$F(\vec{h}) = \sum_j f_j \exp\left(-\frac{1}{4}B_j \vec{h}^{\,t}\vec{h}\right)\exp\left(2\pi i \vec{h}^{\,t}\vec{x_j}\right),$$

where f_j is the scattering factor, B_j is the B-factor for atom j, and $\vec{x_j}$ is the 3-dimensional position of each atom [41].

If we assume that the static atomic electron densities have spherical symmetry (or, more specifically defined by a trivariate Gaussian, $\vec{u}$), this can be converted into the anisotropic temperature factor commonly used, $T(\vec{h})$ [46]:

$$T(\vec{h}) = \exp\left[-2\pi^2\langle(\vec{h}\cdot\vec{u})^2\rangle\right],$$

where the univariate Gaussian form (needing not the direction of $\vec{h}$, but only its magnitude) is described by:

$$T(|\vec{h}|) = \exp\left[-8\pi^2\langle u^2\rangle(\sin^2\theta)/\lambda^2\right]. \quad (13)$$

Finally, the B-factor is defined as $B = 8\pi^2\langle u^2\rangle$. Thus, a B-factor of 20, 80, or 180Å^2 corresponds to a mean positional displacement error of 0.5, 1, and 1.5Å, respectively. (Other metrics, such as R-factor [11] or diffraction-component precision index (DPI) [18] can be used to provide more insight into these uncertainties. However, throughout this paper we will just use the B-factors commonly available in the PDB file.)

2.3.2 Defining and sampling the space of configurations

The statistical framework described above, requires geometric models of the molecules, parameterization of the degrees of freedom available to the molecule, a mapping which allows one to update the model based on sampled degrees of freedom, and an implementation of the QOI. In this section, we describe how we derive the relevant degrees of freedom and set up the joint probability distribution which will be sampled by the QMC protocol.

Given the x-ray structure M containing n atoms of a protein or a complex of two proteins in the PDB file format, we extract the anisotropic B-factors B_i^x, B_i^y and B_i^z for each atom $a_i \in M$. The distribution of the position of the atom in each direction is modeled as a Gaussian distribution whose PDF is defined as $p(x_i) = \frac{1}{\sigma_i^x\sqrt{2\pi}}\exp\left[-\frac{(x_i-\mu_i^x)^2}{2\sigma^2}\right]$ where σ_i^x is the standard deviation derived as $\sqrt{\frac{B_i^x}{8\pi^2}}$ from the B-factor, and the mean μ_i^x is the expected position of the atom. Note that for some x-ray structures only an isotropic B-factor B_i is reported. In that case we simply assume $B_i^x = B_i^y = B_i^z = B_i$.

2.3.3 Sampling

The joint distribution, defined as the product space of the $3n$ independent Gaussian distributions, represents the space of possible configurations for the molecule. Hence, to sample the joint distribution, we first sample the space $[0,1]^{3n}$ using pseudorandom generator which guarantees low discrepancy sampling in high dimensional product spaces [23] to produce a tuple $\langle u_1^x, \ldots, u_n^z\rangle$. Each number u_i^j from this tuple is mapped to get a sample a_i^j from a Normal distribution using the Box-Muller method [10], and finally appropriate translation and scaling is performed to get a sample from the corresponding Gaussian distribution as follows $s_i^j = \mu_i^j + \sigma_i^j a_i^j$.

These samples $\langle s_1^x, \ldots, s_n^z \rangle$ are used to displace the atoms to generate a new configuration.

An important point to note is that the above procedure, while maintaining the constraints implied by the B-factors and the mean positions and only making small perturbations, does not necessarily guarantee that the new configurations will be biophysically feasible. In particular they do not preserve the bond angle and bond lengths. So, we use the Amber force field [17] and energy minimization with implicit solvents to relax the sampled structures. We rejected any sample for which the relaxed structure still had high van der Waals potential indicating that some atoms are too close to each other. Accepted models were protonated and assigned partial charges based on the Amber force field using the PDB2PQR program [19].

One of the major drawbacks with QMC sampling is that the curse of dimensionality prevents naive sampling with a large number of dimensions. Assume that if one wants ϵ discrepancy for each random variable, m samples in a single dimension are required (assuming d r.v.'s from the $3n$ independent Gaussians). Then to maintain the same level of discrepancy, the QMC protocol would require m^d samples. For any reasonable measure of discrepancy (i.e. $< 1\%$), this quickly becomes intractable.

For this reason, it is absolutely crucial to use a sampling method that requires a polynomial number of samples. The low-discrepancy product-space sampler developed by Bajaj *et al.* [2] reduces the number of samples significantly from m^d to only $\left(\frac{d}{\epsilon}\right)^{O(\sqrt{\log\left(\frac{1}{\epsilon}\right)})}$, where $m = \left(\frac{d}{\epsilon}\right)^{3+o(1)}$. Note that this is polynomial in m and d.

3. RESULTS AND DISCUSSION

In this section, we detail the results of applying our QMC based UQ framework to generate Chernoff-like bounds (see Methods) for a set of 57 protein complexes. Additionally, we provide a protocol to determine the number of samples required to guarantee the accuracy of the empirical certificates for specific proteins. The results clearly establish the necessity of rigorous quantification of uncertainties, and also shows that such an endeavor need not be prohibitively time consuming.

Finally, we describe some visualization protocols which provide interactive and intuitive representations of the computed uncertainty measures.

3.1 Uncertainty quantified computation of molecular properties

3.1.1 Benchmark and experiment setup

We applied the QMC approach for empirical UQ of computationally evaluated QOIs to 57 crystal structures with 2 bound chains each. We took the 'Rigid-body' cases of antibody-antigen, antibody-bound and enzyme complexes from the Zlab benchmark 4 [26]. We used this docking benchmark as we were interested in demonstrating how uncertainty in QOI reported on a single proteins is magnified when combined with another protein, such as is often done with computing protein binding affinity.

For each of the complexes, we applied the sampling to the receptor and the ligand (the two chains in the structure) separately, and evaluated the uncertainty measures for the calculation of surface area, volume, and components of free energy including Lennard-Jones, Coulombic, dispersion, GB and PB. We also computed the uncertainties in the binding interface area, and change of free energy. In the following subsections we explore different aspects of this analysis.

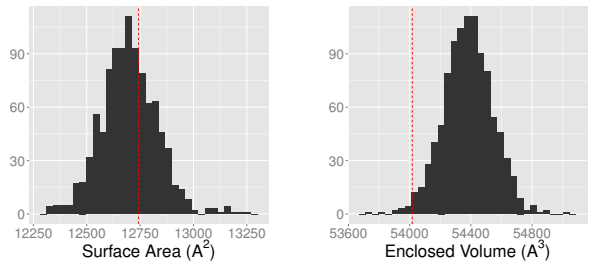

Figure 1: Histogram of sampled QOIs for 1OPH:A. The red vertical line is the value of $\mathcal{F}$ computed using the original coordinates reported in the PDB.

3.1.2 Uncertainty of unperturbed models

Figure 1 shows the distribution of values computed for the sampled models for PDB structure 1OPH-chainA. The red lines in the figures marks the value computed on the original coordinates, and emphasizes the fact that the original coordinates do not always provide the best estimate of the expected value of a QOI. The z-scores for these structures, with respect to the expected values and standard deviations derived from the empirical distribution, are 0.33, -0.82, 1.37, and 0.25 respectively for area, volume, GB, and PB. This emphasizes the importance of applying some form of empirical sampling to find the best representative model (one which minimizes the z-score, for instance).

3.1.3 Certificates for computational models

We also determined the likelihood of producing a large error in the calculation of QOI, due to the presence of uncertainty in the input, in terms of Chernoff-like bounds. For each model in the dataset, we generated the distribution of samples, and then computed the probability, ϵ, of a randomly sampled model having more than 0.1%, 0.5%, 1%, 2%, 5% and 10% error (t), where error is defined as $|x' - E[x]|/E[x]$ such that x' is the value computed for a random model and $E[x]$ is the expected value.

Table 1 lists Chernoff bounds as described above for the two chains of 1OPH, and Table 2 shows corresponding data for the full dataset. The rows named $\Delta area(A)$ represent the quantity $|area(A+B) - area(A) - area(B)|$ computed while keeping B fixed and sampling the distribution of A; rows named $\Delta area(B)$ report the same quantity while keeping A fixed and sampling the distribution of B.

The takeaway from this table is that for most of the QOIs, the probability of incurring more than 5% error is negligible. We also note that the probability of error is higher for Δ QOIs simply because the errors of individual quantities are being propagated and amplified. Uncertainties are also higher in more complex functionals. See Table 1 in [38] for average uncertainties across the entire dataset (instead of just a single protein).

3.2 Number of samples sufficient to provide statistically accurate certificates

Table 1: Chernoff-like bounds for the 1OPH protein. For each value of t, the corresponding values of ϵ are calculated from the 1000 random samples.

t	0.001	0.005	0.010	0.020	0.050	0.100
area(A)	0.911	0.613	0.326	0.050	0.000	0.000
area(B)	0.911	0.582	0.306	0.052	0.000	0.000
Δarea(A)	0.964	0.823	0.650	0.358	0.031	0.000
Δarea(B)	0.973	0.889	0.771	0.560	0.155	0.005
vol(A)	0.727	0.088	0.003	0.000	0.000	0.000
vol(B)	0.765	0.174	0.005	0.000	0.000	0.000
Δvol(A)	0.877	0.485	0.169	0.009	0.000	0.000
Δvol(B)	0.895	0.495	0.176	0.006	0.000	0.000
PB(A)	0.970	0.864	0.727	0.508	0.114	0.014
PB(B)	0.966	0.838	0.672	0.403	0.047	0.004
ΔPB(A)	0.990	0.942	0.863	0.719	0.378	0.106
ΔPB(B)	0.984	0.940	0.870	0.748	0.463	0.164

Table 2: Chernoff bounds for all proteins and both chains, averaged.

t	0.001	0.005	0.010	0.020	0.050	0.100
area	0.910	0.575	0.282	0.059	0.004	0.000
Δarea	0.910	0.578	0.288	0.066	0.009	0.003
vol	0.767	0.191	0.038	0.006	0.001	0.000
Δvol	0.768	0.193	0.038	0.006	0.001	0.000
PB	0.969	0.846	0.708	0.472	0.125	0.023
ΔPB	0.969	0.847	0.710	0.476	0.132	0.030

The results reported in the previous two subsections highlight the importance of UQ and also shows that the mean coordinates do not always correlate well with the QOI computed from the original molecules, so statistical bounds across a set of samples is needed. While the number of samples needed using a pseudo-random number generator is drastically lower than a naive exponential sampling method, it is still prohibitive in practice. (See [38] for the number of sample requirements for different number of atoms.) We believe this theoretical bound overestimates the requirement, and for most practical situations, much fewer samples is sufficient. Now, we seek to determine the number of pseudo-random samples needed to achieve robust certificates, or as the minimal number of samples before the gain achieved through more samples is negligible.

For each QOI, we calculated the mean quantity for a certain number, r, of random samples then calculated ϵ on this reduced dataset. We did the same for s random samples, and computed the L2 distance between the two. Note that, we computed 6 different ϵ (the probabilistic guarantee) for each r, using 6 different values of t (the errors). The expected distance between any two random points in 6d space is 0.9689 (analytic form for such distances has been derived in [37], and the precomputed values for several dimensions are available online [43]. Hence, we chose τ at 0.05, which is much lower that 0.9689, as our measure for convergence.

We used both $s = 1000$ for correlation with the full dataset and $s = r+10$ for an incremental comparison. If the distance between the these was less than a given threshold, τ, then we determined we had reached saturation. For our experiments, we used values of r from 2 to 1000 (full dataset).

Figure 2 shows the number of samples needed before the relative error (when computed on the full dataset) is negligible, compared with the total B-factor of the protein (a statistic that attempts to encorporate the entire molecular uncertainty) and number of atoms. As can be seen in this figure, neither total B-factor nor number of atoms is capable of predicting the number of samples needed. Table 3 shows the number of samples needed to reach saturation for a number of QOI, compared to the number of samples predicted by our incremental method. This table shows our incremental sampling method was sufficient to predict the number of samples necessary on the full dataset. For instance, the Chernoff-bound calculated for on the incremental method for PB energy with 1OPH chain A reached saturation after 287 samples; with only 174 samples, we were able to achieve Chernoff-like bounds with at least 95% accuracy, when compared to the full dataset. This trend was repeated over all 104 individual chains, suggesting that the incremental approach is a good method to use when a "full dataset" has not already been computed. Figure 3 shows a plot of both of these metrics as the number of samples increases.

Table 3: Number of samples necessary to converge on the calculation for the Chernoff-like bound. Here we assume the process to have converged if the difference between the predicted ϵ did not change significantly. Detailed description of this calculation can be found in the text. The pair of entries in each cell of this table refer to the two different variants of convergence we considered. First, incremental-where each value is compared to the value generated when using 10 fewer samples. Second, where the value is compared with the value we get with 1000 samples. The data presented here correspond to PDBID:1OPH, chains A and B.

	B	A	ΔB	ΔA
area	230/134	233/153	215/146	351/197
vol	119/72	102/64	103/55	332/205
LJ	79/49	240/168	315/133	324/192
CP	79/43	114/62	93/71	355/213
GB	281/143	319/202	300/207	326/211
PB	287/174	365/209	357/206	348/196

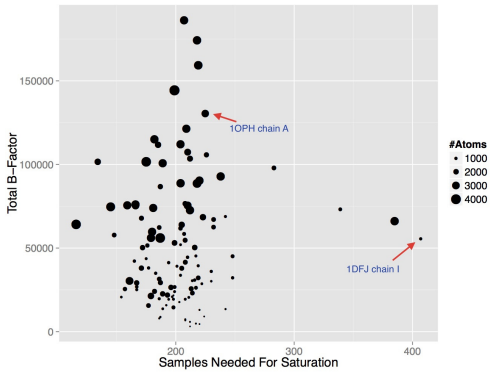

Figure 2: Plot of total B-factor (a measure of both size and uncertainty) against number of samples needed before the relative error is negligible ($\tau = 0.05$). For similar graphs on additional QOIs, please see [38].

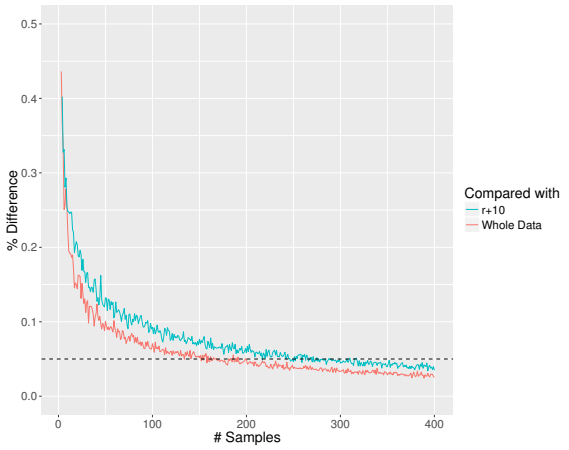

Figure 3: Plot showing the rate of convergence for statistical certificates, computed for the calculation of free energy (MM-PBSA). For each number of samples, r, the Chernoff-like bounds were computed, and then compared with either those computed on the entire dataset (red line) or a partial dataset containing $r + 10$ samples (blue line). Noise in the plotted lines are due to differences in samples selected; reported values are the average over 50 trials. Plot is for 1OPH chain A.

3.3 Visualizing Uncertainties in Molecular Properties

3.3.1 Visualizing uncertainties in computed QOIs

While there are many methods for visualizing uncertainties in molecular structure (i.e. coloring by B-factor), these methods highlight the inherent uncertainties in the molecular structure and their parameterization, but do not directly highlight the effect of these uncertainties on computed properties of the molecule. Specifically, we are interested in bounding the propagated uncertainty in the calculated property, and also localize the origins of uncertainty which disproportionately affect the outcome. This is carried out using the statistical QMC framework described above. Below we discuss some techniques which allows one to visually explore such uncertainties.

Pseudo-electron cloud.

One method for visualizing such uncertainty is a pseudo-electron cloud, where samples are combined into a single volumetric map whose voxels represent the likelihood (over the set of samples) of an atom occupying the voxel. Figure 4(A) shows such a visualization for 1OPH chain A. Note that this data is not simply useful for visualization, but can be used as the representation of the shape of the molecule for docking and fitting exercises to incorporate the input uncertainties directly into the scoring functions.

Localized uncertainty in molecular surface calculations.

In many applications, instead of a volumetric map, one uses a smooth surface model to compute QOIs like area, volume, curvature, interface area etc. In such cases, a visualization like Figure 4(B) can be very descriptive. It shows

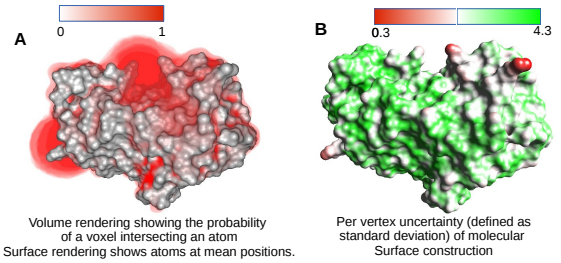

Figure 4: Visualization of molecular surface uncertainties. (A) A volumetric map showing the likelihood of the voxel being occupied by an atom, computed using a sampling of the joint probability distribution of the atom positions. (B) Expected deviation of each point on the surface of a single model, w.r.t. all models sampled based on the joint distribution of the locations of the atoms. Green colored regions are expected to remain more or less stable in any sample, red colored ones may vary a lot.

a single smooth surface model (based on the original/mean coordinates), and the colors at each point on the surface show the average distance of that point from all surfaces generated by sampling the joint distribution. Unsurprisingly, most parts of the surface in the figure has very low deviation, and only the narrow and dangling parts have high deviation. Comparing this with the rendering of B-factors (in Figure 5(A)) shows that even though some parts of the surface are in regions with high B-factors, the uncertainties do not affect the surface computation as much. Hence, higher temperature factors may not always result in a higher uncertainty in computed property, and a sensitivity analysis with low discrepancy sampling is warranted.

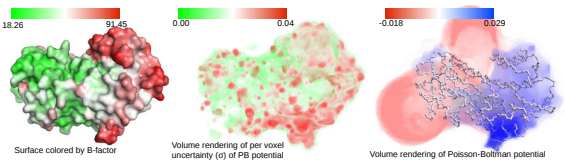

Figure 5: Visualization of molecular energy uncertainties. (A) Simple mapping of B-factors to the surface of the protein. (B) Expected deviation of potential energy of a given voxel, w.r.t. all models sampled based on the joint distribution of the locations of the atoms. Green colored regions are expected to remain more or less stable in any sample, red colored ones may vary a lot. (C) Display of potential energy map averaged over all samples.

Uncertainty in energy calculations.

Now we focus on QOIs that are computed based on other intermediate QOIs. For example, computation of MM-PBSA energy first evaluates the PB potential on a volume which encapsulates the molecule and the solvent. This potential calculation itself requires a smooth surface representation of the molecule as input, along with the positions and charges of the atoms. In this case, as well as bounding the overall uncertainty for the final value of the PB energy, we can also

bound the uncertainties of the intermediate PB potentials calculated at each voxel. We do this by defining the PB potential at a voxel as separate QOIs and apply the QMC sampling to generate an ensemble of atomic models and smooth surfaces, then evaluate the QOIs for each sample. Hence, we derive a distribution for each voxel. The means of these distributions are rendered in Figure 5(C), showing the negative and positive potential regions. The standard deviations of the distributions are rendered in Figure 5(B). Comparing the original uncertainties (B-factors in 5(A)) to these third level propagated uncertainties shows that while in some regions the uncertainties had a cancellation effect, in some other regions they amplified.

4. CONCLUSIONS

In this article we have shown that even subtle uncertainties present in high resolution X-ray structures can lead to significant error in computational modeling. Such errors are propagated and compounded when output from one stage of modeling are used in the next. We considered the uncertainties in atomic position reported through B-factors and evaluated how they create uncertainty in computed quantities of interest (e.g. surface area, van der Waals energy, solvation energy, etc.). While some existing computational protocols attempt to bound the uncertainties/error due to algorithmic or numerical approximations, they do not account for the uncertainties in the input. However, our empirical study on 57 x-ray structures of bound protein complexes showed that there significant probability ($> 10\%$) of having more than 5% error in total energy (PBSA) calculation purely due to the input uncertainties. Hence, one must account for and bound such uncertainties.

We have shown that input uncertainties can be modeled as random variables and the uncertainty of the computed outcome (a dependent random variable) can be bounded using Chernoff-like bounds introduced by Azuma and McDiarmid. We have also shown that such bounds are also applicable when the input random variables are dependent, and show how one can theoretically bound the probability of error for Coulombic potential calculation (and any summation of distance dependent decaying kernels in general, see [38] for full derivation). In the future, we aim to derive similar bounds for other biophysically relevant functions.

We have also introduced an empirical quasi-Monte Carlo approximation method based on sampling the joint distribution of the input random variables to produce an ensemble of models. The ensemble is used to approximate a distribution of values for the quantity of interest. This distribution in turn can be used to bound the uncertainty of the calculation in terms of statistical certificates. A very interesting and promising outcome from application of this framework to a large set of protein structures for a wide variety of calculations showed that one typically needs fewer than 500 samples before the QMC procedure converges, hence it is quite practical to perform and report such certificates in modeling exercises. We are currently working on a graphical model of the input uncertainties, which should possibly lead to even better convergence.

We have also shown that many of the current methods for visualizing protein uncertainty is limited: displaying surface uncertainty simply by B-factor is insufficient, as uncertainty from X-ray crystallography does not necessarily track natural shifts in protein conformation. We have displayed several different visualization techniques for displaying not only atomic uncertainty, but also uncertainty in energy calculations. Displaying 3-dimensional uncertainty of quantities such as the Poisson Boltzman potential can provide valuable information that a single potential map cannot.

5. ACKNOWLEDGEMENTS

This research was supported in part by a grant from NIH (R01-GM117594) and contract (BD-4485) from Sandia National Labs.

6. REFERENCES

[1] K. Azuma. Weighted sums of certain dependent random variables. *Tokuku Mathematical Journal*, 19:357–367, 1967.

[2] C. Bajaj, A. Bhowmick, E. Chattopadhyay, and D. Zuckerman. On low discrepancy samplings in product spaces of motion groups. *arXiv preprint arXiv:1411.7753*, 2014.

[3] C. Bajaj, S.-C. Chen, and A. Rand. An efficient higher-order fast multipole boundary element solution for Poisson-Boltzmann based molecular electrostatics. *SIAM J. Sci. Comput.*, 33(2):826–848, 2011.

[4] C. Bajaj, P. Djeu, V. Siddavanahalli, and A. Thane. TexMol: Interactive visual exploration of large flexible multi-component molecular complexes. In *Proc. IEEE Visual. Conf.*, pages 243–250, Austin, Texas, 2004.

[5] C. Bajaj, H. Lee, R. Merkert, and V. Pascucci. NURBS based B-rep models from macromolecules and their properties. In *Proc. Symp. Solid Model. Appl.*, pages 217–228, 1997.

[6] C. Bajaj, V. Pascucci, A. Shamir, R. Holt, and A. Netravali. Dynamic maintenance and visualization of molecular surfaces. *Discrete Appl. Math.*, 127:23–51, 2003.

[7] C. Bajaj and V. Siddavanahalli. F2Dock: A Fast and Fourier Based Error-Bounded Approach to Protein-Protein Docking. CS Technical report TR-06-57, The University of Texas at Austin, Austin, TX, USA 78712., November 2006.

[8] C. Bajaj and W. Zhao. Fast molecular solvation energetics and forces computation. *SIAM J. Sci. Comput.*, 31(6):4524–4552, 2010.

[9] D. Bashford and D. A. Case. Generalized Born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.*, 51:129–152, 2000.

[10] G. E. Box and M. E. Muller. A note on the generation of random normal deviates. *The annals of mathematical statistics*, (29):610–611, 1958.

[11] A. T. Briinger. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, 355:472–475, 1992.

[12] B. T. Burnley, P. V. Afonine, P. D. Adams, and P. Gros. Modelling dynamics in protein crystal structures by ensemble refinement. *Elife*, 1:e00311, 2012.

[13] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods. The amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16):1668–1688, 2005.

[14] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952.

[15] R. Chowdhury, D. Keidel, M. Moussalem, M. Rasheed, A. Olson, M. Sanner, and C. Bajaj. Protein-protein docking with F^2Dock 2.0 and $GB-rerank$. *Biophys. J.*, 8(3):1–19, 2013.

[16] M. Connolly. Analytical molecular surface calculation. *J. Appl. Cryst.*, 16:548–558, 1983.

[17] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.

[18] D. Cruickshank. Remarks about protein structure precision. *Acta Crystallographica Section D: Biological Crystallography*, 55(3):583–601, 1999.

[19] T. Dolinsky, J. Nielsen, J. McCammon, and N. Baker. Pdb2pqr: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Research*, 32:665–667, 2004.

[20] D. Eisenberg and A. Mclachlan. Solvation energy in protein folding and binding. *Nature (London)*, 319:199–203, 1986.

[21] N. Eswar, B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M.-Y. Shen, U. Pieper, and A. Sali. Comparative protein structure modeling using MODELLER. *Curr. Prot. in Protein Sc.*, Chapter 2:Unit 2.9, 2007.

[22] M. Feig and C. Brooks. Recent advances in the development and application implicit solvent models in biomolecule simulations. *Current Opinion in Structural Biology*, 14:217, 2004.

[23] P. Gopalan, R. Meka, O. Reingold, and D. Zuckerman. Pseudorandom generators for combinatorial shapes. *SIAM J. Comput.*, 42(3):1051–1076, 2013.

[24] M. Habeck, M. Nilges, and W. Rieping. Bayesian inference applied to macromolecular structure determination. *Physical Review E*, 72(3):031912, 2005.

[25] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[26] H. Hwang, T. Vreven, J. Janin, and Z. Weng. Protein–protein docking benchmark version 4.0. *Proteins: Structure, Function, and Bioinformatics*, 78(15):3111–3114, 2010.

[27] F. James, J. Hoogland, and R. Kleiss. Quasi-monte carlo, discrepancies and error estimates. *Methods*, page 9, 1996.

[28] A. Kuzmanic, N. S. Pannu, and B. Zagrovic. X-ray refinement significantly underestimates the level of microscopic heterogeneity in biomolecular crystals. *Nature communications*, 5, 2014.

[29] A. Kuzmanic and B. Zagrovic. Determination of ensemble-average pairwise root mean-square deviation from experimental b-factors. *Biophysical journal*, 98(5):861–871, 2010.

[30] P. T. Lang, H.-L. Ng, J. S. Fraser, J. E. Corn, N. Echols, M. Sales, J. M. Holton, and T. Alber. Automated electron-density sampling reveals widespread conformational polymorphism in proteins. *Protein Science*, 19(7):1420–1431, 2010.

[31] H. Lei, X. Yang, B. Zheng, G. Lin, and N. A. Baker. Quantifying the influence of conformational uncertainty in biomolecular solvation. *arXiv preprint arXiv:1408.5629*, 2014.

[32] Y. Lei and R. R. Mettu. A confidence measure for model fitting with x-ray crystallography data. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, page 489. ACM, 2013.

[33] C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, 141(141):148–188, 1989.

[34] H. Niederreiter. Quasi-monte carlo methods. *Encyclopedia of Quantitative Finance*, 24(1):55–61, 1990.

[35] M. Nina, D. Beglov, and B. Roux. Atomic radii for continuum electrostatics calculations based on molecular dynamics free energy simulations. *J. Phys. Chem. B*, 101:5239–5248, 1997.

[36] A. Onufriev, D. Bashford, and D. Case. Modification of the generalized Born model suitable for macromolecules. *J. Phys. Chem. B*, 104:3712–3720, 2000.

[37] J. Philip. *The probability distribution of the distance between two random points in a box*. KTH mathematics, Royal Institute of Technology, Stockholm, Sweden, 2007.

[38] M. Rasheed, N. Clement, A. Bhowmick, and C. Bajaj. Quantifying and visualizing uncertainties in molecular models. *arXiv preprint arXiv:1508.03882*, 2015.

[39] F. Richards. Areas, volumes, packing, and protein structure. *Annu. Rev. Biophys. Bioeng.*, 6:151–176, 1977.

[40] W. Rieping, M. Habeck, and M. Nilges. Inferential structure determination. *Science*, 309(5732):303–306, 2005.

[41] T. R. Schneider. What can we learn from anisotropic temperature factors. *Proceedings of the CCP4 Study Weekend (Dodson, E., Moore, M., Ralph, A. & Bailey, S., eds)*, pages 133–144, 1996.

[42] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.3r1. PyMOL The PyMOL Molecular Graphics System, Version 1.3, Schrödinger, LLC., August 2010.

[43] N. J. Sloane. Online Encyclopedia of Integer Sequences (OEIS), 2014. http://oeis.org/A103986.

[44] W. G. Touw and G. Vriend. Bdb: databank of pdb files with consistent b-factors. *Protein Eng Des Sel*, 27(11):457–462, Nov 2014.

[45] O. Trott and A. J. Olson. Autodock vina. *J. Comput. Chem.*, 31:445–461, 2010.

[46] K. Trueblood, H.-B. Bürgi, H. Burzlaff, J. Dunitz, C. Gramaccioli, H. Schulz, U. Shmueli, and S. Abrahams. Atomic dispacement parameter nomenclature. Report of a subcommittee on atomic displacement parameter nomenclature. *Acta Crystallographica Section A: Foundations of Crystallography*, 52(5):770–781, 1996.