Quantified Uncertainty of Flexible Protein-Protein Docking Algorithms

Nathan L. Clement
Department of Computer Science, University of Texas at Austin, USA
nclement@cs.utexas.edu

Abstract

The strength or weakness of an algorithm is ultimately governed by the confidence of its result. When the domain of the problem is large (e.g. traversal of a high-dimensional space), an exact solution often cannot be obtained, so approximations must be made. These approximations often lead to a reported quantity of interest (QOI) which varies between runs, decreasing the confidence of any single run. When the algorithm further computes this QOI based on uncertain or noisy data, the variability (or lack of confidence) of the QOI increases. Unbounded, these two sources of uncertainty (algorithmic approximations and uncertainty in input data) can result in a reported statistic that has low correlation with ground truth.

In molecular biology applications, this is especially applicable, as the search space is generally large and observations are often noisy. This research applies *uncertainty quantification* techniques to the difficult protein-protein docking problem, where uncertainties arise from the explicit conversion from continuous to discrete space for protein representation (introducing some uncertainty in the input data), as well as discrete sampling of the conformations. It describes the variability that exists in existing software, and then provides a method for computing *probabilistic certificates* in the form of Chernoff-like bounds. Finally, this paper leverages these probabilistic certificates to accurately bound the uncertainty in docking from two docking algorithms, providing a QOI that is both robust and statistically meaningful.

2012 ACM Subject Classification Applied computing → Molecular structural biology; Mathematics of computing → Hypothesis testing and confidence interval computation; Computing methodologies → Uncertainty quantification

Keywords and phrases protein-protein docking, uncertainty quantification, protein flexibility, low-discrepancy sampling, high-dimensional sampling

Digital Object Identifier 10.4230/LIPIcs.WABI.2019.3

Related Version A full version of the paper is available at `http://arxiv.org/abs/1906.10253`.

Acknowledgements I would like to thank all those who have supported and helped advise on this work, for their valuable feedback and suggestions for improvement.

1 Introduction

Predicting the bound conformation of two proteins (protein-protein docking) has many applications in medicine and biology [26, 19]. The simpler form of this problem is the so-called "bound-bound" case, where the 3-dimensional coordinates of the *in situ* protein complex is resolved (via e.g. X-ray crystallography, NMR, etc.), and atoms corresponding to individual proteins are then extracted from the complex. The more difficult version is the "unbound-unbound" case, where each protein in the pair is imaged in its separate native state, and the algorithm must predict the correct *in situ* bound complex [18]. Importantly, the final quantity of interest (QOI) in many cases is the change in binding free energy: protein complexes with a high change in free energy are more likely to be found as a bound complex, and are likely good targets for drug discovery pathways. The difficulty of the unbound-unbound case then arises from the inherent flexibility of proteins: large-scale movements

19th International Workshop on Algorithms in Bioinformatics (WABI 2019).
Editors: Katharina T. Huber and Dan Gusfield; Article No. 3; pp. 3:1–3:12
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

may occur along the pathway from a closed conformation (unbound) to an open (bound) one, or visa versa. If docking is performed on only the unbound complexes, the final delta energy could be completely misleading. (To aid in discussion, here and through the paper we will refer to one protein, typically the bigger, as the *receptor*, and the other as the *ligand*.)

A subsequent difficulty of the unbound-unbound docking problem is that computational approaches must search two high-dimensional spaces. The first is that of possible protein structures, a naive description of which is $\mathbb{R}^{3m} \times \mathbb{R}^{3n}$, where m and n are the number of atoms in the ligand and receptor, respectively. The second is the space of possible docked conformations. In rigid-body docking (each protein is static), this is the 6-dimensional real space of 3 rotational + 3 translational degrees of freedom, $SE(3) = SO(3) \times \mathbb{R}^3$ [1, 25]. Without approximation, searching this high-dimensional space is computationally intractable. To achieve meaningful results, successful algorithms must employ some sort of simplification.

One of the biggest issues that arises from these simplifications is *uncertainty propagation*. A computational representation of a protein is, by nature, an approximation (discrete representation of a continuous space). Computing a simple statistic, or *quantity of interest* (QOI), on these representations is then by nature uncertain [27]. Algorithmic approximations (due to randomness or variations in the inputs) in one stage of a protein docking pipeline lead to uncertainty in the input for the next stage. If these uncertainties are not *quantified* at each stage, the uncertainties propagate to future levels of the pipeline, leading to a result or QOI that is unbounded, and may contain little valuable information.

This paper provides a framework for bounding the uncertainty of protein-protein docking. For a docking procedure where the QOI, $f(\mathbf{X})$, is some complicated function or optimization functional involving noisy data $\mathbf{X}$, we seek to provide a *probabilistic certificate* as a function of parameter t that the computed value $f(\mathbf{X})$ is not more that t away from the true value, with high probability. This certificate is expressed as a Chernoff-Hoeffding like bound [8]:

$$\Pr\left[|f(X) - E[f]| > t\right] \leq \epsilon, \tag{1}$$

where $E[f]$ is the expectation of f, computed over all permutations of X. The primary QOI we are interested in bounding is the change in Gibbs free energy, or ΔG, as this is the metric most useful for real-world experiments. However, we also consider the interface RMSD (iRMSD), which is defined as the RMSD between $C\alpha$ atoms on the interface of the bound pair.

Instead of providing a new docking algorithm as a solution to bounding the above certificate, this research instead considers the docking algorithm, $f(\mathbf{X})$, as a *black box*, exploring the landscape of possible structures, $X \in \mathbf{X}$, as inputs to f and computing the certificates from the output. This then provides a framework by which any two algorithms can be compared, and by which conclusive results can be reported.

In this work, we expand upon our previous research [27, 10] in the following manner. First, the model used in the previous research was simplistic, and, while useful for modeling small uncertainties, does not provide insight into uncertainty of large-scale protein movement. Second, we consider the impact of this conformational uncertainty to provide certificates for black-box docking functions. This second contribution can be used when trying to interpolate results of a given docking algorithm to biological equivalents.

The only known research that applies uncertainty quantification to protein-protein docking is a recent preprint by Cau and Shen [5]. The authors use Bayesian active learning to explore protein-protein docking samples using a black-box energy function. Once the energy landscape has been sufficiently sampled, they provide posterior distributions of the desired QOI, which enables computing confidence intervals for each model. The major differences between this work and our own work is 1) the treatment of the entire *docking algorithm* as a black box (instead of just the energy function), and 2) the use of a hierarchical model convolved with a von Mises distribution to generate samples local to the unbound input.

The paper is organized as follows. First, we provide the theoretical and technical details of our approach, including probabilistic certificates through effective sampling, protein representation, sampling protocol, and benchmark dataset. Second, we show that the protein sampling protocol used improves upon the results of both rigid-body and flexible docking metrics, computing the probabilistic certificates for the change in Gibbs free energy for sets of docked proteins. Finally, we discuss the importance of these results both in terms of UQ for docking algorithms and biological relevance.

2 Materials and methods

2.1 Computing Chernoff-like bounds

Our primary motivation in this work is to compute a *probabilistic certificate* to bound the uncertainty in a computed statistics. We are most concerned with providing the Chernoff-Hoeffding like bound expressed in Equation 1, which provides a probabilistic guarantee for the moments of a QOI computed on noisy data.

We can provide a theoretic bound for the uncertainty by using the McDiarmid inequality, defined in [22] and extended to support summations of decaying kernels such as the Leonnard-Jones potential in [27]. Let (X_i) be independent random variables with discrete space A_i, let $f : \Pi_i A_i \to \mathbb{R}$, and let $|f(x_1, \ldots, x_k, \ldots, x_n) - f(x_1, \ldots, x'_k, \ldots, x_n)| \leq c_k$, or c_k is the degree of change influenced on f over all variations of x_k. Then, for any $t > 0$:

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

Thus, to provide theoretic bounds, all that is required is to determine the value of c_k for each x_k. However, computing c_k analytically may be difficult, and even if it were possible, theoretical bounds these often overestimate the error. An alternate approach is then to empirically compute these certificates using quasi-Monte Carlo (QMC) methods [24, 16]. Assuming the distribution of (X_i) is known, we sample this space and evaluate f at each sample. This leads to an estimate of the distribution of f over the joint space of all A_i, which provides sufficient data to compute certificates on the uncertainty, as defined in Equation 1.

Correctness of this approach relies on the correctness of the QMC methods and the description of the joint sampling space. For this reason, we will spend the next section describing our protein representation and the corresponding sampling space. In the Results and discussion section we will show that our sampling space is accurate (e.g. a good representation of the distribution of (X_i)), and thus the provided certificates are also sound.

2.2 Protein representations

The base structure of a protein is a linear chain of *amino acids* (also called “residues”). Each amino acid consists of a set of atoms, and all the atoms connected by covalent bonds into a single 3-dimensional structure. Such atoms divide into two groups: *backbone* atoms: two carbons, one nitrogen, and one oxygen; and zero or more *side-chain* atoms. The carbon connecting the backbone to the side-chain atoms is called the $C\alpha$ atom, and the first side-chain carbon (if it exists) is called the $C\beta$ atom (see Figure 1). The *native representation* of a protein is thus a graph in 3-dimensional Cartesian space, where each node of the graph

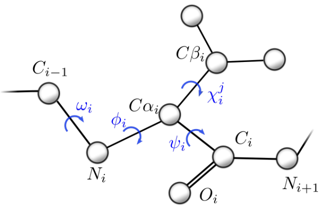

Figure 1 Torsion angles for a protein chain. The backbone atoms are labeled C_i, O_i, $C\alpha_i$, and N_i. For a constrained internal coordinate representation, only the ψ_i, ϕ_i, and potentially χ_i torsion angles are considered (ω_i is fixed at 180°).

represents atoms and edges represent bonds. The position of each node/atom is represented by a vector in $\mathbb{R}^3$, requiring three parameters for each atom. If $\hat{t}$ is the average number of side-chain atoms per residue for a given protein, then this representation requires a total of $n = 3(\hat{t} + 4)N$ parameters (3 degrees of freedom each for the $\hat{t}$ side-chain and 4 backbone atoms), or degrees of freedom (DOFs), for a protein with N amino acids.

An alternate representation of proteins, employed by most sampling protocols (e.g. [23, 12] and others) is the *internal coordinates* representation. Under this representation, the position of each atom is only defined *in relation* to the atoms around it, and the free variables are bond angles, bond lengths, and torsion angles (the degree of "twist" defined by 2 planes or 4 atoms, see Figure 1). This does not immediately reduce the total degrees of freedom (since in general, each *atom* needs to be described by bond angles, bond lengths, and torsion angles); however, if small-scale atomic vibrations are ignored, then bond lengths and angles can be approximated as constant, leaving the only DOFs as the ψ, ϕ, and χ_i torsion angles (the ω torsion angle on the backbone is held at $\sim 180°$ by the sp^2 partial double bond [4]). If $\hat{k}$ is the average number of χ_i angles for a given residue (k varies from 0 to 5 in the standard 20 amino acids), then the number of DOFs for this representation for a protein with N amino acids is $m = (\hat{k} + 2)N$. Since in most cases $\hat{k} + 2 \ll 3(\hat{t} + 4)$, this *constrained* internal coordinate method allows for a lower-dimensional specification of the protein conformational space without a loss in representation [13].

2.3 Hierarchical domain decomposition and motion graph

Roughly speaking, proteins decompose into rigid and flexible parts. Rigid contiguous parts are called domains, which exhibit little movement in several conformations. In turn, flexible parts, also known as hinges, interconnect domains. These flexible parts show three types of motion: shearing or gliding (i.e. a lateral movement along domain interfaces), bending (i.e. an angular movement between axes of two connected domains), and twisting (i.e. a rotational movement around the longitudinal axis of a domain).

When representing large-scale protein motion, we are primarily interested in hinges, or flexible regions connecting large mostly-rigid bodies or domains. However, since there may be multiple levels of motion, we use a *hierarchical representation* of the constrained internal angles representation of the protein. The hierarchical representation is not a recursive subdivision of the protein, but rather a description of (possibly overlapping) protein motions. This allows us to represent motions at one level that consist of atoms from different domains in the previous level.

To obtain this hierarchical domain decomposition for a given protein, we model the protein as a $C\alpha$ (one node per residue) GNM (Gaussian network model), and compute the NMA (normal modes analysis) decomposition of the protein (in this work, we use the

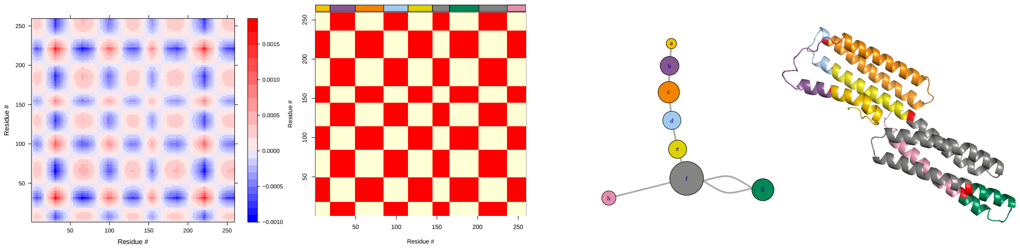

Figure 2 The NMA decomposition of 1RKE receptor, for the second non-trivial mode. From left to right: the cross-correlation fluctuation matrix, $[F]_2$; the sign of entries of $[F]_2$, with short domains removed; the domain graph representation of the protein, where the size of each node represents the number of residues in that domain; and the domain graph representation mapped onto the 3d structure of the protein, colored according to domain with hinge residues colored red. Hinges that are also flexible connectors separate all domains but f (gray) and g (green), which are connected by segments (hinges) that would not form a cut in the domain graph representation.

implementation from the R Bio3D package [14]). Each of the k modes represents a separate direction of motion, from large-scale motions (the smallest eigenvalues) to the high-frequency vibrations of hydrogen atoms. Each mode corresponds to a different level in our hierarchical representation; that is, each hierarchical level corresponds to a distinct rigidity threshold.

Hinges are obtained in a similar fashion to that demonstrated by HingeProt[11], as follows. For each mode, i, we compute the mean square fluctuation matrix as follows:

$$[F]_i = \frac{3k_BT}{\gamma}\lambda_i^{-1}\mathbf{u}_i\mathbf{u}_i^{\mathsf{T}}, \tag{2}$$

where λ_i and $\mathbf{u}_i$ is the eigenvalue and eigenvector of mode i, and k_B, T, and γ are the Boltzmann constant, temperature, and uniform force constant, respectively. Regions of this matrix with the same sign form the rigid domains, and individual residues where the sign changes (from positive to negative) become hinges. For practical purposes, we collapse domains with only a small number of residues.

The final stage at a given level is to construct a domain graph representation, where nodes in the graph represent rigid domains and edges in the graph exist wherever two domains are in contact with each other, i.e. any atom from one rigid domain is within r_c of any atom from another domain (see Figure 2 for the decomposition a single level in the hierarchy). From this graph, we categorize each hinge as *flexible connector* if the removal of the hinge would form a cut of the domain graph representation, i.e. its removal would result in two disjoint subgraphs.

Once we have obtained the domain graph representation of the protein for each of the k NMA modes, we construct a multi-graph of the domain hierarchy for the entire protein [2]. At the top level of the hierarchy are the hinges and domains computed by the first non-trivial mode (i.e. with the smallest eigenvalue), representing more broad, global motions. The next level of the hierarchy is represented by the second smallest eigenvalue, and so on until all k modes have been used. We also assign a weight, w_k, to all hinges at level k of the hierarchy, arising from Equation 2:

$$w_k = 3k_BT\lambda_k^{-1}. \tag{3}$$

The final dimension of the product space of sampling is then $K_R + K_L$, where K_R (K_L) is the number of hinges from all k levels for the receptor (ligand), creating a product space of $SO(3)^{K_R+K_L}$ ($SO(3)$ is the special orthogonal group of rotations about a fixed axis). For

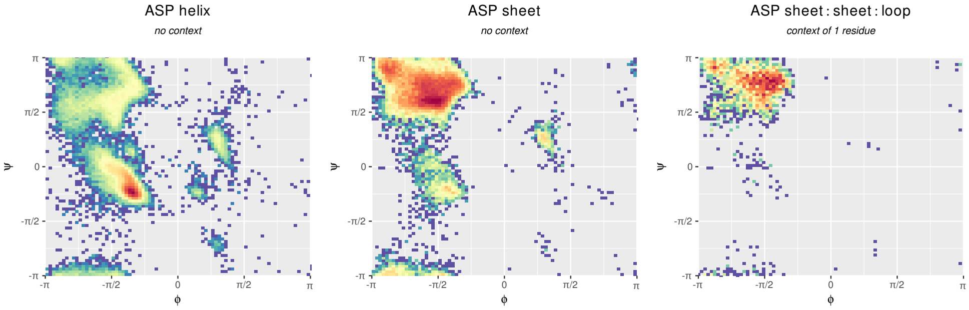

Figure 3 Ramachandran distributions for aspartic acid under different parameterizations. Left: ss_i is a helix; middle: ss_i is a loop; right: ss_{i-1}, ss_i, and ss_{i+1} are respectively sheet-sheet-loop. Note that the distributions are more tightly clustered with the gain of additional context.

the dataset used in this paper, the value of $K_R + K_L$ range from 21 hinge residues (3FN1) to 70 (1BKD). It is well known that generating a small number of good (i.e. low discrepancy) samples is difficult in high dimensions, so to overcome this issue, we use the low-discrepancy sampling protocol developed by [3] when generating samples.

2.4 Sampling protocol and Ramachandran distributions of amino acids

Based on the hierarchical protein decomposition described above, we now describe how to obtain a set of representative samples of the protein. Even with a good low-discrepancy sampling, this high-dimensional product space still requires a large number of samples to completely cover the product space. However, most of these samples will lead to physically impossible protein structures: clashes between nearby atoms, steric strain, or even a protein that is no longer biologically active. We would like to reduce the sampling space for a given torsion angle from all of $SO(3)$ to only the relevant, low-energy regions.

To establish a set of generic neighbor-dependent Ramachandran probability distribution, we compute the torsion angles from a set of ~15k high-quality, non-homologous protein structures obtained from the Pisces server [29]. From this set, we generate discrete probability distributions for each backbone torsion angle pair, conditioned on the amino acid type and secondary structure type of the previous and following residues. In other words,

$$Prob_N(\phi, \psi, i) = \Pr(\phi, \psi | ss_{i-1}, ss_i, ss_{i+1}, aa_i), \quad (4)$$

where ss_i and aa_i are the secondary structure and amino acid types of residue i, respectively, and ϕ and ψ are the backbone torsion angles (see Figure 1). Figure 3 shows the conditional distributions for aspartic acid.

To generate samples of a given protein, we would like to draw samples for each flexible residue from the neighbor-dependent Ramachandran distributions. However, we also recognize that the input protein has important structural elements that should be preserved. For this reason, we convolve the discrete Ramachandran distribution with a bivariate von Mises distribution (the two-dimensional variant of the approximately-Gaussian distribution on a unit circle, e.g. $[-\pi, \pi)^2$ [21]), centered at the given backbone torsion angle. The cosine variant of the bivariate von Mises distribution is given as follows:

$$\Pr(\phi, \psi) = Z_c(\kappa_1, \kappa_2, \kappa_3) \exp(\kappa_1 \cos(\phi - \mu) + \kappa_2 \cos(\psi - \nu) + \kappa_3 \cos(\phi - \mu - \psi + \nu)), \quad (5)$$

where μ and ν describe the mean for ϕ and ψ, κ_1 and κ_2 describe their concentration, and κ_3 describes their correlation. If κ_3 is zero and $\kappa_1 = \kappa_2 = \sigma$, then σ can be used to increase or decrease the amount of bias the input structure has on the Ramachandran distributions. Lower values of σ (lower concentration) bias more toward the general Ramachandran distributions, while higher values of σ bias more towards the input protein structure.

The final probability of a given (ϕ, ψ) pair at position i, $Prob(i, \phi, \psi)$, is the convolution of the neighbor-dependent Ramachandran distribution with the specific von-Mises distribution:

$$Prob\left(\phi, \psi, i\right) \propto \Pr\left(\phi, \psi | ss_{i-1}, ss_i, ss_{i+1}, aa_i\right) * \exp\left[\sigma \cos\left(\phi - \hat{\phi}_i\right) + \sigma \cos\left(\psi - \hat{\phi}_i\right)\right], \quad (6)$$

where $\hat{\phi}_i$ and $\hat{\psi}_i$ are the values of ϕ and ψ for residue i in the input protein.

With the internal angles representation and hierarchical decomposition of the protein as input, we perform the following importance sampling protocol on each level, l:

1. For each hinge at level l, $h_j^{(l)}$, let i be the index of the residue corresponding to this hinge.
   a. Generate the pair $(\hat{\phi}, \hat{\psi})$, drawn from the von Mises-convolved neighbor-dependent Ramachandran distributions
   b. Let ρ_l be the probability of a given hinge residue changing, arising from w_l in Equation 3: $\rho_l = \min(1, w_l)$
   c. If $h_j^{(l)}$ is a cut or no other non-cut hinges have been sampled at level l, set (ϕ_i, ψ_i) to $(\hat{\phi}, \hat{\psi})$ with probability ρ_l; otherwise, keep the original (ϕ_i, ψ_i) pair
2. From the internal angle sample, generate the explicit structure in $\mathbb{R}^3$
3. Compute the number of clashes caused by hinges at level l, and accept the torsion angle changes for level l if the number of clashes are less than some parameter c. We define a *clash* as two atoms occupying the same space in $\mathbb{R}^3$.

As we are most interested in modeling the large-scale uncertainty that arises from domain movements, we then find the optimal placement of side-chain atoms using SCWRL4 [17], followed by a brief energy minimization step with Amber16 [6] to remove any steric strain. Finally, we rank each sample by free energy, and keep only the samples with the lowest energy. These final two steps (minimization and ranking by energy) prevent us from using samples that are biologically irrelevant.

2.5 Benchmark dataset

In this research, we are interested in 1) modeling the uncertainty of a given protein-protein docking algorithm, but also 2) improving the existing docking results in the unbound-unbound case. The Zlab benchmark 5 [28] contains a set of proteins that have had the X-ray structure determined both in isolation and together, and consist of 254 protein pairs classified as either *difficult*, *medium difficulty*, or *rigid-body*, depending on the interface RMSD (iRMSD). The difficult class of proteins have an iRMSD of > 2.2Å, which means there is typically some movement between bound and unbound conformations.

To select our set of input structures, we docked each protein classified as “difficult” in both the bound and unbound conformations with F2Dock [1, 9], a rigid-body docking algorithm. We selected those proteins that performed well in the bound structure but poorly with the unbound structure as candidates in our benchmark. The criteria we use for differentiating between success and failure is whether there exists a “hit” in the top 1000 reported poses. We define a “hit” as a bound pair with iRMSD within 5Å of the actual bound conformation.

Since we are primarily interested in the single-body docking problem (and not the multi-body docking problem), we only kept the single-chain proteins for our experiment, which led to 10 single-chain proteins that perform well when using the bound conformation but

Table 1 Protein structures used in dataset, labeled according to the ID from the ZLab benchmark 5 [28]. The top section contain those that performed well when bound, the bottom section containing those that did not.

ID	# Residues			iRMSD (Å)	ΔEnergy (J)
	receptor	ligand	contact		
1ATN	372	258	36	42	131
1F6M	320	108	62	16	87
1FQ1	183	295	53	53	367
1BKD	439	166	97	20	425
1R8S	160	187	61	28	439
1RKE	262	176	68	52	524
1ZLI	306	74	77	13	212
2C0L	292	122	92	17	366
2I9B	265	122	101	29	387
2J7P	292	265	80	20	370
2OT3	253	157	69	17	428
3FN1	160	90	38	14	315
1H1V	368	327	74	33	180
1Y64	411	357	66	39	192
3AAD	264	153	42	37	66

not when unbound. In addition, we also included the 3 single-chain proteins that performed poorly when both the bound and unbound conformation were used. Statistics on the size and free energy of each protein are given in Table 1.

3 Results and discussion

3.1 Conformational sampling distributions

Our primary concern for generating a good set of samples is that the samples cover a good portion of the feasible set of the protein conformational space. We consider two metrics for measuring coverage: 1) the free energy of individual proteins, and 2) the iRMSD from the sample to the bound conformation. The first of these metrics is an unbiased measure of protein stability: if all samples have abnormally high energy, they are unlikely to be biologically feasible. However, it is possible that the bound conformation lies in an energy well made more available when in combination with the second protein. For this reason, we are not interested in only finding the energy minimum, but also the distance from the bound conformation. Figure 4 shows the energy vs iRMSD for 1000 samples of each protein.

3.2 Improvements in unbound-unbound docking

By generating a set of proteins that have a closer iRMSD to the bound conformation, we are able to improve on the blind unbound-unbound docking protocol, for both rigid-body and flexible docking algorithms. We compare the results for the bound and unbound case for F2Dock (a rigid-body docking algorithm) [1], Rosetta (a semi-flexible docking algorithm) [15, 7], and SwarmDock (a flexible docking algorithm) [20]. We perform bound-bound and unbound-unbound docking for each program, and compute the iRMSD on the reported poses. For F2Dock and Rosetta, the number of reported poses is variable, which we set to 1000. SwarmDock reports a fixed number of results, so this number varies from 465–548 poses.

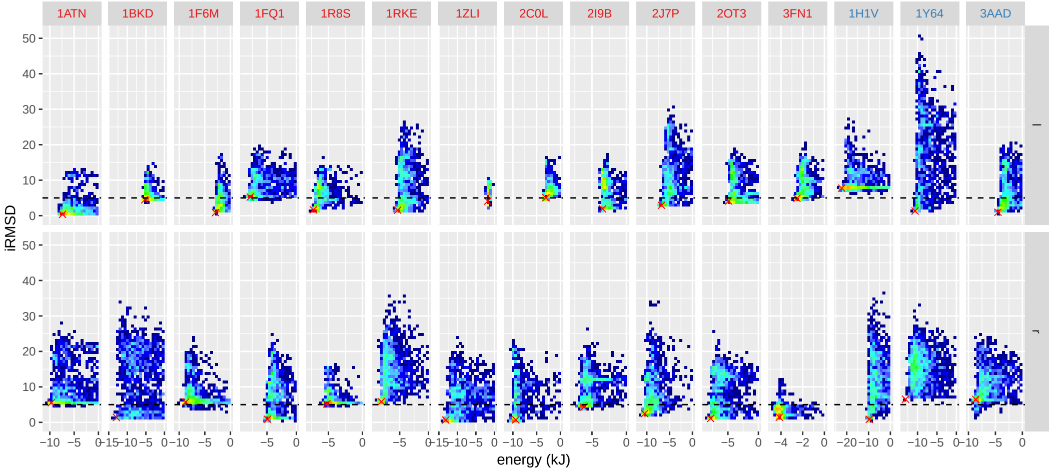

Figure 4 Plot of iRMSD (against the bound conformation) vs energy for samples generated from the unbound conformation. Proteins are separated by ligand (top) and receptor (bottom) to show the difference in individual protein movement. The black dashed line shows iRMSD= 5, or the value at which a match is considered a "hit," and the red "X" marks the spot of the original unbound protein. For all proteins, there exist some samples that improve on both the iRMSD and energy; some of the proteins, such as 3AAD receptor and 1ZLI ligand, improve upon the iRMSD greatly. Strong convergence is shown by a funnel-shaped energy landscape, and is seen for many protein pairs. Protein labels are colored red (good when bound) and blue (bad when bound).

Since F2Dock and Rosetta both have command-line interfaces, we also perform docking 50 samples of the unbound conformation of each protein. The minimum iRMSD for each protein (bound, unbound, and samples for F2Dock and Rosetta) are found in Table 2.

The results from the iRMSD statistics suggest a few findings. First, the flexible algorithms (Rosetta and SwarmDock) are better at docking the bound-bound conformations than the rigid-body one (F2Dock). This is potentially due to the fact that clashes in side-chain atoms prevent the rigid body docking algorithm from correctly identifying the best conformation, but also could be due to the fact that each program uses a different energy function, and may be better tuned for these specific proteins in the flexible programs. The most important observation, however, is the huge difference in iRMSD between the bound and unbound pairs. This suggests that the input to the algorithm (e.g. unbound or bound) is an important characteristic of the docking result, and variations in input structure must be accounted for and described in the output QOI as empirical certificates. Finally, we also note that for each protein, using many different sampled proteins always improves the iRMSD of the docking result (sometimes drastically), suggesting that the sampling protocol is sound (leading to better results).

3.3 Probabilistic certificates from Quasi-Monti Carlo samples

To describe the uncertainty of the results of the docking algorithm, we compute the probabilistic certificates arising from the Chernoff-like bounds of the sampled algorithms, given in Equation 1. This provides a metric that can compare across proteins (for the same docking algorithm) and across docking algorithms (for the same protein, or over all proteins). We could provide probabilistic certificates for any QOI; however, we are primarily interested in bounding the binding free energy. If the reported free energy is tightly bound by a probabilistic certificate, we are more confident that we have identified the correct free energy.

Table 2 Best RMSD (over top 1000 poses for F2Dock and Rosetta and all poses for SwarmDock) for proteins included in this dataset. A single asterisk marks proteins not in the bound form with at least one hit (iRMSD < 5Å) in the top poses. F2Dock and Rosetta statistics for sampled proteins are also included. Note the great improvement on Rosetta docking when using the sampled proteins, and that the sampled proteins are always better than the unbound case. The large differences between bound and unbound input suggests the output QOI is highly dependent on the input to the model.

	F2Dock			Rosetta			SwarmDock	
ID	bound	unbound	sampled	bound	unbound	sampled	bound	unbound
1ATN	1.2	7.2	*4.6	0.08	8.6	6.8	0.98	*4.6
1BKD	1.3	8.7	*4.9	0.23	16.9	5.4	0.68	8.7
1F6M	1.2	8.1	*5.0	0.11	17.9	13.4	0.69	5.6
1FQ1	2.3	6.4	*4.4	0.48	15.0	8.2	3.55	5.6
1R8S	1.7	9.4	6.2	0.22	14.5	6.2	0.72	5.1
1RKE	0.8	7.1	5.8	0.15	15.1	12.8	0.65	5.4
1ZLI	0.6	10.0	*4.6	0.13	10.6	7.1	0.71	9.0
2C0L	0.5	*4.8	*4.0	0.30	12.2	8.2	0.75	*3.8
2I9B	1.1	8.4	*4.7	0.09	13.6	8.8	7.93	6.5
2J7P	1.4	7.2	*3.4	1.12	17.2	14.9	0.60	6.6
2OT3	1.1	5.1	*3.9	0.16	15.6	6.8	0.92	6.0
3FN1	1.0	5.5	*4.9	0.10	9.9	*4.8	0.53	*4.1
1H1V	8.8	11.0	8.0	0.27	18.8	13.5	0.68	9.1
1Y64	9.3	11.4	10.7	1.9	35.0	15.6	1.37	11.7
3AAD	7.0	8.2	5.3	0.39	22.0	9.2	2.36	7.1

Figure 5 shows a comparison of the certificate for ΔG of each protein (at Pr = 0.9, see Equation 1), and includes the true QOI (red dot), computed on the bound-bound conformation. For some of the proteins (e.g. 3FN1), the provided certificate is much tighter than others (e.g. 2C0L). This also allows us to directly compare the two different programs in terms of docking uncertainty. While the rigid F2Dock algorithm occasionally has higher bounds, with high probability the true statistic lies within the Pr = 0.9 certificate range. The Rosetta results usually contain the true QOI, but is not contained within the min/max range for 3FN1.

4 Conclusion

In this work, we provide a framework for providing probabilistic certificates on uncertainty in a docking algorithm. Fundamental to these certificates is the contribution of a low-discrepancy hierarchical sampling protocol that includes general amino acid information in the form of Ramachandran plots, but also structural information that is specific to the input protein in the form of a bivariate von Mises distribution. We show that the low-discrepancy samples generated by this protocol explore the energy landscape for the unbound protein, which includes samples closer to the bound conformation.

With these samples, we compare three different docking algorithms, ranging from rigid-body to completely flexible. We show that docking results vary substantially depending on the input protein structure – even for the flexible docking algorithms – further substantiating our claim that uncertainty quantification is essential to protein-protein docking.

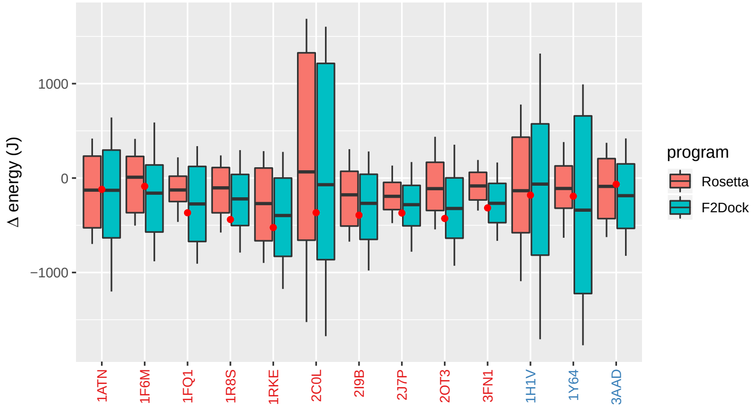

Figure 5 Probabilistic bounds compared to ground truth for values of ΔG. Protein labels are colored red (good when bound) and blue (bad when bound). The box shows the value of the certificate at Pr = 0.9 and the tails show the min/max values, and the red point shows the true statistic, computed on the bound-bound form of the protein.

Finally, we repeat the protein-protein docking experiments with different structures from our hierarchical sampling protocol and assess the variations in reported binding free energy. We compute a probablistic certificate for the binding free energy, and compare the 90% confidence interval with the value computed on the bound complex. This provides a tool for comparing not only uncertainty across proteins, but also across docking algorithms.

References

1 C. Bajaj, R. Chowdhury, and V. Siddahanavalli. F2Dock: Fast Fourier Protein-protein Docking. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 8(1):45–58, 2011.

2 Chandrajit Bajaj, Rezaul Alam Chowdhury, and Vinay Siddavanahalli. F3Dock: A fast, flexible and Fourier based approach to protein-protein docking. *The University of Texas at Austin, ICES Report*, pages 08–01, 2008.

3 Chandrajit L. Bajaj, Abhishek Bhowmick, Eshan Chattopadhyay, and David Zuckerman. On Low Discrepancy Samplings in Product Spaces of Motion Groups. *arXiv e-prints*, page arXiv:1411.7753, November 2014. arXiv:1411.7753.

4 Ian David Brown. Recent developments in the methods and applications of the bond valence model. *Chemical reviews*, 109(12):6858–6919, 2009.

5 Yue Cao and Yang Shen. Bayesian active learning for optimization and uncertainty quantification in protein docking. *arXiv preprint*, 2019. arXiv:1902.00067.

6 D.A. Case, R.M. Betz, D.S. Cerutti, et al. *AMBER 2016.* University of California, San Francisco, 2016.

7 Sidhartha Chaudhury, Monica Berrondo, Brian D Weitzner, Pravin Muthu, Hannah Bergman, and Jeffrey J Gray. Benchmarking and analysis of protein docking performance in Rosetta v3. 2. *PloS One*, 6(8):e22477, 2011.

8 Herman 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.

9 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.

10 Nathan Clement, Muhibur Rasheed, and Chandrajit Lal Bajaj. Viral capsid assembly: A quantified uncertainty approach. *Journal of Computational Biology*, 25(1):51–71, 2018.

11 Ugur Emekli, Dina Schneidman-Duhovny, Haim J Wolfson, Ruth Nussinov, and Turkan Haliloglu. HingeProt: automated prediction of hinges in protein structures. *Proteins: Structure, Function, and Bioinformatics*, 70(4):1219–1227, 2008.

12 Vamshi K. Gangupomu, Jeffrey R. Wagner, In-Hee Park, Abhinandan Jain, and Nagarajan Vaidehi. Mapping Conformational Dynamics of Proteins Using Torsional Dynamics Simulations. *Biophysical Journal*, 104(9):1999–2008, 2013.

13 Bryant Gipson, David Hsu, Lydia E Kavraki, and Jean-Claude Latombe. Computational models of protein kinematics and dynamics: Beyond simulation. *Annual Review of Analytical Chemistry*, 5:273–291, 2012.

14 Barry J Grant, Ana PC Rodrigues, Karim M ElSawy, J Andrew McCammon, and Leo SD Caves. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*, 22(21):2695–2696, 2006.

15 Jeffrey Gray, Stewart Moughon, Chu Wang, Ora Schueler-Furman, Brian Kuhlman, Carol Rohl, and David Baker. Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of Molecular Biology*, 331(1):281–299, 2003.

16 Fred James, Jiri Hoogland, and Ronald Kleiss. Quasi-Monte Carlo, discrepancies and error estimates. *Methods*, page 9, 1996. `arXiv:physics/9611010`.

17 Georgii G Krivov, Maxim V Shapovalov, and Roland L Dunbrack. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins: Structure, Function, and Bioinformatics*, 77(4):778–795, 2009.

18 Daisuke Kuroda and Jeffrey J Gray. Pushing the backbone in protein-protein docking. *Structure*, 24(10):1821–1829, 2016.

19 Loren M LaPointe, Keenan C Taylor, Sabareesh Subramaniam, Ambalika Khadria, Ivan Rayment, and Alessandro Senes. Structural organization of FtsB, a transmembrane protein of the bacterial divisome. *Biochemistry*, 52(15):2574–2585, 2013.

20 Xiaofan Li, Iain H Moal, and Paul A Bates. Detection and refinement of encounter complexes for protein–protein docking: taking account of macromolecular crowding. *Proteins: Structure, Function, and Bioinformatics*, 78(15):3189–3196, 2010.

21 Kanti Mardia, Charles Taylor, and Ganesh Subramaniam. Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. *Biometrics*, 63(2):505–512, 2007.

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

23 R. J. Milgram, G. Liu, and J. C. Latombe. On the structure of the inverse kinematics map of a fragment of protein backbone. *Journal of Computational Chemistry*, 29(1):50–68, 2008.

24 Harald Niederreiter. Quasi-Monte Carlo methods. *Encyclopedia of Quantitative Finance*, 24(1):55–61, 1990.

25 Dzmitry Padhorny, Andrey Kazennov, Brandon S Zerbe, Kathryn A Porter, Bing Xia, Scott E Mottarella, Yaroslav Kholodov, David W Ritchie, Sandor Vajda, and Dima Kozakov. Protein–protein docking by fast generalized Fourier transforms on 5D rotational manifolds. *Proceedings of the National Academy of Sciences*, 113(30):E4286–E4293, 2016.

26 Muhibur Rasheed, Radhakrishna Bettadapura, and Chandrajit Bajaj. Computational Refinement and Validation Protocol for Proteins with Large Variable Regions Applied to Model HIV Env Spike in CD4 and 17b Bound State. *Structure*, 23(6):1138–1149, 2015.

27 Muhibur Rasheed, Nathan Clement, Abhishek Bhowmick, and Chandrajit L Bajaj. Statistical framework for uncertainty quantification in computational molecular modeling. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2017.

28 Thom Vreven, Iain H Moal, Anna Vangone, Brian G Pierce, et al. Updates to the Integrated Protein–Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *Journal of Molecular Biology*, 427(19):3031–3041, 2015.

29 Guoli Wang and Roland L Dunbrack Jr. PISCES: a protein sequence culling server. *Bioinformatics*, 19(12):1589–1591, 2003.