Statistical Framework for Uncertainty Quantification in Computational Molecular Modeling

Muhibur Rasheed, Nathan Clement, Abhishek Bhowmick, and Chandrajit L. Bajaj

Abstract—As computational modeling, simulation, and predictions are becoming integral parts of biomedical pipelines, it behooves us to emphasize the reliability of the computational protocol. For any reported quantity of interest (QOI), one must also compute and report a measure of the uncertainty or error associated with the QOI. This is especially important in molecular modeling, since in most practical applications the inputs to the computational protocol are often noisy, incomplete, or low-resolution. Unfortunately, currently available modeling tools do not account for uncertainties and their effect on the final QOIs with sufficient rigor. We have developed a statistical framework that expresses the uncertainty of the QOI as the probability that the reported value deviates from the true value by more than some user-defined threshold. First, we provide a theoretical approach where this probability can be bounded using Azuma-Hoeffding like inequalities. Second, we approximate this probability empirically by sampling the space of uncertainties of the input and provide applications of our framework to bound uncertainties of several QOIs commonly used in molecular modeling. Finally, we also present several visualization techniques to effectively and quantitavely visualize the uncertainties: in the input, final QOIs, and also intermediate states.

Index Terms—Uncertainty quantification, sampling, molecular modeling

1 Introduction

Computational modeling of any physical system is inherently imperfect due to a myriad of shortcomings. A computational model is often a discrete representation of a continuous model of reality. Additionally, to achieve computational tractability, one uses simplified model formulations, and employs algorithmic approximations, often using coarse samplings of parameter/search spaces. Furthermore, the available observed data of the physical system may itself be noisy, incomplete, or from a different context, resulting in our inability to capture all relevant factors of the system. Moreover the ones we do capture, all possess a level of uncertainty. 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 in a multi-stage protocol—they can create a result that is very unreliable.

Computational molecular modeling is a sub-field of research that is especially susceptible to the accumulation of cascadic errors for many computed molecular properties, generally defined as quantities of interest (QOI). The computations for these QOI include multi-step methods for protein sequence alignment and homology modeling, implicit solvation interfaces (i.e., molecular surfaces) generation [1], [2], [3], [4], [5], [6], configuration-dependent binding affinity calculations [7], [8], molecular docking and structure refinement via molecular substructure replacement and fitting [9], [10], [11], [12], 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 and a bound on its uncertainty.

Unfortunately, the majority of current computational structure modeling and prediction protocols do not report the confidence on the final model, quantity or prediction they compute, and even when they do, they fail to rigorously consider uncertainties in the input. For instance, structure prediction protocols (including fitting, docking, homology modeling, etc.) addresses uncertainties in an indirect way by reporting several structures ranked under some scoring function, f, with the assumption that at least one of the predicted structural models would be close to the truth. However, it is not clear how to ascertain the quality or confidence on individual models in the ranked list or whether a near-accurate structure model is present in the entire ranked list at all.

Similarly, protocols for computing specific properties of molecules like surface area, binding free energy, solvation, etc., usually 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, [14] for X-ray crystallography or [15], [16] for NMR structure prediction using probabilistic analysis),

• *The authors were with the Department of Computer Science, University of Texas at Austin, Austin, TX 78705. E-mail: muhibur@utexas.edu, {nclement, bajaj}@cs.utexas.edu, ab.abhishek.bhowmick@gmail.com.*

Manuscript received 4 Dec. 2016; revised 22 May 2017; accepted 25 July 2017. Date of publication 22 Nov. 2017; date of current version 5 Aug. 2019.
(Corresponding author: Chandrajit L. Bajaj.)
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2017.2771240

determining how this uncertainty propagates to future stages in a pipeline is left unaddressed. An exception is recent work by Lei et al. [17] 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 article, we present a mathematical and an empirical framework, both of which take into account uncertainties present in the input to any computational step and provide an upper bound on the uncertainty of the outcome. We also provide intuitive visualization tools to visually inspect uncertainties on a structural model at different stages of a pipeline. We believe these to be invaluable additions to rational design and analysis protocols in molecular modeling.

We define statistical uncertainty quantification as a certificate expressed as a tail bound $Pr[|f - E[f]| > t] < \epsilon$. In other words, a probabilistic certificate is a function of a user-defined 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 our framework, we treat each component of $\mathbf{X}$ as random variables (RVs). Then we adopt a method of bounded differences, which is a modification of Markov and Chebyshev inequalities, used for independent RVs to provide Chernoff-like bounds. However, this is not quite applicable to most biophysical QOIs since the components of $\mathbf{X}$ need not necessarily be independent. We show that for such cases, a variation of the Chernoff-Hoeffding bound, namely the Azuma inequality [18], 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 McDiarmid's inequality [19]. In the methods section, we describe this framework in greater detail and also show an example application to compute such certificates for functions with decaying kernels. In molecular biology applications, this family of functions include the van der Waals interaction energy, atom-atom contact potentials based on distance cutoffs, integrals over point neighborhoods, etc.

McDiarmid bounds often tend to be too conservative (capturing the worst possible case). Furthermore, analytically deriving such bounds is quite challenging for more complicated functions. To address this, we have also developed an alternative technique where we approximate the distribution of values of the QOIs over the space of input uncertainty, and then estimate the tail bounds based on the distribution. The reliability of this estimation hinges upon a sufficient and low-discrepancy sampling of the uncertainty space. We employed a recently-developed pseudo-random sampling algorithm which requires fewer samples to achieve the desired accuracy. Our empirical analysis over a diverse set of proteins (from the Zlab benchmark [20]) showed that a fairly small number of samples often suffices to generate a robust approximation of the distribution of the QOI.

Our framework is general, and could theoretically be applied to model the uncertainties of any QOI under uncertainties from any source. For the purposes of this article we have chosen to limit our applications and examples to a selected few QOIs and sources of uncertainties, leaving our software tools open to researchers who would like to explore any other of the many possible sources of uncertainty and QOIs. We have chosen to understand the effect of small positional uncertainties of atoms in high resolution crystal structures. We used B-factors reported in PDB files as an implicit description of positional uncertainty of an atom. 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-Boltzmann (PBSA) models for single molecules; as well as interface area, and binding free energy calculation for pairs of bound molecules. (For completeness, we also describe and provide some results from a protocol to identify large-scale conformational changes based on changes in internal angles.)

The use of B-factors as a representation of the positional uncertainty comes with two simplifying assumptions. First, we assume that the reported B-factors are accurate, even though there can be other equally-good or better assignments of positional uncertainty based on the same electron scattering data. For methods that improve the B-factor estimation see, among others, [21], [22], [23], [24], [25]. Second, even though there may be correlations between the B-factors of collections of atoms—primarily due to the way positions and uncertainties are resolved from raw data—we treat each coordinate as independent random variables. This assumption can be relaxed, but would require a slightly more involved definition of the distribution of uncertainties (see Methods for details).

Our empirical study on 57 x-ray structures of bound protein complexes showed that positional uncertainties translate to relatively low uncertainty for simple quantities such as exposed surface area (e.g., $\sim$5 percent probability of having more than 2 percent error), but significant uncertainties for complex QOIs such as total energy (e.g., >10 percent probability of having more than 5 percent error). Our study clearly establishes the value in computing and reporting such certificates, to add credibility (or caution) to reported values, especially for complex QOIs which involve *propagation* of uncertainty from simpler QOI in the calculation. Furthermore, we found that 500 samples were more than sufficient to compute reliable certificates for any QOI we considered, even for molecules involving more than a thousand atoms (having thousands of dimensions in the uncertainty space).

The distribution of values from the quasi-Monte Carlo protocol can be used as a rich source of data for quantitative visualization of uncertainties at different levels of granularity. Current visualization and modeling tools (for example, PyMol [13], Chimera [26], Coot [27], JMol [28], etc.) allow one to visualize B-factors using color maps on the atoms, smooth surface, or scalar field (or volume). However, they fail to highlight the functional relevance of these input uncertainties. For example, if the QOI is the optimal conformation of a ligand when it binds to a particular protein found using computational means (typically by minimizing a scoring function such as PBSA energy (see Fig. 1C)), the uncertainties of the atoms near the binding site would have a higher effect on the QOI. We present visualization techniques that reflect exactly how the uncertainties of different atoms affect the QOI (see Fig. 1B, for example). These provide functionally important information about the molecule and aid rational design by focusing on more significant sets

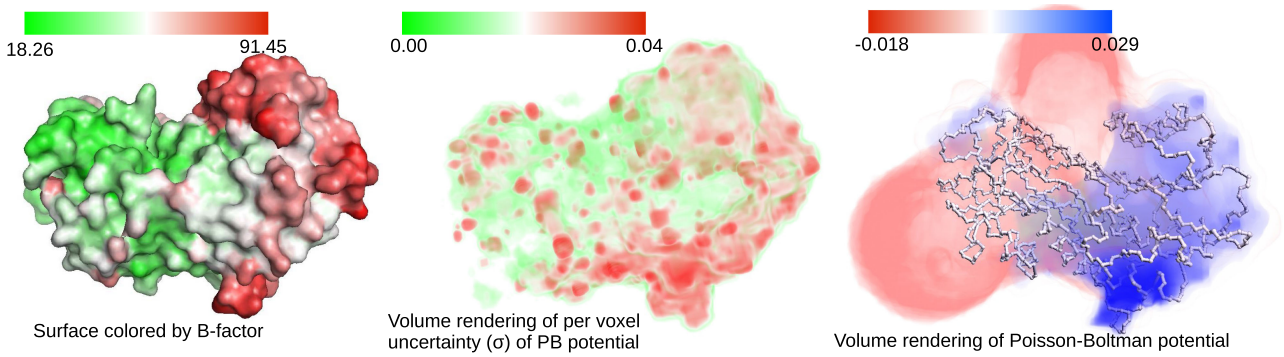


Fig. 1. *An illustrative example of quantifying and visualizing uncertainties.* Atomic coordinates of molecules, are often reported with a measure of uncertainty (e.g., B-factors). Currently available software hardly incorporates the effect of such uncertainties into their results or visualization. For example, currently, one is able to see the distribution of high and low B-factors on the structural model of a molecule (1OPH) as is shown in (A), rendered using PyMol [13]. However, we would like to understand the uncertainty in a computed quantity of interest (QOI), and also understand how the input uncertainties affect this outcome. Consider, for example, a specific QOI, the Poisson-Boltzman (PB) electrostatic potential inside and outside the same molecule. We show this potential field in (C), the molecule itself is rendered as its backbone only. The potential is rendered using blue (positive), white (neutral), and red (negative) colors. Now, using our empirical uncertainty quantification model, we can (and should) further compute the uncertainty, expressed as the standard deviation, σ, computed across an ensemble of slightly perturbed samples of the molecule, of the potential computed at every voxel. This uncertainty is visualized in (B). Note that the standard deviation is quite high in some regions. Notice that while (A) and (B) have some correlation (high uncertainty in structure tend to correlate with high uncertainty in function), it provides a more accurate and application-specific picture of the underlying uncertainty.

of atoms. In this article we present several novel, relevant, quantitative and easy to interpret visualization techniques and enhance existing ones to aid different steps of a typical molecular modeling pipeline.

We have developed software which implements the mathematical framework of sampling needed for statistical UQ, use existing tools to compute the QOIs [7], [29], [30], [31], [32], and compute uncertainty bounds as well the visualization directives which can be directly loaded into existing molecular, surface/volume visualization software [13], [33]. We envision our methods and tools would enable the end users tools to achieve both quantitative and visual evaluation of various molecular modeling QOIs for correctness—or the lack thereof.

2 Methods

Statistical uncertainty quantification aims to provide a certificate bounding the probability of error in the QOI. Such a certificate can be expressed mathematically as a Chernoff-Hoeffding [34] like bound as follows:

$$Prob(f, X, t, \epsilon) \triangleq \Pr[|f(X) - E[f]| > t] \leq \epsilon, \qquad (1)$$

where $f(\mathbf{X})$ is the QOI computed on noisy data, $\mathbf{X}$, and t is a user-defined threshold. The certificate reports ϵ, a probability that the error in $f(\mathbf{X})$ is greater than t. In other words, we want to guarantee that the probability of the error being over the threshold is very small.

In this article, we adopt a loose uncertainty bound on Doob martingales as introduced by Azuma [18] and Hoeffding [35] and later extended by McDiarmid [19]. 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 $t > 0$*

$$\Pr[|f(\mathbf{X}) - \mathbf{E}[f]| > t] < 2 \exp\left(-2t^2 \Big/ \sum_k c_k^2\right).$$

We present the derivation of the McDiarmid bound and prove that it is applicable to general cases, even when the variables are not independent; followed by an example application for a function involving summations over decaying kernels in the next couple of sections.

McDiarmid bounds often overestimate the error, and it is often not easy to compute c_k analytically for many QOIs with complex functionals. An alternate approach is to empirically compute the certificates using quasi-Monte Carlo (QMC) methods [36], [37]. Assuming that the distributions of the input RVs are known, we sample the joint space and evaluate f for each, leading to an estimation of the distribution of f over the joint uncertainty space. Then, it becomes trivial to compute the uncertainty of individual values of the QOI, as well as providing certificates like Equation (1). Correctness of this empirical approach depends on the quality and size of the samples. We discuss our characterization of the input uncertainty space and the sampling technique in later sections, and provide experimental results which show that certificates can be robustly computed using significantly few samples than the dimension of the space would imply.

2.1 Thoretical Framework for Statistical Uncertainty Quantification

One often uses Chernoff-Hoeffding style bounds to provide uncertainty bounds theoretically when the underlying RVs are independent and one is analyzing the sum of the random variables. In practical situations the random variables have dependencies. In such cases, one can still prove large deviation bounds using the theory of martingales, specifically Doob martingales and their extension.

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 ith observable is the same as the observed value of the $(i-1)$th, irrespective of

the values of all other observables. The Azuma inequality for martingales bounds errors as follows:

Claim 2.3 (Azuma inequality). Let (X_i) be martingale with respect to (Z_i). Suppose $|X_i - X_{i-1}| \leq c_i$. Then

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

Now consider a variation, the Doob martingale, constructed in the following way:

Claim 2.4. 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).

The weak form of the McDiarmid inequality (below) is an analog, and derives from the Azuma inequality:

Claim 2.5. Let (X_i) be independent random variables. Let $f : \prod_i A_i \to \mathbb{R}$ for sets A_i. Also, suppose that $|f(x_1, \ldots, x_k, \ldots, x_n) - f(x_1, \ldots, x'_k, \ldots, x_n)| \leq c_k$. Then, for $t > 0$

$$\Pr[|f(\mathbf{X}) - \mathbf{E}[f]| > t] < 2\exp\left(-2t^2 \Big/ \sum_k c_k^2\right).$$

This now completes the derivation of the McDiarmid bound given in Definition 2.1.

2.1.1 Relaxing Independence Requirements by Extensions of Doob Martingale

McDiarmid's inequality assumes that the function has Lipschitz properties and the RVs are independent of each other. This lead to relatively clean bounds. However, if one wishes to analyze a very general scenario, then we can proceed as follows.

First, there is a sequence of random variables $(X_i)_{i=1}^n$ taking values in some set A. They can be dependent in any way. Consider any function $f : A^n \to \mathbb{R}$. Then, by Azuma's inequality (mentioned earlier), we can have a certificate bound of the form

$$\Pr[|f(\mathbf{X}) - \mathbf{E}[f]| > t] \leq \exp\left(-t^2/2\sum_i c_i^2\right),$$

where the only assumption we need is

$$|\mathbf{E}_{X_{i+1},\ldots,X_n}[f(\mathbf{X})|X_1, \ldots, X_i] - \mathbf{E}_{X_i,\ldots,X_n}[f(\mathbf{X})|X_1, \ldots, X_{i-1}]| \leq c_i.$$

Thus, the change in expectation on fixing the ith random variable should not be too large. One can easily see that the conditions required for McDiarmid's inequality immediately imply the above hypothesis and thus the certificate bound follows. However, one can also put practically minimal restriction on the random variables and do the above computation on the amount of perturbation in the expectation, at the cost of notational aesthetics.

We give more details now. As before there is a sequence of random variables $(X_i)_{i=1}^n$ taking values in a set A, say, arbitrarily dependent on each other. Consider any function $f : A^n \to \mathbb{R}$. Define a sequence of random variables for $i = 1$ to $n - 1$

$$B_i = \mathbf{E}_{X_{i+1},\ldots,X_n} f(\mathbf{X}).$$

By definition, B_i is a function of $X_1, \ldots, X_i$. Note that the sequence $(B_i)_{i=1}^n$ forms a martingale with respect to X_i.

Suppose we had $|B_{i+1} - B_i| \leq c_i$ for $i = 1$ to $n - 1$. Note that

$$B_0 = \mathbf{E}_{\mathbf{X}} f(X),$$

and

$$B_n = f(\mathbf{X}).$$

Then, by using Azuma's inequality on the martingale sequence (B_i), we get

$$\Pr[|f(\mathbf{X}) - \mathbf{E}[f]| > t] \leq \exp\left(-t^2/2\sum_i c_i^2\right).$$

Note that the only requirement we needed for the certificate bound was $|B_{i+1} - B_i| \leq c_i$. We placed no restriction on the underlying random variables. We will now reduce this to a slightly stricter, albeit easier to analyze requirement.

Suppose for every i, and every $x_1, \ldots, x_n, x'_i$, we have

$$\left|\mathbf{E}_{X_i|X_{i+1},\ldots,X_n} f(\mathbf{X}) - f(x_1, \ldots, x'_i, x_{i+1}, \ldots, x_n)\right| \leq c_i.$$

Then

$$\begin{aligned} B_{i-1} - B_i &= \mathbf{E}_{X_i,\ldots,X_n} f(\mathbf{X}) - \mathbf{E}_{X_{i+1},\ldots,X_n} f(\mathbf{X}) \\ &= \mathbf{E}_{X_{i+1},\ldots,X_n} \mathbf{E}_{X_i|X_{i+1},\ldots,X_n} f(\mathbf{X}) \\ &\quad - f(x_1, \ldots, x'_i, x_{i+1}, \ldots, x_n), \end{aligned}$$

which is bounded by c_i by assumption. Therefore, we can then use Azuma's inequality on the above martingale.

We highlight this with a simple illustration. Consider the single kernel model at a single point. This will illustrate the main point. We will again consider the 2 dimensional kernel defined by

$$f(x, y) = \frac{a}{(x^2 + y^2)^{b/2}}.$$

Let (X, Y) be a variable following a joint distribution. Note that we do not require independence between X and Y. Define

$$\begin{aligned} B_0(x, y) &= f(x, y), \\ B_1(x) &= \mathbf{E}_Y[f(x, Y)], \\ B_2 &= \mathbf{E}_{X,Y}[f(X, Y)]. \end{aligned}$$

We will demonstrate that for most reasonable distributions (e.g., those that are Lipschitz-bounded), it is relatively straightforward to prove

$$|\mathbf{E}_{Y|X=x}[f(x', Y)] - f(x, Y)| \leq c.$$

This immediately implies that both

$$|B_1 - B_0|, |B_2 - B_1| \leq c.$$

Using Azuma's inequality, we conclude the required large deviation bound certificate.

2.2 Analytical Uncertainty Bounds for Biophysical Quantities

The theoretical framework presented above can be applied to functions that possess the Lipschitz property. Here we consider a function which is expressed as a summation over decaying kernels of the form shown below

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

where b_k are non-negative constants, a_k are constants, and A and B are two sets of points, such that each of the points are uncertain.

We chose this function since it fits many molecular QOIs. For example, this maps easily to the 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 quantities of interest. In such applications the uncertain quantities will be the positions of the atoms.

In the following, we introduce some notation 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

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

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

Lemma 2.6. *For a given set of a_k and b_k, $f_{\mathbf{x_1}}(\mathbf{x_2}) = f_{\mathbf{0}}(\Delta\mathbf{x})$ where $\Delta\mathbf{x} = (\mathbf{x_2}-\mathbf{x_1})$.*

When both $\mathbf{x_1}$ and $\mathbf{x_2}$ are uncertain but bounded such that every component x_{1i} of $\mathbf{x_1}$ is sampled from the interval $[l_{1i}, u_{1i}]$, and every component x_{2i} of $\mathbf{x_2}$ is sampled from the interval $[l_{2i}, u_{2i}]$, then we can assume that every component Δx_i of $\Delta\mathbf{x}$ is also bounded by 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 $f_{\mathbf{0}}(\Delta\mathbf{x})$ due to the uncertainty of $\Delta\mathbf{x}$. 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 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

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

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

Lemma 2.7. *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, $f_{2_i}(\mathbf{x}, y)$ represent $f_2(\mathbf{x})$ such that the value of the ith 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 2.8. *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 2.7 and 2.8 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 kth kernel.

Lemma 2.9. *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 2.10 follows immediately from Lemma 2.9.

Lemma 2.10. *For the sum of decaying kernel $f_4(A, \mathbf{y})$ given in Equation (11)*

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

such that $D_{3i}(\Delta \mathbf{x})$ 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 2.11. *For the sum of decaying kernel $F(A, B)$ given in Equation (11)*

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

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

2.3 Empirical Uncertainty Quantification

The basis of empirical uncertainty quantification is to approximate the distribution of the function $f(\mathbf{X})$ under the space of perturbations of the uncertain variables $\mathbf{X}$. This approximation is done by sufficiently exploring the space and accumulating the measurements/observations or computed values of f. Depending on $\mathbf{X}$ and the function f, different exploration methods can be beneficial. We have chosen to use the quasi-Monte Carlo method. This is general and one can prove that the approximation of the distribution of $f(\mathbf{X})$ produced by QMC has bounded error.

Bounded Error of Estimation Through Low-Discrepancy Sampling. We define the modulus of continuity, $\omega(f, t)$, for a function f on a d-dimensional product space $\mathcal{I}^d$, as

$$\omega(f, t) = \max_{u, v \in \mathcal{I}^d \& \delta(u, v) \leq t} |f(u) - f(v)|, \tag{13}$$

where $\delta(u, v)$ is the distance between two samples. Essentially, $\omega(f, t)$ is the maximum change of f between two close samples.

Now, we define the discrepancy of a set P of N samples, with respect to a collection of subsets, $\mathcal{X}$, as

$$D_N(P, \mathcal{X}) = \max_{X \in \mathcal{X}} \left(\frac{|P \cap X|}{|P|} - \frac{\mu(X)}{\mu(\mathcal{U})} \right), \tag{14}$$

where μ is the Lebesgue measure (high-dimensional volume), and $\mathcal{U}$ is the universe. In other words, $D_N(P, \mathcal{X})$ captures the how evenly the points cover the universe.

The following theorem establishes the bound on the approximation error for the distribution of f (adapted from Theorem 2.13 of [38]).

Theorem. *If f is continuous in $\mathcal{I}^d$, then, for any set of samples $P = \{x_1, x_2, \ldots, x_N\}$ such that $x_i \in \mathcal{I}^d$, we have*

$$\left| \int_{\mathcal{I}^d} f(u) du - \frac{1}{N} \sum_{n=1}^{N} f(x_n) \right| \leq 4\omega\left(f; \left(D_N^*(P) \right)^{1/d} \right). \tag{15}$$

In our case, we want to approximate a distribution. Notice that the above theorem guarantees that if low-discrepancy sampling is performed, the cumulative distribution function (CDF), as well as the moments, will be approximated with bounded error. Note that a simpler QMC strategy can be applied to find the minimum c_k for each x_k, and hence derive the loose McDiarmid bound.

In the next few sections we detail the application of empirical UQ on molecular modeling scenarios. In particular, we discuss the identifying the sources of uncertainties, defining their distributions, defining the joint space and finally the specific sampling techniques.

2.3.1 Structural Uncertainties in Molecular Structures

The two most common representation of molecular structure in atomic detail both express the position of each atom- one using Cartesian coordinates, and the other using internal coordinates (which is a series of bond lengths, bond angles and dihedral angles). 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 can, in most cases, identify the expected locations of each atom by analyzing a 3D reconstructed electron density map of a molecule derived from the diffraction pattern from a crystal lattice of the molecule. Clearly, the expected location is determined as one from a distribution of possibilities: typically, one that best fits the density while satisfying other constraints including the protein's primary, secondary and tertiary structure, as well as biophysical interactions. To capture the inherent uncertainty and the distribution of other possible locations of the atom, a temperature factor or B-factor is also reported.

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 [39].

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, $T(\vec{h})$ [40]

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

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

$$T(|\vec{h}|) = \exp\left[-8\pi^2 \langle u^2 \rangle (\sin^2 \theta)/\lambda^2\right]. \tag{16}$$

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 [41] or diffraction-component precision index (DPI) [42] can be used to provide more insight into these uncertainties. However, throughout this paper we will use B-factors as they are commonly available in the PDB.

When using an internal coordinate system, the assumption is that the entire molecule is like a connected graph embedded in 3D space. Each node of the graph is an atom, and each edge represents a bond. So, knowing the position of any one atom, the location of all other atoms can be uniquely determined, given the bond length, bond angle (angle at a node between two bonds), and dihedral angles (given three successive bonds, the dihedral angle is defined as the angle between the two planes formed by bonds $1-2$ and $2-3$) of all other atoms. Moreover, internal coordinates successfully capture the dependence of one atom's position on the positions of all its neighbors. Also, since bond lengths and angles have been empirically observed to be constants (or nearly so), this representation allows one to reduce the number of degrees of freedom to only the change of dihedral angles (henceforth called torsion angles since the change is similar to a twisting motion around a bond).

2.3.2 Parametrizing the Space of Uncertainties

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

$$p(x_i) = \frac{1}{\sigma_i^x \sqrt{2\pi}} \exp\left[-\frac{(x_i - \mu_i^x)^2}{2\sigma^2}\right], \tag{17}$$

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 $B_i^x = B_i^y = B_i^z = B_i$.

To represent a protein with internal coordinates, the ϕ and ψ angles of the backbone are considered the random variables. We shall express each such degree of freedom using a random variable distributed uniformly between a range $[lower, upper]$ where the limits of the range are derived from the so-called Ramachandran plot [43], which is an empirical study of dihedral angle values observed in protein structures.

2.3.3 Sampling and Curating Configurations

The joint distribution, either defined as the product space of Gaussian distributions for the B-factor case, or of independent uniform distributions for the torsion angles case, represents the space of possible configurations for the molecule. Let N be the number of degrees of freedom in either representation, and start by generating a sample from the uniform distribution $U_k \sim \mathcal{U}(0,1)^N$ using a pseudorandom generator which guarantees low discrepancy sampling in high dimensional product spaces (described in the next section).

To convert each uniform sample to the joint distribution for the B-factor case, we use entries from U_k to produce a tuple $\langle u_1^x, \ldots, u_n^z \rangle$. Each number u_i^j from this tuple is mapped to a sample from a Normal distribution, a_i^j, using the Box-Muller method [44], and finally appropriate translation and scaling is performed to get a sample from the corresponding Gaussian distribution as $s_i^j = \mu_i^j + \sigma_i^j a_i^j$. These samples $\langle s_1^x, \ldots, s_n^z \rangle$ are used as displacement values for each of the n atoms in the new configuration.

For torsion angles, each entry in U_k is mapped from $[0,1]$ to the $[lower, upper]$ range by simple uniform scaling, and the desired configuration in Cartesian space is computed using the Denavit-Hartenberg transform [45].

It is possible that for both methods outlined above some sampled configurations might result in severe clashes between non-consecutive atoms. Those configurations with high number of clashes are discarded. Additionally, as bond angles and lengths are not maintained under the B-factor sampling model, we subject each sample configuration to a brief minimization round with Amber 12 [46] using the ff03 forcefield [47]. Finally, we prepare each sample for further calculations by protonating and assigning partial charges using PDB2PQR [48] in amber mode.

2.3.4 Efficient Low-Discrepancy Sampling

A major issue with the practicality of QMC is that it suffers from the curse of dimensionality. Specifically, if m samples

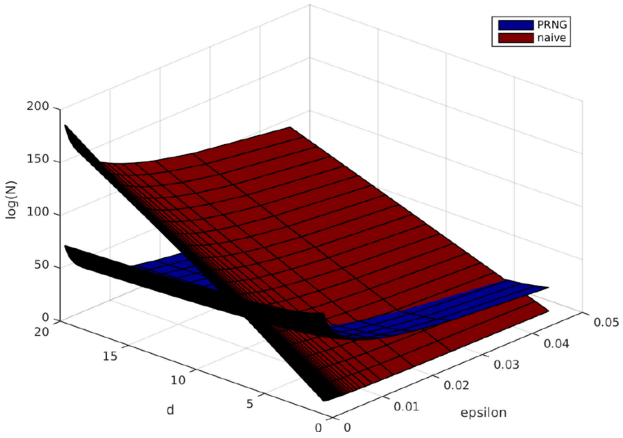


Fig. 2. Number of samples needed to maintain ϵ discrepancy. Note that the y-axis is log-scale, so the number of samples (N) required for the naive method grows exponentially in d (number of dimensions), whereas the pseudorandom method (PRNG) grows as a function of $\log d$.

are required to achieve a discrepancy ϵ in one dimension, then at least m^d samples will be necessary to achieve the same level of discrepancy in d dimensions, if the sampling is carried out naively (i.e., product of samples in each dimension). For the type of applications we are interested in (i.e., a required discrepancy of < 1 percent and dimension $\sim$1000), the number of samples would be prohibitively high.

The low-discrepancy product-space sampler developed by Bajaj et al. [49] reduces the number of samples significantly: from m^d to only $\left(\frac{d}{\epsilon}\right)^{O\left(\sqrt{\log\left(\frac{1}{\epsilon}\right)}\right)}$, where $m = \left(\frac{d}{\epsilon}\right)^{3+o(1)}$, while still ensuring the same discrepancy bounds. Note that this is polynomial in m and d. (See Fig. 2 for a summary of the number of samples required for different values of d and ϵ.) Furthermore, this method guarantees that for any given number of samples, the discrepancy of the sampled set is optimal. Hence, it can be applied iteratively.

The basic intuition behind the sampling strategy of Bajaj et al. is as follows. Low-discrepancy sampling in one dimension (e.g., the interval $[0, 1]$) is trivial: for ϵ discrepancy, sample at $m = 1/\epsilon$ equal points along the interval. In d dimensions, using this naive approach independently in each dimension requires m^d samples, and the discrepancy over the entire hyper-rectangle is highest along the diagonal: $\epsilon\sqrt{d}$. If, instead, the coordinates of a single sample are generated *dependently*, the same discrepancy can be achieved using much fewer samples. Deterministic sequences with these guarantees exist (such as the Sobol or Halton sequences [50]), and the work of Bajaj et al. improves upon these through randomization, thus reducing some of the bias observed in deterministic methods.

In our application, we have explored iterative sampling and considered the sampling to be sufficient when the approximation of the moments of the distribution of a QOI converge. We found that convergence was achieved with relatively few samples in practice. Please see Results for details.

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.

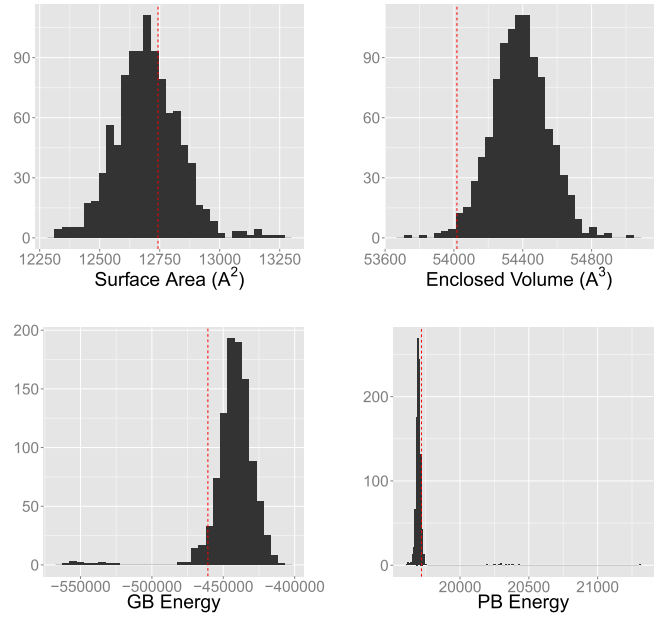


Fig. 3. Histogram of sampled QOIs for 1OPH:A. The red vertical line is the value of $\mathcal{F}$ computed using the original coordinates reported in the PDB.

3.1 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 [20]. We used this docking benchmark as we were interested in demonstrating how uncertainty in QOI reported on a single protein 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 sections we explore different aspects of this analysis.

3.1.2 *Uncertainty of Unperturbed Models*

Fig. 3 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 value and standard deviation

TABLE 1
Chernoff-Like Bounds for PDBID:1OPH

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
GB(A)	0.970	0.835	0.660	0.396	0.061	0.012
GB(B)	0.964	0.817	0.618	0.310	0.019	0.003
ΔGB(A)	0.986	0.929	0.863	0.725	0.364	0.099
ΔGB(B)	0.982	0.934	0.866	0.726	0.372	0.094
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

For each value of t, corresponding values of ϵ are calculated from 1,000 samples.

derived from the empirical distribution, are 0.33, −0.82, 1.37, and 0.25 for area, volume, GB, and PB, respectively. 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 1,000 low-discrepancy samples, and then computed the probability, ϵ, of a randomly sampled model having more than 0.1, 0.5, 1, 2, 5 and 10 percent error (t). Error is defined as $|x' - E[x]|/E[x]$, where x' is the value computed for a random model and $E[x]$ is the expected value over all samples.

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.

As can be seen in these tables, for most of the QOIs, the probability of incurring more than 5 percent error is negligible. 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.

3.2 Number of Samples Sufficient to Provide Statistically Accurate Certificates

The results reported in the previous two sections highlight the importance of UQ and also shows that mean values of QOI do not always correlate well with the those computed on the original molecules, so statistical bounds across a set of samples are needed. However, it is important to note that single statistics that deviate greatly from the mean ("rare" events that are actually plausible biological configurations) can have a great effect on these first and second moments. Therefore, it is important to ensure that the distributions generated through sampling accurately represent the underlying distributions, instead of just a poorly-sampled subset. While the number of samples required with our pseudo-random number generator is drastically lower than a naive exponential sampling method, it is still prohibitive to generate all possible samples. However, in all the simulations we ran, we have found that this theoretical bound overestimates the requirement and much fewer samples is sufficient.

TABLE 2
Average Chernoff-Like Bounds for 1,000 Samples of All Protein Chains

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
GB	0.963	0.817	0.648	0.391	0.100	0.029
ΔGB	0.963	0.818	0.651	0.394	0.103	0.030
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

In this section, we present a protocol for determining the number of pseudo-random samples needed to achieve robust certificates, or the minimal number of samples before the gain achieved through additional sampling is negligible.

For each QOI, we select a subset of r samples and calculated ϵ for given values of t (see Equation (1)) on this reduced dataset. We did the same for s random samples (where $s > r$), and computed the distance (L2 norm) between the two. If the distance was less than a given threshold, τ, then we determined we had reached saturation; otherwise, we increased the number of reduced samples, r (potentially also s), and repeated the above calculations.

As the full set of samples is not always available, we determined the saturation for both the full dataset (where $s = 1{,}000$) and a reduced, incremental comparison ($s = r + 10$, averaged over 50 trials). For the experiments in this section, we calculated ϵ for 6 different values of τ. Since the expected distance between any two random points in 6d space is 0.9689 (an analytic form for such distances has been derived in [51], and the precomputed values for several dimensions are available online [52]), we chose τ at 0.05, which is much lower than 0.9689, as our measure of convergence.

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. For instance, the Chernoff-bound calculated for on the incremental method for PB energy with 1OPH chain A (top table) reached saturation after 287 samples; with only 174 samples, we were able to achieve Chernoff-like bounds with at least 95 percent 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. Fig. 4 shows a plot of both metrics as the number of samples is increased.

Fig. 5 shows the number of samples needed before the relative error (when computed on the full dataset) is

TABLE 3
Number of Samples Necessary to Converge on Statistical Bounds

1OPH	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

1DFJ	E	I	ΔE	ΔI
area	265/157	471/295	421/240	337/199
vol	156/69	385/273	428/257	343/218
LJ	319/223	462/361	507/361	363/206
CP	235/138	652/401	607/509	353/214
GB	341/219	335/193	296/217	303/227
PB	360/235	546/375	567/430	338/225

The process has converged if differences between computed values of ϵ do not change (detailed description of convergence can be found in the text). The pair of entries in each cell refer to the two variants of convergence considered: incremental, where each value is compared with the value computed with 10 additional samples; and global, where the value is compared with that computed on all 1,000 samples. Data presented here correspond to PDBID:1OPH, chains A and B (top table) and PDBID:1DFJ, chains E and I (bottom).

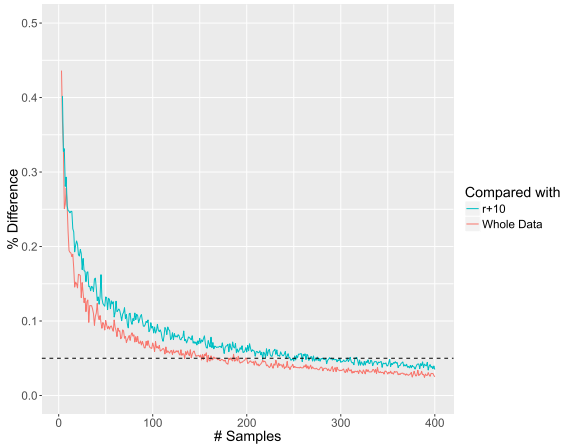


Fig. 4. Plot showing the rate of convergence for statistical certificates, computed for the calculation of free energy (MM-PBSA). For each number of samples, r, the Chernoff-like bounds were computed, and then compared with either those computed on the entire dataset (red line) or a partial dataset containing $r + 10$ 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.

negligible for several QOI, compared with the total B-factor of the protein (a statistic that attempts to incorporate the entire molecular uncertainty) and number of atoms. As can be seen in these figures, neither total B-factor nor number of atoms are capable of predicting the number of samples needed. Indeed, the number of samples required to predict accurate distributions of values is not linked to any extrinsic property (such as size of the protein, total B-factor, etc.), but is instead linked to intrinsic properties of the distribution and how often the "rare" events happen. Thus, if the rare events are more common (such as chain I of PDBID:1DFJ in Fig. 5), it is necessary to generate more samples. The incremental sampling approach provides an accurate procedure for computing confident probabilistic bounds by generating more samples for proteins that have higher intrinsic uncertainty.

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 localizing 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. Fig. 6A 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 (see also Fig. 10).

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 Fig. 6B can be very descriptive. It shows a single smooth

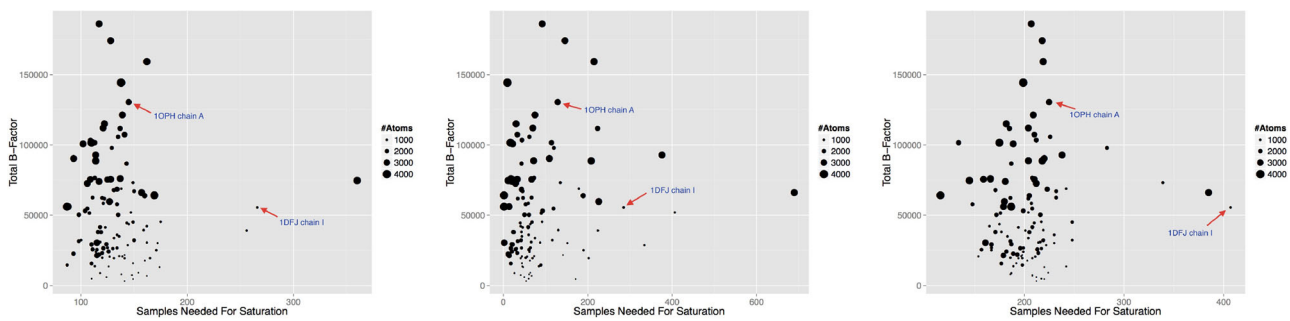


Fig. 5. *Convergence of sampling protocol across all samples.* Plot of total B-factor (a measure of both size and uncertainty) against number of samples needed before the relative error is negligible ($\tau = 0.05$). QOIs are exposed surface area (left), LJ potential (center), and PB energy (right).

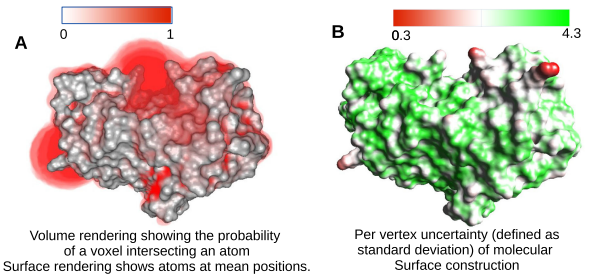


Fig. 6. *Visualization of molecular surface uncertainties.* (A) A volumetric map showing the likelihood of the voxel being occupied by an atom, computed using a sampling of the joint probability distribution of the atom positions. (B) Expected deviation of each point on the surface of a single model, 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.

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 show very low deviation, and only the narrow and dangling parts show high deviations. Comparing this with the rendering of B-factors (in Fig. 7A) 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.

Uncertainty in Energy Calculations. We use the same technique to show the local impact of compounded QOIs, or those using primary QOIs as input. 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, not only can we bound the bound the overall uncertainty of the final value of PB energy, but can also bound the uncertainties of intermediate PB potentials calculated at each voxel. We do this by defining the PB potential at each voxel as a separate QOI, and evaluate the QOI (potential at a given voxel) for each sample generated with the QMC sampling protocol. Hence, we derive a *distribution* for each voxel. The means and standard deviations of these distributions are rendered in Figs. 7C and 7B (respectively), showing the positive and negative potential regions, as well as regions that have varying potential. Comparing the input uncertainties (B-factors in Fig. 7A) to these third level propagated uncertainties

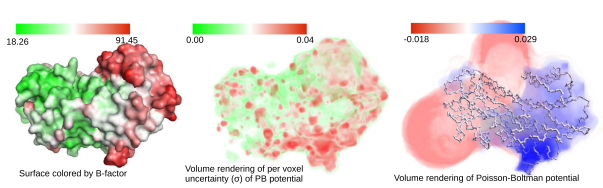


Fig. 7. *Visualization of molecular energy uncertainties.* (A) Simple mapping of B-factors to the surface of the protein. (B) Expected deviation of potential energy of a given voxel, w.r.t. all models sampled based on the joint distribution of the locations of the atoms. Green colored regions are expected 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.

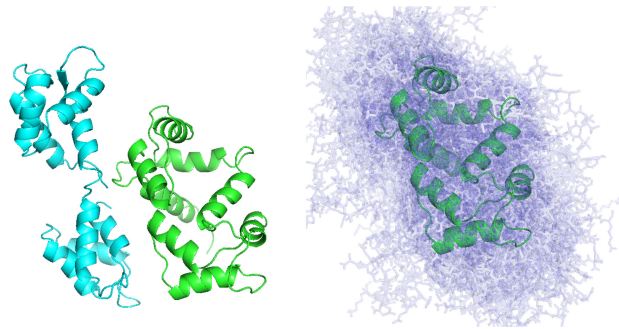


Fig. 8. *Large-scale conformational samples resulting from torsion angle sampling.* Samples of Calmodulin, which undergoes massive conformational shifts when bound with calcium. Left, the open and closed structures (corresponding to PDBID:1CFD and PDBID:1CKK, respectively) are colored blue and green (respectively). Right, torsional samples of closed conformation (transparent), showing large conformational shifts from input molecule.

shows that while in some regions the uncertainties had a cancellation effect, in other regions they were amplified.

3.4 Applications to Internal Angle Sampling

Throughout this paper, we have identified a protocol for showing the uncertainty in QOI that exists in the presence of atomic uncertainty. We have stressed that this same protocol could also be used to model uncertainty on QOI when the underlying molecules are sampled from their internal (torsional) angles. However, there is one major distinction.

For proteins, small changes in torsion angles can result in huge conformational shifts; therefore, the QMC sampling protocol defined in this paper does not produce *uncertain states of the same protein*, but *completely new configurations*. To test our QMC sampling protocol with torsional angles, we used Calmodulin, a protein that undergoes large conformational shifts when bound with calcium (see Fig. 8). We sampled from the closed structure of Calmodulin (PDBID:1CFD), which contains 148 residues, providing 296 degrees of freedom, and generated 1,000 samples according to the QMC protocol defined above.

Unlike with the atomistic sampling protocol, providing bounds for QOI like as surface area or volume (that are expected to change substantially with each new configuration) do not bound *uncertainty*, but instead deviations due to configurational shifts (see Fig. 9, where the surface area

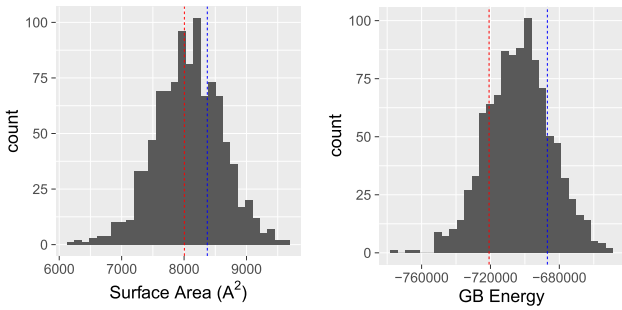


Fig. 9. Distribution of simple QOI (surface area, left, and energy (GBSA), right) for samples of closed Calmodulin conformation PDBID:1CKK. Red dotted line shows the QOI on the input closed conformation, and blue shows the open (PDBID:1CFD). Note that these histograms do not represent *structural* uncertainty but changes in QOI due to *conformational changes*. Compare with Fig. 3 where structural uncertainty for a different protein is bounded.

and energy vary much more than the surface area and energy of atomistic sampling, as found in Fig. 3). Hence, for internal angle sampling, we focus on uncertainties for QOIs which are computed over a collection of configurations instead. For example, define a QOI as the possible binding site on a protein (receptor). We define this as the probability, over possible configurations and transformations of the binding partner (ligand), that an atom (i.e., a point on the surface) of the receptor would be in contact (within a distance cut-off).

Given a calibrated scoring function $\mathcal{F}$ with bounded errors, specific configurations s_A and s_B of two molecules A and B, and a low dispersion sampling of the space of relative orientations $SE(3)$, we can compute a list of the top k ranked orientations of s_A with s_B (e.g., through protein-protein docking). Let the ith orientation be expressed using a transformation T_i which is applied to s_B (denoted $T_i(s_B)$). Now, for each atom a on molecule A, let $BS(a, s_B)$ be a random variable denoting the event that a is in contact with s_B upon binding; i.e., a is on the binding site. Now, we define the probability of $BS(a, s_B)$ as follows:

$$p_{BS}(a, s_B) = Pr[BS(a, s_B)] = \frac{1}{k}\sum_i cont(a, T_i(s_B)), \qquad (18)$$

where $cont(a, T_i(s_B))$ is 1 if at least one atom in $T_i(s_B)$ is within a cutoff distance from a, otherwise it is 0.

Equation (18) establishes the binding site probabilistically. Given an accurate docking tool and molecules in favorable configurations, the probabilities would be high for small contiguous regions of protein A, and low in other regions. On the other hand, almost equal probability across A would indicate poor docking prediction, and/or poor affinity between the molecules.

3.4.1 Inhibitor Selection Based on Binding Site Overlap

We can perform the samp procedure for protein-ligand docking, where the ligand is typically a small (non-protein) molecule. During ligand optimization for binding, one needs to sample the configuration for the ligand and then, for each configuration, apply docking to predict the best ranked orientations (of the sample configuration). In such cases, we augment the definition of the probability of an atom being on the binding site by summing over all configurations. In other words, let R be the receptor and L be the ligand, and let $r \in R$ be an atom on R, then

$$p_{BS}(r, \mathcal{C}^L) = \frac{1}{kN}\sum_{s_L \in \mathcal{C}^L_N}\sum_i^k (cont(r, T_i(s_L))), \qquad (19)$$

where $\mathcal{C}^L$ represents the configurational space of the ligand, L, and $\mathcal{C}^L_N$ is the N discrete samples taken from this space.

Given this probabilistic estimate of the binding site, we define a new scoring function to rank the ligand configuration+orientations. For a given orientation T_i of a given configuration s_L, we define its score as follows:

$$bindScore(s_L, T_i) = \sum_{r \in R} p_{BS}(r, \mathcal{C}^L) \cdot cont(r, T_i(s_L)). \qquad (20)$$

Hence, configurations of the ligand which can bind at highly probably binding sites are rewarded.

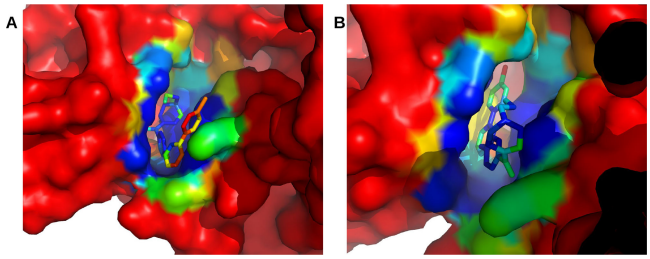


Fig. 10. *Comparing different ligand conformations using probabilistic binding site on the same protein.* The smooth surfaces in the figures show the probabilistic binding site on the kB kinase β (PDBID:3RZF) for binding with an inhibitor ligand, where red is not likely and blue is very likely. The ligand atoms are colored according to the average probability of receptor atoms within a given cutoff (4Å). Hence, a ligand configuration which has a high proportion of blue and low amount of red/orange is a better configuration (model on the right).

A visualization of Equation (20) is shown in Fig. 10. For this particular example, we obtained a correct configuration of a ligand that inhibits the κB kinase β, from the protein data bank model 3RZF, and a wrong configuration of the same ligand from the model 3QAD. We started with the wrong configuration, generated 1,000 samples, and docked each of them using Autodock [32] to the kinase. Fig. 10 shows two ligand conformations. The one on the left has a poor RMSD (3.24) with respect to the known correct configuration (and orientation) of 3RZF, and the one on the right has a favorable RMSD (1.069). Interestingly, both of them received the same score, -8.2, from AutoDock. However, the left one had a $bindScore$ of 55,213, and the right one had 55,667, showing that $bindScore$ correctly identifies the correct configuration.

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 is 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 percent) of having more than 5 percent error in total energy calculation (PBSA) purely due to 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). 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. Currently, we are working on providing bounds for binding free energy of large-scale conformational shifts through torsion angle sampling.

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.

Acknowledgments

This research was supported in part by a grant from NIH (R01-GM117594) and NIH (R41-GM116300).

References

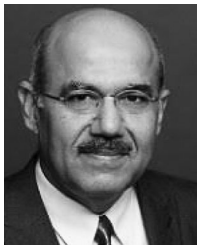
[1] M. Connolly, "Analytical molecular surface calculation," *J. Appl. Crystallography*, vol. 16, pp. 548–558, 1983.
[2] C. Bajaj, V. Pascucci, A. Shamir, R. Holt, and A. Netravali, "Dynamic maintenance and visualization of molecular surfaces," *Discrete Appl. Math.*, vol. 127, pp. 23–51, 2003.
[3] D. Eisenberg and A. Mclachlan, "Solvation energy in protein folding and binding," *Nature*, vol. 319, pp. 199–203, 1986.
[4] M. Nina, D. Beglov, and B. Roux, "Atomic radii for continuum electrostatics calculations based on molecular dynamics free energy simulations," *J. Phys. Chemistry B*, vol. 101, pp. 5239–5248, 1997.
[5] D. Bashford and D. A. Case, "Generalized born models of macromolecular solvation effects," *Annu. Rev. Phys. Chemistry*, vol. 51, pp. 129–152, 2000.
[6] C. Bajaj and W. Zhao, "Fast molecular solvation energetics and forces computation," *SIAM J. Sci. Comput.*, vol. 31, no. 6, pp. 4524–4552, 2010.
[7] R. Chowdhury, et al., "Protein-protein docking with F^2Dock 2.0 and $GB\text{-}rerank$," *Biophys. J.*, vol. 8, no. 3, pp. 1–19, 2013.
[8] C. Bajaj and V. Siddavanahalli, "F2Dock: A fast and fourier based error-bounded approach to protein-protein docking," Univ. Texas at Austin, Austin, TX 78712, USA, CS Tech. rep. TR-06-57, Nov. 2006.
[9] F. Richards, "Areas, volumes, packing, and protein structure," *Annu. Rev. Biophys. Bioeng.*, vol. 6, pp. 151–176, 1977.
[10] 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.*, 1997, pp. 217–228.
[11] M. Feig and C. Brooks, "Recent advances in the development and application implicit solvent models in biomolecule simulations," *Current Opinion Structural Biol.*, vol. 14, 2004, Art. no. 217.
[12] A. Onufriev, D. Bashford, and D. Case, "Modification of the generalized born model suitable for macromolecules," *J. Phys. Chemistry B*, vol. 104, pp. 3712–3720, 2000.
[13] Schrödinger, LLC, "The PyMOL molecular graphics system, version 1.3r1," Aug. 2010, PyMOL The PyMOL Molecular Graphics System, Version 1.3, Schrödinger, LLC, pymol.org
[14] Y. Lei and R. R. Mettu, "A confidence measure for model fitting with x-ray crystallography data," in *Proc. Int. Conf. Bioinf. Comput. Biol. Biomed. Inform.*, 2013, Art. no. 489.
[15] W. Rieping, M. Habeck, and M. Nilges, "Inferential structure determination," *Sci.*, vol. 309, no. 5732, pp. 303–306, 2005.
[16] M. Habeck, M. Nilges, and W. Rieping, "Bayesian inference applied to macromolecular structure determination," *Phys. Rev. E*, vol. 72, no. 3, 2005, Art. no. 031912.
[17] H. Lei, X. Yang, B. Zheng, G. Lin, and N. A. Baker, "Quantifying the influence of conformational uncertainty in biomolecular solvation," arXiv:1408.5629, 2014, arxiv.org
[18] K. Azuma, "Weighted sums of certain dependent random variables," *Tokuku Math. J.*, vol. 19, pp. 357–367, 1967.
[19] C. McDiarmid, "On the method of bounded differences," *Surveys Combinatorics*, vol. 141, no. 141, pp. 148–188, 1989.
[20] H. Hwang, T. Vreven, J. Janin, and Z. Weng, "Protein–protein docking benchmark version 4.0," *Proteins: Struct. Function Bioinf.*, vol. 78, no. 15, pp. 3111–3114, 2010.
[21] B. T. Burnley, P. V. Afonine, P. D. Adams, and P. Gros, "Modelling dynamics in protein crystal structures by ensemble refinement," *Elife*, vol. 1, 2012, Art. no. e00311.
[22] W. G. Touw and G. Vriend, "BDB: Databank of PDB files with consistent B-factors," *Protein Eng. Des. Sel.*, vol. 27, no. 11, pp. 457–462, Nov. 2014.
[23] P. T. Lang, et al., "Automated electron-density sampling reveals widespread conformational polymorphism in proteins," *Protein Sci.*, vol. 19, no. 7, pp. 1420–1431, 2010.
[24] A. Kuzmanic, N. S. Pannu, and B. Zagrovic, "X-ray refinement significantly underestimates the level of microscopic heterogeneity in biomolecular crystals," *Nature Commun.*, vol. 5, 2014, Art. no. 3220.
[25] A. Kuzmanic and B. Zagrovic, "Determination of ensemble-average pairwise root mean-square deviation from experimental B-factors," *Biophys. J.*, vol. 98, no. 5, pp. 861–871, 2010.
[26] E. F. Pettersen, et al., "UCSF Chimera–A visualization system for exploratory research and analysis," *J. Comput. Chemistry*, vol. 25, no. 13, pp. 1605–1612, 2004.
[27] P. Emsley, B. Lohkamp, W. G. Scott, and K. Cowtan, "Features and development of coot," *Acta Crystallographica Section D - Biological Crystallography*, vol. 66, pp. 486–501, 2010.
[28] R. M. Hanson, "Jmol—A paradigm shift in crystallographic visualization," *J. Appl. Crystallography*, vol. 43, pp. 1250–1260, 2010.
[29] D. A. Case, et al., "The amber biomolecular simulation programs," *J. Comput. Chemistry*, vol. 26, no. 16, pp. 1668–1688, 2005.
[30] 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.*, vol. 33, no. 2, pp. 826–848, 2011.
[31] N. Eswar, et al., "Comparative protein structure modeling using MODELLER," *Current Protocols Protein Sci.*, Chapter 2, 2007, Art. no. Unit 2.9.
[32] O. Trott and A. J. Olson, "AutoDock Vina," *J. Comput. Chemistry*, vol. 31, pp. 445–461, 2010.
[33] C. Bajaj, P. Djeu, V. Siddavanahalli, and A. Thane, "TexMol: Interactive visual exploration of large flexible multi-component molecular complexes," in *Proc. IEEE Vis. Conf.*, 2004, pp. 243–250.
[34] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Ann. Math. Statist.*, vol. 23, no. 4, pp. 493–507, 1952.
[35] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statistical Assoc.*, vol. 58, no. 301, pp. 13–30, 1963.
[36] H. Niederreiter, "Quasi-Monte Carlo methods," *Encyclopedia Quantitative Finance*, vol. 24, no. 1, pp. 55–61, 1990.
[37] F. James, J. Hoogland, and R. Kleiss, "Quasi-Monte Carlo, discrepancies and error estimates," *Methods*, 1996, Art. no. 9. [Online]. Available: http://arxiv.org/abs/physics/9611010
[38] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1992.
[39] T. R. Schneider, "What can we learn from anisotropic temperature factors," in *Proc. CCP4 Study Weekend*, 1996, pp. 133–144.
[40] K. Trueblood, et al., "Atomic dispacement parameter nomenclature. Report of a subcommittee on atomic displacement parameter nomenclature," *Acta Crystallographica Section A: Found. Crystallography*, vol. 52, no. 5, pp. 770–781, 1996.

[41] A. T. Briinger, "Free R value: A novel statistical quantity for assessing the accuracy of crystal structures," *Nature*, vol. 355, pp. 472–475, 1992.
[42] D. Cruickshank, "Remarks about protein structure precision," *Acta Crystallographica Section D: Biol. Crystallography*, vol. 55, no. 3, pp. 583–601, 1999.
[43] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, "Stereochemistry of polypeptide chain configurations," *J. Mol. Biol.*, vol. 7, pp. 95–99, 1963.
[44] G. E. Box and M. E. Muller, "A note on the generation of random normal deviates," *Ann. Math. Statist.*, vol. 29, pp. 610–611, 1958.
[45] R. S. Hartenberg and J. Denavit, "A kinematic notation for lower-pair mechanisms based on matrices," *Trans. ASME J. Appl. Mech.*, vol. 22, pp. 215–221, 1955.
[46] D. A. Case, et al., "Amber 12," Univ. California, San Francisco, 2012. [Online]. Available: http://ambermd.org/
[47] Y. Duan, et al., "A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations," *J. Comput. Chemistry*, vol. 24, no. 16, pp. 1999–2012, 2003.
[48] T. Dolinsky, J. Nielsen, J. McCammon, and N. Baker, "PDB2PQR: An automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations," *Nucleic Acids Res.*, vol. 32, pp. 665–667, 2004.
[49] C. Bajaj, A. Bhowmick, E. Chattopadhyay, and D. Zuckerman, "On low discrepancy samplings in product spaces of motion groups," arXiv:1411.7753, 2014, arxiv.org
[50] L. Kuipers and H. Niederreiter, *Uniform Distribution of Sequences*. New York, NY, USA: Wiley, 1974.
[51] J. Philip, *The Probability Distribution of the Distance Between Two Random Points in a Box*. Stockholm, Sweden: KTH mathematics, Royal Institute of Technology, 2007.
[52] N. J. Sloane, "Online encyclopedia of integer sequences (OEIS)." [Online]. Available: 2014, http://oeis.org/A103986.

Muhibur Rasheed received the BSc and MSc degrees in computer science from the Bangladesh University of Engineering and Technology, and the PhD degree from the Computer Science Department, University of Texas at Austin, in 2014. His research interest include geometric modeling and simulations of complex physical systems, especially in molecular biology. He has developed geometric data structures for fast neighborhood searches in arbitrary dimensions, designed algorithms for efficient computations of molecular surfaces and properties, and prediction of molecular structures at atomic resolutions. He is currently working on algorithms for geometry discretization at CD-Adapco.

Nathan Clement received the BS degree in computer science with a bioinformatics emphasis from Brigham Young University, in 2010, and the MS degree in computer science from the University of Texas at Austin, in 2014. He is currently working toward the PhD degree in computer science with the University of Texas at Austin. Past research projects have included design and implementation of massively parallel algorithms for next-generation DNA sequencing data, including hash-based search and metric space indexing. His dissertation is focused on developing programs for generating and testing low-discrepancy sequences in high-dimensional spaces, and providing suitable probabilistic models for biological systems.

Abhishek Bhowmick received the undergraduate degree from the Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur, in 2010. He received the PhD degree supported by the Homer Lindsay Bruce Endowed Graduate Fellowship from the Department of Computer Science, University of Texas at Austin. His research interests lie in the study of analytic and algebraic aspects of theoretical computer science. More specifically, he is interested in the paradigm of structure versus randomness of polynomials and its implications in coding theory, algebraic geometry, and correlation bounds. Since January 2017, he has been a research fellow in the Simons Institute for the Theory of Computing at UC Berkeley.



Chandrajit L. Bajaj received the bachelor of technology degree in electrical engineering from IIT Delhi, in 1980 and the PhD degree in computer science from Cornell University, in 1984. He is presently the director of the Computational Visualization Center, the University of Texas at Austin. He has been a professor of computer sciences with the University of Texas at Austin since 1997 and also has held the Computer Applied Math (CAM) Chair of Visualization since 1997. His research pursuits are focused on the algorithmic and computational mathematics underpinnings of imaging and geometry data sciences, computer graphics, bio-informatics and visualization with applications stemming from bio-medical engineering and physical and chemical sciences. He is a fellow of the American Association for the Advancement of Science (AAAS), fellow of the Institute of Electrical and Electronics Engineers (IEEE), fellow of the Society of Industrial and Applied Mathematics (SIAM), and also fellow of the Association of Computing Machinery (also known as ACM), which is the world's largest education and scientific computing society. He has won the University of Texas Faculty research award, the Dean Research Assignment award, the Indian Institute of Technology, Delhi's Distinguished Alumnus Award, and also thrice won the University of Texas, Institute of Computational Engineering and Sciences, Moncreif Grand Challenge research award.