HHS Public Access

Author manuscript

J Chem Theory Comput. Author manuscript; available in PMC 2021 August 11.

Published in final edited form as:

J Chem Theory Comput. 2020 August 11; 16(8): 5334–5347. doi:10.1021/acs.jctc.0c00476.

Bayesian active learning for optimization and Uncertainty quantification in protein docking†

Yue Cao‡, **Yang Shen**‡,¶

‡Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, United States

¶TEES-AgriLife Center for Bioinformatics and Genomic Systems Engineering, Texas A&M University, College Station, TX 77840, United States

Abstract

Ab initio protein docking represents a major challenge for optimizing a noisy and costly "black box"-like function in a high-dimensional space. Despite progress in this field, there is a lack of rigorous uncertainty quantification (UQ). To fill the gap, we introduce a novel algorithm, Bayesian Active Learning (BAL), for optimization and UQ of such black-box functions with applications to flexible protein docking. BAL directly models the posterior distribution of the global optimum (i.e. native structures) with active sampling and posterior estimation iteratively feeding each other. Furthermore, it uses complex normal modes to span a homogeneous, Euclidean conformation space suitable for high-dimensional optimization and constructs funnel-like energy models for quality estimation of encounter complexes.

Over a protein-docking benchmark set and a CAPRI set including homology docking, we establish that BAL significantly improves against starting points from rigid docking and refinements by particle swarm optimization, providing a top-3 near-native prediction for one third targets. Quality assessment empowered with UQ leads to tight quality intervals with half range around 25% of actual interface RMSD and confidence level at 85%. BAL's estimated probability of a prediction being near-native achieves binary classification AUROC at 0.93 and AUPRC over 0.60 (compared to 0.50 and 0.14, respectively, by chance), which also improves ranking predictions. This study represents the first UQ solution for protein docking, with rigorous theoretical frameworks and comprehensive empirical assessments.

Graphical Abstract

†Source codes, trained machine learning-based energy models, and supplemental videos are available at https://github.com/Shen-Lab/BAL.

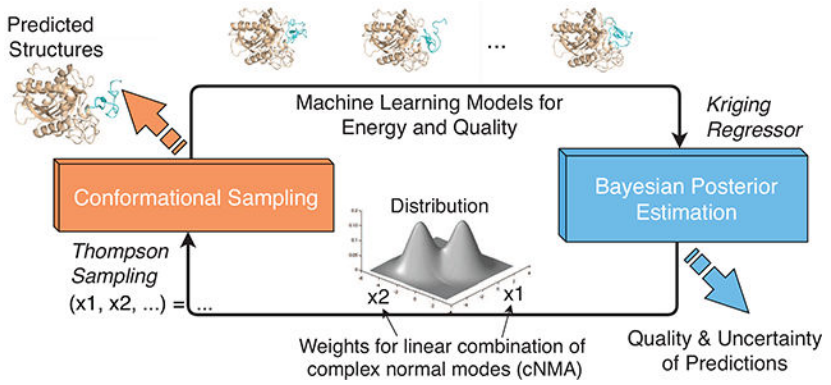
yshen@tamu.edu.

Conflict of interest: None declared.

Associated Contents

Supporting Information

The Supporting Information is available free of charge at http://pubs.acs.org. Theoretical and empirical comparisons between Bayesian Active Learning and Nonparametric Conjugate Prior Distribution; Fast RMSD calculation; Prediction of the extent of conformational changes; Constraints on the extent for the ligand; List of protein complexes used in this study; Training energy models; Numerical comparisons between PSO and BAL over test functions and protein docking; Sampled energy landscapes for all test cases; Parameter values; Running time statistics; and Video information.



1 Introduction

Protein-protein interactions underlie many cellular processes, which has been increasingly revealed by the quickly advancing high-throughput experimental methods. However, compared to the binary information about which protein pairs interact, the structural knowledge about how proteins interact remains relatively scarce[1]. Protein docking helps close such a gap by computationally predicting the 3D structures of protein-protein complexes given individual proteins' 1D sequences or 3D structures[2].

Ab initio protein docking is often recast as an energy (or other objective functions) optimization problem. For the type of objective functions relevant to protein docking, neither the analytical form nor the gradient information would help global optimization as the functions are non-convex and extremely rugged thus their gradients are too local to inform the global landscapes. So the objective functions are often treated as *de facto* "black-box" functions for global optimization. Meanwhile, these functions are very expensive to evaluate. Various protein-docking methods, especially refinement-stage methods, have progressed to effectively sample the high-dimensional conformational space against the expensive functional evaluations[3–9].

While solving such optimization problems still remains a great challenge, quantifying the uncertainty of numerically-computed optima (docking solutions) is even more challenging and has not been addressed by any protein-docking method. Even though the uncertainty information is much needed by the end users, current protein-docking methods often generate a rank-ordered list of results without giving quality estimation with uncertainty to individual results and without providing the confidence in whether the entire list contains a quality result (for instance, a near-native protein-complex model with iRMSD ≤ 4 Å).

Uncertainty quantification (UQ), if addressed, would lead to two benefits. First, for individual optimization trajectories, uncertainty awareness would improve the robustness of their optimization outcomes. Second, uncertainty of the solutions can be easily fed to any quality assessment tools for distributions rather than point estimates of some quality of interest (very often iRMSD); and comparing these distributions would provide more robust model ranking across trajectories.

Sources of uncertainty in protein docking methods include the objective function as well as the sampling scheme, which can be classified as epistemic uncertainty[10]. For instance, energy models as objective functions provide noisy and approximate observations of the assumed ground truth — the Gibbs free energy; and iterative sampling techniques suffer from both the approximation of the search space (e.g. rotamerized side chains) and insufficient data in the approximated space (e.g. small numbers of samples considering the high dimensionality of the search space). In addition, uncertainty in protein structure data (e.g. X-ray crystal structures of proteins being "averaged" versions of their native conformations and derived from fitting observed diffraction patterns), which can be classified as aleatoric uncertainty, also enters protein docking methods when crystal structures are used as ground-truth native structures for training objective functions or tuning parameters in protein-docking methods.

Whereas the forward propagation of aleatoric uncertainty in protein structure data to structure-determined quantities has been studied empirically[11] and theoretically[12], the much more difficult, inverse quantification of uncertainty in predicted protein or protein-complex structures originating from epistemic uncertainty in computational methods, is still lacking a mathematically rigorous solution. A unique challenge for uncertainty quantification (UQ) in protein docking is that the desired quality of interest here is directly determined by the optimum itself rather than the optimal value. In other words, closeness to native structures (for instance, measured by iRMSD) is an indicator for the usefulness of the docking results, but closeness to native structures' energy values is not necessarily the case. Therefore, UQ in protein docking has to be jointly solved with function optimization when finding the inverse mapping from a docking objective function to its global optimum is neither analytically plausible nor empirically cheap.

In this study, we introduce a rigorous Bayesian framework to simultaneously perform function optimization and uncertainty quantification for expensive-to-evaluate black-box objective functions. To that end, our Bayesian active learning (BAL) iteratively and adaptively generates samples and updates posterior distributions of the global optimum. Specifically, we propose a posterior in the form of the Boltzmann distribution building upon a non-parametric kriging regressor and a novel adaptive-annealing schedule. The iteratively updated posterior carries the belief (and uncertainty as well) on where the global optimum is given historic samples and guides next-iteration sampling, which presents an efficient data-collection scheme for both optimization and UQ. Compared to typical Bayesian optimization methods[13] that first model the posterior of the objective function and then optimize the resulting functional, our BAL framework directly models the posterior of the global optimum and overcomes the intensive computations in both steps of typical Bayesian optimization methods. Compared to another work[14] that also models the posterior of the global optimum, Nonparametric Conjugate Prior Distribution (or NCPD in short), we provide both theoretical and empirical results that our BAL has a consistent and unbiased estimator as well as a global uncertainty-aware and dimension-dependent annealing schedule.

We also make innovative contributions in the application domain of protein docking. Specifically, we design a machine learning-based objective function that estimates binding

affinities for docked encounter complexes as well as assesses the quality of interest, iRMSD, for docking results. And we re-parameterize the search space for both external rigid-body motions[15] and internal flexibility[5], into a low-dimensional homogeneous and isotropic space suitable for high-dimensional optimization, using our (protein) complex normal modes (cNMA)[16]. Considering that protein docking refinement often starts with initial predictions representing separate conformational clusters/regions, we use estimated local posteriors over individual regions to construct local and global partition functions; and then calculate the probability that the prediction for each conformation, each conformational cluster, or the entire list of conformational clusters is near-native.

The rest of the paper is organized as following. In Materials and Methods, we first give a mathematical formulation for the optimization and the UQ, then introduce our Bayesian active learning (BAL) that iteratively updates sampling and posterior estimation. We next introduce the parameterization of the search space that allows concurrent and homogeneous sampling of external rigid-body and internal flexible-body motions as well as newly-developed machine learning models as the noisy energy function that estimates the binding free energy for encounter complexes. And we end the Materials and Methods with uncertainty quantification for protein docking.

In Results and Discussion, using a comprehensive protein docking benchmark set involving unbound docking and a CAPRI set involving homology docking, we assess optimization results for BAL with comparison to starting structures from ZDOCK and refined structures by particle swarm optimization (PSO). We further assess the uncertainty quantification results: accuracy of the confidence levels and tightness of the confidence region, as well as the confidence scores for the near-nativeness of predictions. Case studies further reveal the causes of success and failure. Lastly, before reaching conclusions, we visualize the estimated energy landscape and confirm that the funnel-like energy landscapes do exist near native structures in the homogeneous conformational space blending external rigid-body and internal flexible-body motions.

2 Materials and Methods

2.1 Mathematical Formulation

We consider a black-box function $f(\mathbf{x})$ (e.g. ΔG, the change in the Gibbs free energy upon protein-protein interaction) that can only be evaluated at any sample with an expensive yet noisy observation $y(\boldsymbol{x})$ (e.g. modeled energy difference or scoring function for conformation $\boldsymbol{x}$). Our goal in optimization is

$$\boldsymbol{x}^* = \arg \min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$$

(e.g. $\boldsymbol{x}^*$ denotes the native structure of a protein complex and $\mathcal{X}$ denotes the domain of sample space for conformations). And our goal in uncertainty quantification is the probability distribution of $\boldsymbol{x}^*$ around its prediction $\hat{\boldsymbol{x}}$, rather than the single point estimation $\hat{\boldsymbol{x}}$ itself (in which the distributions is equivalent to an impulse function at $\hat{\boldsymbol{x}}$).

Once the uncertainty of the solution is quantified, the uncertainty of the solution quality can be subsequently derived. A summary of the latter is simply the probability that the prediction $\hat{\boldsymbol{x}}$ falls in an interval [*lb*,*ub*] of quality relative to $\boldsymbol{x}^*$:

$$P(lb \leqslant Q(\hat{\boldsymbol{x}}, \boldsymbol{x}^*) \leqslant ub) = 1 - \sigma,$$

where 1-σ is the confidence level; [*lb*,*ub*] is the confidence interval in the solution quality; and Q(·,·), the quality of interest measuring some distance or dissimilarity, can be an Euclidean norm (as in our assessment for test functions), another distance metric, or other choices dependent on the user (for instance, iRMSD as in our assessment for protein docking with $ub = 4$ Å). Note that Q(·,·) can be any quality assessment (QA) tool that does not assume the knowledge of $\boldsymbol{x}^*$ as well.

2.2 Bayesian active learning with a posterior of x*

We address the problem above in a Bayesian perspective: instead of treating $\boldsymbol{x}^*$ as a fixed point, we model $\boldsymbol{x}^*$ as a random variable and construct its probability distribution, $p(\boldsymbol{x}^*|\mathscr{D})$, given samples $\mathscr{D} = \{(\boldsymbol{x}, y)\}$. This probability distribution, carrying the belief and the uncertainty on the location of $\boldsymbol{x}^*$, is a prior when $\mathscr{D} = \emptyset$ (no sample) and a posterior otherwise. Considering the cost of function evaluation, we iteratively collect new samples in iteration t (where all samples collected by the end of the t-th iteration are denoted $\mathscr{D}^{(t)}$) based on the latest estimated posterior, $p(\boldsymbol{x}^*|\mathscr{D}^{(t-1)})$; and we update the posterior $p(\boldsymbol{x}^*|\mathscr{D}^{(t)})$ based on $\mathscr{D}^{(t)}$. An illustration of the iterative approach is given in Fig. 1.

For optimization, we set $\hat{\boldsymbol{x}}$ to be the best sample with the lowest y value given a computational budget (reflected in the number of samples or iterations). For UQ, given the posterior $p(\boldsymbol{x}^*|\mathscr{D}^{(t)})$ in the final iteration, one can propagate the inferred uncertainty in $\boldsymbol{x}^*$ forwardly to that in the quality of interest, $Q(\hat{\boldsymbol{x}}, \boldsymbol{x}^*)$, for the found $\hat{\boldsymbol{x}}$, using techniques such as Markov chain Monte Carlo.

2.2.1 Non-parametric posterior of x*—

We propose to use the Boltzmann distribution to describe the posterior

$$p(\boldsymbol{x}^*|\mathscr{D}^{(t)}) \propto \exp(-\rho \cdot \hat{f}(\boldsymbol{x}))$$

where $\hat{f}(\boldsymbol{x})$ is an estimator for $f(\boldsymbol{x})$, and ρ is a parameter (sometimes $\frac{1}{RT}$ where R is the gas constant and T the temperature of the molecular system).

To iteratively guide the expensive sampling and balance between exploration and exploitation in a data efficient way, we choose ρ to follow an adaptive annealing schedule over iteration t:

$$\rho_t = \rho_0 \cdot \exp((h_p^{(t-1)})^{-1} \frac{1}{n_t^d})$$

where ρ_0, the initial ρ, is a parameter; $h_p^{(t-1)}$ is the (continuous) entropy of the last-iteration posterior, a shorthand notation for $h(p(\boldsymbol{x}^*|\mathscr{D}^{(t-1)}))$; $n_t = |\mathscr{D}^{(t)}|$ is the number of samples collected so far; and d is the dimensionality of the search space $\mathscr{X}$.

This annealing schedule is inspired by the adaptive simulated annealing (ASA)[17], especially the exponential form and the $n_t^{\frac{1}{d}}$ term. However, we use the $(h_p^{(t-1)})^{-1}$ term rather than a constant as in ASA so that we exploit all historic samples $\mathscr{D}^{(t)}$. In this way, as the uncertainty of $\boldsymbol{x}^*$ decreases, ρ_t increases and shifts the search toward exploitation.

The function estimator $\widehat{f}(\boldsymbol{x})$ also updates iteratively according to the incrementally increasing n_t samples $\mathscr{D}^{(t)} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_t}$. We Use a consistent and unbiased kriging regressor[18] which is known to be the best unbiased linear estimator(BLUE):

$$\widehat{f}(\boldsymbol{x}) = f_0(\boldsymbol{x}) + (\kappa^{(t)}(\boldsymbol{x}))^T (\boldsymbol{K}^{(t)} + \varepsilon^2 \boldsymbol{I})^{-1} (\boldsymbol{y}^{(t)} - \boldsymbol{f}_0^{(t)})$$

where $f_0(\boldsymbol{x})$ is the prior for $E[f(\boldsymbol{x})]$; $\boldsymbol{\kappa}^{(t)}(\boldsymbol{x}) \in \mathscr{R}^{n_t}$ is the kernel vector with the ith element being the kernel, a measure of similarity, between $\boldsymbol{x}$ and $\boldsymbol{x}_i \in \mathscr{D}^{(t)}$; $\boldsymbol{K}^{(t)} \in \mathscr{R}^{n_t \times n_t}$ is the kernel matrix with the(i,j) element being the kernel between $\boldsymbol{x}_i \in \mathscr{D}^{(t)}$ and $\boldsymbol{x}_j \in \mathscr{D}^{(t)}$; $\boldsymbol{y}^{(t)}$ and $\boldsymbol{f}_0^{(t)}$ are the vector of $y_1, \ldots, y_{n_t}$ and $f_0(\boldsymbol{x}_1), \ldots, f_0(\boldsymbol{x}_{n_t})$, respectively; and ϵ reflects the noise in the observation and is estimated to be 2.1 as the prediction error for the training

We derive the kriging regressor in the Supporting Information (S1) Sec. 1.2.2. And we will use the regressor to evaluate binding energy and estimate iRMSD for UQ over multiple regions in Sec. 2.5.

2.2.2 Adaptive sampling based on the latest posterior—For a sequential sampling policy that balances exploration and exploitation during the search for the optimum, we choose Thompson sampling[19] which samples a batch of points in the t-th iteration based on the latest posterior $p(\boldsymbol{x}^*|\mathscr{D}^{(t-1)})$. This seemingly simple policy has been found to be theoretically[20] and empirically[21] competitive compared to other updating policies such as Upper Confidence Bound[13]. In our case, it is actually straightforward to implement given the posterior on $\boldsymbol{x}^*$.

There are multiple reasons to collect in each iteration a batch of samples rather than a single one. First, given the high dimension of the search space, it is desired to collect adequate data before updating the posterior. Second, the batch sampling weakens the correlation among samples and make them more independent, which benefits the convergence rate of the kriging regressor. Last, parallel computing could be trivially applied for batch sampling, which would significantly improve the algorithm throughput.

Fig. 1 gives an illustration of the algorithm behavior. The initial samples drawn from a uniform distribution leads to a relatively flat posterior whose maximum is off the function

optimum (Fig. 1F). As the iteration progresses, the uncertainty about the optimum gradually reduces(Fig. 1E) and newer samples are increasingly focused (Fig. 1C,D) as the posteriors are becoming narrower with peaks shifting toward the function optimum (Fig. 1G,H).

In our docking study, $d = 12$ for a homogeneous space spanned by complex normal modes (see Sec. 2.3). We construct a prior and collect 30 samples in the first iteration and 20 in each of the subsequent iterations. We limit the number of iterations (samples) to be 31 (630) for optimization and posteriors as a way to impose a computational budget (6–13 CPU hours for protein complexes of typical sizes). The reason is that it costs minutes to locally minimize each conformational sample using CHARMM[22] and remove bond distortions in flexible perturbations, before energies can be evaluated. More samples, albeit more expensive, would improve the quality of energy minimization and posterior estimation. At the end of all iterations, an additional set of 1,000,000 samples will be generated according to the final posterior, for quality estimation and uncertain quantification of the final prediction. Note that these 1,000,000 samples do not drive the optimization process and do not need to be locally minimized anymore; and this stage of post-optimization UQ costs about a CPU hour.

2.2.3 Kernel with customized distance metric

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2l^2}\right),$$

The kernel in the kriging regressor for the posterior is a measure of similarity. For test functions defined in an Euclidean space, we use the radial basis function (RBF) kernel: where $||\boldsymbol{x}_i - \boldsymbol{x}_j||$, a measure of dissimilarity, is the Euclidean distance and l, the bandwidth of the kernel, is set as $l = l_0 \cdot n_t^{\frac{1}{d}}$ following Györfi et al.[23]. l_0,dependent on search space,is set at 2.0 for docking without particular optimization.

For protein docking we replace the Euclidean distance in the RBF kernel with the interface RMSD (iRMSD) between two sample structures. iRMSD captures sample dissimlarity relevant to function-value dissimilarity and is independent of search-space parameterization. For this purpose, we also have to address two technical issues. First, protein interface information is determined by $\boldsymbol{x}^*$ and thus unknown. We instead use the putative interface seen in the samples. Each iRMSD is calculated using the same set of C_α atoms, the union of interface C_α atoms derived from 50 random perturbations of the starting structure (see more details in the SI Sec. 2.4). Second, kernel calculation with iRMSD is time consuming. The time complexity of iRMSD calculation is $O(N)$ and that of regressor update is $O(Nn^2)$, where N, the number of interfacial atoms, can easily reach hundreds or thousands, and n, the number of samples, can also be large. To save computing time, we develop a fast RMSD calculation method that reduces its time complexity from $O(N)$ down to $O(1)$ (see details in SI Sec. 2.1).

2.2.4 Related methods—

Current Bayesian optimization methods typically model the posterior distribution of f(**x**) rather than that of $\boldsymbol{x}^* = \arg\min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$ directly. After modeling the posterior distribution over the functional space (a common non-parametric way is through Gaussian processes), they would subsequently sample the functional space and optimize sample functions. For instance, Villemonteix et al. [24] used Monte Carlo sampling; and Hernández-Lobato et al. [25] discretized the functional space to approximate the sample paths of Gaussian processes using a finite number of basis functions then optimized each sample path to get one sample of $\boldsymbol{x}^*$. The two-step approach of current Bayesian optimization methods involve intensive sampling and (non-convex) optimization that is computationally intensive and not pragmatically useful for protein docking.

To our knowledge, Ortega et al. presented the only other study that directly models the posterior distribution over the optimum[14]. Both their method NCPD and our BAL fall in the general category of Bayesian optimization and use consistent non-parametric regressors. However, we prove in Sec. 1.2 of the SI that their regressor is biased whereas our kriging regressor is unbiased. We explain in Sec. 1.1 of the SI that their annealing schedule (temperature control) only considers the pairwise distance between samples without location awareness and is independent of dimensionality d; whereas ours has a term involving location-aware global uncertainty and generalizes well to various dimensions. Beyond those theoretical comparisons, we also included empirical results to show the superior optimization and UQ performances of BAL.

The rest of the Materials and Methods section involve methods specific to the protein docking problem: parameterization, dimensionality reduction, and range reduction of the search space $\mathcal{X}$; machine learning model as $y(\boldsymbol{x})$, i.e., an energy model for encounter complexes; quality assessment with uncertainty quantification for a predicted structure or a list of predictions; and the use of such assessment metrics for scoring purposes: ranking predictions or classifying their nativeness.

2.3 Conformational Sampling in $\mathcal{X}$

In protein docking the full search space $\mathcal{X}$ captures the degrees of freedom for all atoms involved. Let one protein be receptor whose position is fixed and the other be ligand (the larger one is often chosen as the receptor, as done in a protein-docking benchmark set[26]). And let N_R, N_L, and N be the number of atoms for the receptor, the ligand, and the complex respectively. Then $\mathcal{X} = \mathbb{R}^{3N-6}$ is a Euclidean space whose dimension easily reaches 10^4 for a small protein complex without surrounding solvent molecules. If accuracy is sacrificed for speed, proteins can be (unrealistically) considered rigid and $\mathcal{X} = SE(3) = \mathbb{R}^3 \times SO(3)$ is a Riemannian manifold[27] of ligand translations and rotations. Docking methods fall in the spectrum between these two ends that are represented by all-atom molecular dynamics and FFT rigid docking, respectively. For instance, one can consider locally rigid pieces of a protein rather than a globally rigid protein, then $\mathcal{X}$ becomes the product of many $SE(3)$ for local rigidity[28]; or one can model individual proteins' internal flexible-body motions using normal modes on top of the ligand rigid-body motions, thus $\mathcal{X}$ becomes the product of $\mathbb{R}^K$ (where $K << N_{R/L}$) and $SE(3)^4$.

From the perspective of optimization and UQ, both the high-dimensionality of $\mathbb{R}^{3N-6}$ and the geometry of the lower-dimensional manifold present challenges. Almost all dimensionality reduction efforts in protein docking impose conditions (such as aforementioned local or global rigidity) in the full Euclidean space and lead to embedded manifolds difficult to (globally) optimize over. The challenge from the manifold has been either disregarded in protein docking or addressed by the local tangent space[15,27,28].

Could and how could the dimensionality of the conformational space be reduced while its geometry maintains homogeneity and isotropy of a Euclidean space and its basis vectors span conformational changes of proteins upon interactions? In this subsection we give a novel approach to answer this question for the first time. In contrast to common conformational sampling that separates internal flexible-body motions (often Euclidean) and external rigid-body motions (a manifold)[7], we re-parameterize the space into a Euclidean space spanned by complex normal modes[16] blending both flexible- and rigid-body motions. The mapping preserves distance metric in the original full space. We further reduce the dimensionality and the range in the resulting space[29].

2.3.1 Complex normal modes blend flexible- and rigid-body motions—We previously introduced complex normal mode analysis, cNMA[16], to model conformational changes of proteins during interactions. Using encounter complexes from rigid docking, cNMA extends anisotropic network model (ANM) to capture both conformational selection and induced fit effects. After the Hessian matrix is projected to remove the rigid-body motion of the receptor, its non-trivial eigenvectors $\boldsymbol{\mu}_j$ ($j = 1,..., 3N - 6$) form orthonormal basis vectors. We showed that $\boldsymbol{\mu}_j^{\mathrm{R}}$, the components of the complex normal modes, better capture the direction of individual proteins' conformational changes than conventional NMA did[16]. We also showed that the re-scaled eigenvalues for these components, $\lambda_j^{\mathrm{R}} = \frac{\lambda_j}{||\boldsymbol{\mu}_j^{\mathrm{R}}||^2}$, can be used to construct features for machine learning and predict the extent of the conformational changes.

2.3.2 Dimensionality reduction—In this study we focus on the motions of a whole complex rather than individual proteins and develop sampling techniques for protein docking. Each complex normal mode[16] simultaneously captures concerted flexible-body motions of individual proteins (receptor and ligand) and rigid-body motion of the protein whose position is not fixed (ligand). Such modes together span a homogeneous and isotropic Euclidean space where the distance between two points is exactly the RMSD between corresponding complex structures. The Euclidean space is friendly to high-dimensional optimization. In this study, complex normal modes are precomputed using the starting structure of each conformational cluster and not updated while sampling the cluster to save computational costs.

For dimensionality reduction in the resulting space, we choose the first K_1 non-trivial eigenvectors $\boldsymbol{\mu}_j$ ranked by increasing eigenvalues λ_j; and we additionally include K_2 $\boldsymbol{\mu}_j$ (not in the first K_1) ranked by increasing λ_j^R (λ_j rescaled using the receptor's contribution to this complex normal mode $\boldsymbol{\mu}_j$[16]). In other words, we sample in a($K_1 + K_2$)-dimensional

Euclidean space spanned by complex normal modes and denote the set of basis vectors as $\mathscr{B}$. K_1 of these complex normal modes are the slowest for the whole complex (judging by λ_j) and the rest K_2 different ones are the slowest for the receptor portion of the complex (judging by λ_j^R). In this study K_1 and K_2 are set at 9 and 3, respectively, leading to the dimension of the reduced space to d=12. Empirically, we find that the first 9 non-trivial complex normal modes often contain six with dominant rigid-body motions of the ligand and three with dominant ligand flexibility; and the other 3 in the basis set are, by definition, with dominant receptor flexibility. Supplemental videos illustrate the motions of the two types of complex normal modes and are available along with our codes.

Our framework of Bayesian active learning, using kernels in its unbiased kriging regressor, is applicable to any choice of the basis set $\mathscr{B}$. Naturally, it faces more challenge in optimization, let alone uncertainty quantification, as the dimensionality of $\mathscr{B}$ increases (see empirical results for test functions in Sec. 3.1). Although current basis vectors are low-frequency backbone flexibility derived from normal mode analysis, more vectors can be considered for the set $\mathscr{B}$, such as higher-frequency normal modes, local conformational rearrangements including loop and helix motions, and large conformational changes such as hinge motions. In this work, side-chain flexibility is considered by locally minimizing every conformational sample $\boldsymbol{x}$.

2.3.3 Range reduction—For range reduction in the dimension-reduced space, we perturb a starting complex structure $\vec{C}_0$ along aforementioned basis vectors to generate sample $\vec{C} \in \mathbb{R}^{3N-6}$ while enforcing a prior on the scaling factor s in the first iteration. Specifically

$$\vec{C} = \vec{C}_0 + \sum_{j \in \mathscr{B}} r_j \frac{s}{\sqrt{\lambda_j}} \cdot \boldsymbol{\mu}_j$$

where r_j, the coefficient of the jth normal mode $\boldsymbol{\mu}_j$, is uniformly sampled on S^d, the surface of a d-dimensional standard sphere with a unit radius. The scaling factor s is given by

$$s = \frac{\tau_{\mathrm{R}}}{\frac{1}{\sqrt{N_{\mathrm{R}}}} \left\| \sum_{j \in \mathscr{B}} \frac{r_j}{\sqrt{\lambda_j}} \cdot \boldsymbol{\mu}_j^R \right\|},$$

where $\boldsymbol{\tau}_R$ is the estimated conformational change (measured by RMSD in all C_α atoms) between the unbound and the bound receptor. Note that vectors $\boldsymbol{\mu}_j^{\mathrm{R}}$ (the receptor portion of the jth complex normal mode) are not orthonormal to each other.

We previously predicted τ_{R} by a machine learning model giving $\widehat{\mathrm{RMSD}}_{\mathrm{R}}$, a single value for each receptor[29]. Here we replace $\widehat{\mathrm{RMSD}}_{\mathrm{R}}$ with a predicted distribution by multiplying it to a truncated normal distribution $N(\mu = 0.99, \sigma^2 = 0.31^2)$ within[0,2.5].Thelatter distribution is derived by fitting the ratios between the actual and the predicted values, $\mathrm{RMSD}_{\mathrm{R}}/\widehat{\mathrm{RMSD}}_{\mathrm{R}}$, for 50 training protein complexes (see more details about dataset in Sec. 2.7 and these about

distribution fitting in Sec. 2.2 of the SI). Therefore, our parameterization produces $\boldsymbol{x} = s \cdot \mathbf{r} \in \mathbb{R}^d$ whose prior is derived as above.

Since the ligand component of complex normal modes include simultaneous flexible-and rigid-body motions, conformational sampling could lead to severely distorted ligand geometry. We thus further restrict the ligand perturbation δ_L (flexible- and rigid-body together) to be within $\overline{\Delta}_L$

$$\Delta_L = \sqrt{\frac{1}{N_L}} \left\| \sum_{j \in \mathcal{B}} r_j \frac{s}{\sqrt{\lambda_j}} \cdot \boldsymbol{\mu}_j^L \right\| \leqslant \overline{\Delta}_L,$$

where $\boldsymbol{\mu}_j^L$ denotes the ligand portion of the jth complex normal mode.

We set $\overline{\Delta}_L$ at 6Å according to the average size of binding energy at traction basins seen in conformational clusters[30]. For samples generated from the aforementioned prior or the updated posterior, we reject those violating the ligand perturbation limit. We discuss about the feasibility of the search region in SI Sec.2.3.

Every conformational sample, generated through sampling the prior or the iteratively-updated posteriors of $\boldsymbol{x}^*$, is locally minimized through CHARMM[22] to remove possible bond distortions before energy evaluation. This setting could be changed in future, along with energy models, to reduce the cost of energy evaluation for each sample, allow for more samples, and improve energy minimization and posterior estimation.

2.4 Energy Model $y(\boldsymbol{x})$

We have so far introduced search strategies for functions defined in a Euclidean space or specifically for protein docking. Energy models $y(\boldsymbol{x})$ are at least as important as search strategies for protein docking. In fact, an improved search strategy might expose more deficiencies of energy models, such as false-positive energy wells. We therefore have developed a “funnel-like” energy model to not only mitigate the issue but also to estimate model quality (iRMSD) of encounter complexes.

2.4.1 Binding affinity prediction for sampled encounter complexes—

We introduce a new energy model based on binding affities $K_d'(\boldsymbol{x})$ of structure samples $\boldsymbol{x}$ that are often encounter complexes. The model assumes that K_d' correlates with K_d, the binding affinity of the native complex, and deteriorates with the increase of the sample’s iRMSD (the encounter complex being less native-like):

$$K_d'(\boldsymbol{x}) = K_d \cdot \exp(\alpha \cdot (\text{iRMSD}(\boldsymbol{x}))^q),$$

where α and q are hyper-parameters optimized through cross-validation. In other words, we assume that the fraction of binding affinity loss is exponential in a polynomial of iRMSD. Therefore, the binding energy, a machine learning model $y(\boldsymbol{x};\boldsymbol{w})$ of parameters $\boldsymbol{w}$ can be represented as

$$y(\boldsymbol{x};\boldsymbol{w}) = RT\ln(K'_d(\boldsymbol{x})) = RT\ln(K_d)+RT\alpha\cdot(\mathrm{iRMSD}(\boldsymbol{x}))^q$$

2.4.2 iRMSD prediction for sampled encounter complexes—Given an observed or regressed $y(\boldsymbol{x})$ value for an encounter complex sample, one can estimate iRMSD($\boldsymbol{x}$) with given K_d using the equation above inversely, which provides quality assessment (QA) without native structures.

2.4.3 Machine learning—We train machine learning models, including ridge regression with linear or RBF kernel and random forest, for $y(\boldsymbol{x};\boldsymbol{w})$. The 8 features include changes upon protein interaction in energy terms such as internal energies in bond, angle, dihedral and Urey-Bradley terms, van der Waals, non-polar contribution of the solvation energy based on solvent-accessible surface area (SASA), and electrostatics modeled by Generalized Born with a simple SWitching (GBSW), all of which are calculated in a CHARMM27 force field.

We use the same training set of 50 protein pairs (see details in SI Sec. 2.5) as in predicting the extent of conformational change. From rigid docking and conformational sampling we generate 13,004 complex samples for 50 protein pairs, including 6,464 near-native and 6,540 worse examples in the training set. We balance the near-native and non-near native samples in order to make the binding-energy model, as well as the resulting posterior estimation and uncertainty quantification, focus on near-native or slightly worse encounter complexes as opposed to those with very high iRMSD (say, above 10 Å). Hyper-parameters of ridge regression with RBF kernel as well as random forest are optimized by cross-validation. And model parameters $\boldsymbol{w}$ are trained again over the entire training set with the best hyper-parameters. More details can be found in Sec. 2.6 of the SI. For the assessment, we use the test set **a** of 26 protein pairs (again in SI Sec. 2.5) and generate 20 samples similarly for each of the 10 initial docking results for each pair, leading to 5,200 cases.

2.5 Quality Assessment with Uncertainty Quantification for Protein Docking

A unique challenge to protein docking refinement is that, instead of optimization and UQ in a single region $\mathcal{X}$, we may do so in K separate ones $\mathcal{X}_i$ ($i = 1,...,K$) where each $\mathcal{X}_i$ is a promising conformational region/cluster represented by a initial docking result. This is often necessitated by the fact that the extremely rugged energy landscape is populated with low-energy basins separated by frequent high-energy peaks in a high-dimensional space, thus preferably searched over multiple stages[31].

One benefit of UQ for protein docking results is to determine, for each $\hat{\boldsymbol{x}}_i$ – the prediction in $\mathcal{X}_i$ (the ith structure model),its quality bounds [lb, ub] such that

$$P(lb \leqslant Q(\hat{\boldsymbol{x}}_i, \boldsymbol{x}^*) \leqslant ub) = 1 - \sigma$$

where the quality of interest $Q(\hat{\boldsymbol{x}}, \boldsymbol{x}^*)$ here is iRMSD between a predicted and the native structure and 1-σ is a desired confidence level. Again, $Q(\cdot,\cdot)$ can be any quality assessment

(QA) function that does not necessarily need information about the native structure $\boldsymbol{x}^*$. We used our iRMSD predictor in this study.

To that end, we forwardly propagate the uncertainty from $\boldsymbol{x}^*$ (native structure) to iRMSD, given the final posterior $p(\boldsymbol{x}^*|\mathscr{D}^{(t)})$ in individual regions (local posteriors). Specifically, we generate 1,000,000 samples following the local posterior using Markov chain Monte Carlo, evaluate their binding energies using the kriging regressor, and estimate their iRMSD using our binding affinity prediction formula inversely (as described in Sec. 2.4.2). We then use these sample iRMSD values to determine confidence intervals [*lb, ub*] for various confidence score 1 - σ so that P(iRMSD < *lb*) = *P*(iRMSD > *ub*) = σ/2.

2.6 Confidence scores for near-nativeness

We next calculate the probability that a prediction $\hat{\boldsymbol{x}}_i$ is near-native, i.e., $P(Q(\hat{\boldsymbol{x}}_i, \boldsymbol{x}^*) \leqslant 4\ \text{Å})$[32]. Calculating this quantity would demand the probability that the native structure lies in the *i*th conformational region / cluster, $P(\boldsymbol{x}^* \in \mathscr{X}_i)$ ($P(\mathscr{X}_i)$ in short) as well as that the probability that it lies in all the K regions, $P(\boldsymbol{x}^* \in \cup_{i=1}^{K} \mathscr{X}_i)$ ($P(U_K)$ in short). By following the chain rule we easily reach

$$\begin{aligned} & P(\text{iRMSD}(\hat{\boldsymbol{x}}_i, \boldsymbol{x}^*) \leqslant 4) \\ = \ & P(\text{iRMSD}(\hat{\boldsymbol{x}}_i, \boldsymbol{x}^*) \leqslant 4|\mathscr{X}_i) P(\mathscr{X}_i|U_K) P(U_K) \end{aligned}$$

Here we use the fact that $\left\{\boldsymbol{x}^* \in \mathscr{X}_i\right\} \subset \left\{\boldsymbol{x}^* \in \cup_{i=1}^{K} \mathscr{X}_i\right\}$ and assume that $\{\text{iRMSD}(\hat{\boldsymbol{x}}_i, \boldsymbol{x}^*) \leqslant 4\} \subseteq \{\boldsymbol{x}^* \in \mathscr{X}_i\}$ (the range of conformational clusters in iRMSD is usually wider than 4 Å).

We discuss how to calculate each of the three terms for the product.

2.6.1 $\boldsymbol{P}(\textbf{iRMSD}(\hat{\boldsymbol{x}}_i, \boldsymbol{x}^*) \leqslant 4|\mathscr{X}_i)$**—**If the native structure $\boldsymbol{x}^*$ (unknown) is contained in the ith region/cluster $\mathscr{X}_i$, what is the chance that the predicted structure $\hat{\boldsymbol{x}}_i$ (known)is within 4 Å We again use forward uncertainty propagation starting with the local posterior $p(\boldsymbol{x}^*|\mathscr{D}_t)$ in $\mathscr{X}_i$. We sample 100,000 structures following the posterior with Markov chain Monte Carlo, calculate their iRMSD to the prediction $\hat{\boldsymbol{x}}_i$, and empirically determine the portion within 4 Å for the probability of interest here. Notice that the native interface is unknown thus the putative interface is used instead.

2.6.2 $\boldsymbol{P}(\mathscr{X}_i|U_K)$**—**If the native structure is contained in at least one of the K regions, what is the chance that it is in $\mathscr{X}_i$? Following statistical mechanics, we reach

$$\begin{aligned} P(\mathscr{X}_i|U_K) = & \frac{Z_i}{Z} \\ = & \frac{\int_{\boldsymbol{x} \in \mathscr{X}_i} \exp(-\frac{1}{RT}\hat{f}_i(\boldsymbol{x}))d\boldsymbol{x}}{\sum_{j=1}^{K} \int_{\boldsymbol{x} \in \mathscr{X}_j} (-\frac{1}{RT}\hat{f}_i(\boldsymbol{x}))d\boldsymbol{x}} \end{aligned}$$

where Z_i and Z are local and global partition functions, respectively; and $\widehat{f}_i(\boldsymbol{x})$ is the final kriging regressor. Different regions are assumed to be mutually exclusive. The integrals are calculated by Monte Carlo sampling.

Another approach is to replace $\frac{1}{RT}$ above with ρ_i, the final ρ for temperature control in $\mathcal{X}_i$. In practice we did not find significant performance difference between the two approaches, partly due to the fact that final ρ_i in various clusters / regions reached similar values for the same protein complex.

2.6.3 P(U_K)—What is the chance that the native structure is within the union of the initial regions, i.e., at least one initial region or model is near-native? The way to calculate $P(U_K)$ is very similar to that in uncertainty quantification. Specifically, 100,000 structures are sampled following the posterior of each region $\mathcal{X}_i$, evaluated for binding energy using the kriging regressor $\widehat{f}_i(\boldsymbol{x})$, and estimated with iRMSD using the binding affinity predictor formula inversely. We empirically calculate the portion q_i in which sample iRMSD values are above 4Å. Assuming the independence among regions with regards to near-nativeness, we reach $P(U_K) = 1 - \prod_{i=1}^{K} q_i$. However, if conformational regions $\mathcal{X}_i$, presumed to be separate before search, are overlapping afterwards, $1 - \prod_{i=1}^{K} q_i$ would underestimate $P(U_K)$. One possible approach to address the issue, which could sacrifice optimization, is to introduce constraints on $\mathcal{X}_i$ and keep them separate during search.

2.7 Data sets

We use a comprehensive protein docking benchmark set 4.0[26] of 176 protein pairs that diversely and representatively cover sequence and structure space, interaction types, and docking difficulty. We split them into a training set, test sets a and b with stratified sampling to preserve the composition of difficulty levels in each set. The “training” set is not used for tuning BAL parameters (Sec. 2.2). Rather, it is just for training energy model ($y(\boldsymbol{x};\boldsymbol{w})$ in Sec. 2.4) and conformational-change extent prediction (τ_R in Sec. 2.3.3). The training and test a sets contain 50 and 26 pairs with known K_d values[33], respectively. And the test set **b** contain 100 pairs with K_d values predicted from sequence alone[34].

We also use a smaller yet more challenging CAPRI set of 15 recent CAPRI targets[29]. Unlike the benchmark set for unbound docking, the CAPRI set contains 11 cases of homology docking, 8 of which start with just sequences for both proteins and demand homology models of structures before protein docking. Compared to the benchmark test set of 86 (68%), 22 (18%) and 18 (14%) cases classified rigid, medium, and flexible, respectively; the corresponding statistics for the CAPRI set are 4 (27%), 5 (33%) and 6 (40%), respectively. Their Kd values are also predicted from sequence alone.

The complete lists of the benchmark sets and the CAPRI set, with difficulty classification, are provided in Sec. 2.5 of the SI.

For each protein pair, we use 10 distinct encounter complexes as starting structures ($K = 10$). As reported previously[29], those for the benchmark sets are top-10 cluster representatives by

ZDOCK, kindly provided by the Weng group; and those for the CAPRI set are top-10 models generated by the ZDOCK webserver.

3 Results and Discussion

We briefly summarize our contributions in the Methods section and connect them to individual experiments in the Results section:

- The generic framework of Bayesian active learning for function optimization, posterior estimation, and uncertainty quantification was introduced in Sec. 2.2 and will be assessed over test functions in Sec. 3.1;
- The complex normal modes-based conformational sampling, introduced in Sec. 2.3, was based on our previous study16 and will be used in both PSO and BAL for fair comparison; thus will not be assessed separately;
- The funnel-like energy models for both affinity and quality estimation of encounter complexes was introduced in Sec. 2.4 and their accuracy will be assessed in Sec. 3.2;
- All the contributions in Sec. 2.2–2.4 proposed for optimization in protein docking will be assessed together, using a protein-docking benchmark set and recent CAPRI targets, in Sec. 3.3;
- Quality assessment with uncertainty quantification was introduced for protein docking in Sec. 2.5 and will be assessed in Sec. 3.4;
- Confidence scores for the near-nativeness of each prediction, as well as conditional probabilities of each prediction and each conformational cluster, were introduced in Sec. 2.6 and they will be used for ranking predictions and assessed in Sec. 3.5;
- With all the contributions elaborated in Methods, we will report the overall docking performance (after optimization and UQ-empowered ranking) in Sec. 3.6, examine the causes of success or failures using case studies in Sec. 3.7, and report energy landscapes revealed by BAL in Sec. 3.8.

3.1 Optimization and UQ Performance over Test Functions

We first tested our BAL algorithm on four non-convex test functions of various dimensions and compared it to particle swarm optimization (PSO)[35,36], an advanced optimization algorithm behind a very successful protein-docking method SwarmDock[4]. Detailed settings are provided in Sec. 2.7 of the SI.

For optimization we assess $\|\hat{\boldsymbol{x}} - \boldsymbol{x}^*\|$, the distance between the predicted and actual global optima, a measure of direct relevance to the quality of interest in protein docking – iRMSD. Compared to PSO, BAL made predictions that are, on average, closer to the global optima with smaller standard deviations (except for 2D Griewank where BAL had larger standard deviation); and the improvement margins increased with the increasing dimensions (Fig. 2 and Table S8).

For quality assessment with UQ we assess r90, the distance upper bound of 90% confidence, i.e., $P(||\hat{\boldsymbol{x}} - \boldsymbol{x}^*|| \leqslant r_{90}) = 1 - \sigma = 90\%$. The metrics to assess r_{90} include η, the relative error ($\eta = |\frac{r_{90}}{||\hat{\boldsymbol{x}} - \boldsymbol{x}^*||} - 1|$); and $\widehat{P}$, the portion of the confidence intervals from 100 runs that actually encompass the corresponding global optimum. We found in Table S9 that our confidence intervals are usually tight judging from η and they contain the global optima with portions $\widehat{P}$ close to 90%, the desired confidence level. The portions agreed less with the desired confidence level for some functions as the dimensionality increase, which suggests the challenge of optimization and UQ in higher dimensions.

The rest of the results are on protein docking, which presents more challenges in objective function, feasible set, and more, compared to the aforementioned test functions.

3.2 Evaluation of Models for Energy and Quality Estimation

We compare the performances of three machine learning models over the training and test sets for energy model (Sec. 2.4.2). As no actual binding affinities of encounter complexes are available, we estimated the iRMSD values based on the random forest model's binding energy prediction (Sec. 2.4.1) and compared them to the actual iRMSD (of native interfaces) using RMSE for absolute error. Random forest gave the best performances thus used as the energy model y(x) hereinafter. Specifically, performances are split to encounter complexes of varying quality (iRMSD) in Fig. 3A and Fig. 3B. The random forest model (blue bars) led to RMSE of 0.70 Å (1.0 Å) for the near-native samples in the training (test) set. The RMSEs increased slowly as iRMSD ≤ 10 Å and did sharply beyond (a region too far from the native for refinement), which matches our design rationale to focus on energy model accuracy, and as a result uncertain quantification, in the lower iRMSD region.

We also assess how "funnel-like" the energy model is. We thus calculated for each protein pair the Spearman's ranking coefficient ρ between the energy model and the actual iRMSD. The random forest of MM-GBSW features showed the highest ρ of 0.72 and 0.60 for the training and the test sets, respectively (Fig. 3C), albeit with large deviation across protein pairs.

We lastly assess the energy model's ability to rank across protein pairs. Specifically, we estimated each native protein-complex's binding energy by setting iRMSD to be zero in the energy model and compared the estimated and actual binding energy using RMSE and Pearson's r in the supplemental Table S11. The random-forest energy model achieved 2.45 (4.78) Kcal/mol in RMSE and Pearson's r of 0.79 (0.75) in binding energy $\delta G = -RTln(K_d)$ for the training (test) set.

3.3 Docking Performance: Optimization

We show the improvements in PSO and BAL solution quality (measured by the decrease of iRMSD) against the starting ZDOCK solutions in Fig. S4 of the Supplemental Material. Speaking of the amount of improvement, BAL improved iRMSD by 1.2 Å, 0.74 Å, and 0.76 Å for the training, test, and CAPRI sets, respectively, outperforming PSO's corresponding measures of 0.82 Å, 0.45 Å, and 0.49 Å. It also outperformed PSO for the more challenging near-native cases (note that BAL's iRMSD improvement for the near-native test set or

CAPRI set was almost neutral). Speaking of the portion with improvement, BAL improved iRMSD in the near-native cases for 75%, 68%, and 73% of the training, test, and CAPRI sets, respectively; whereas the corresponding statistics for PSO were 59%, 50%, and 53%, respectively. More split statistics based on docking difficulty can be found in Fig. S5.

We also compared BAL and PSO solutions head-to-head over subsets of varying difficulty levels for protein docking (Fig. 4). Overall, BAL's solutions are better (or significantly better by at least 0.5 Å) than those of PSO for 70%–80% (31%–45%) of the cases, which was relatively insensitive to the docking difficulty level.

Both PSO and BAL use a single trajectory of 31 iterations and 630 samples for each region/cluster. Most time is on local structure minimization and energy evaluations using CHARMM[22]. The BAL running time for optimization of each cluster thus almost linearly grows with the size of the protein pair (Fig. S8), ranging from 7 hours for a 200-residue complex to 13 hours for a 1700-residue one.

3.4 Docking Performance: Quality Assessment with Uncertainty Quantification

We next assess the solution-quality UQ results for protein docking. Similar to that for test functions, we assess $r_{1-\sigma}$, the half length of (1 - σ) confidence interval [*lb, ub*], i.e., $P(lb \leqslant \text{iRMSD}(\widehat{\boldsymbol{x}}, \boldsymbol{x}^*) \leqslant ub) = 1 - \sigma$. The metrics to assess $r_{1-\sigma}$ include η, the relative error ($\eta = |\frac{r_{1-\sigma}}{\text{iRMSD}(\widehat{\boldsymbol{x}}, \boldsymbol{x}^*)} - 1|$); and $\widehat{P}$, the portion of the confidence intervals that actually contain the corresponding native structure across all docking runs (10 for 10 models of each protein pair in each set).

Table 1 shows that the portions matched well with the confidence levels over all four data sets. Test set b and the CAPRI set did not have actual K_d values available and were thus impacted further by the uncertainty of K_d prediction, although the impact did not appear significant. There was a trade off between the confidence level and the length of the confidence interval, as narrower confidence intervals (with less η) corresponded to lower confidence levels. A balance seems to be at the 85% confidence level where the relative iRMSD uncertainty is around 25%.

3.5 Scoring Performance Empowered by UQ

For scoring models or predictions, two metrics are used for assessing the performance. The first is Spearman's ρ for ranking protein-docking predictions (structure models) for each pair. The second is the area under the Precision Recall Curve (AUPRC), for the binary classification of each prediction being near-native or not. Considering that the near-natives are minorities among all predictions, AUPRC is a more meaningful measure than the more common AUROC.

With these two metrics we assess four scoring functions on predictions $\widehat{\boldsymbol{x}}_i$: (1) $\Delta E(\widehat{\boldsymbol{x}}_i)$, the MM-GBSW binding energy, i.e., the sum of the 8 features; (2) our random-forest energy model $y(\widehat{\boldsymbol{x}}_i)$; (3) $P(X_i|U_K)$, the conditional probability that the ith prediction's region is near-native give that there is at least such one in the top K predictions; and (4) $P(\text{iRMSD}(\widehat{\boldsymbol{x}}_i, \boldsymbol{x}^*) \leqslant 4)$, the unconditional probability that the ith prediction is near-native.

For ranking assessment, from Fig. 5 we find that, whereas the original MM-GBSW model achieved merely 0.2 for Spearman's ρ, our energy model using the same 8 terms as features in random forest drastically improved the ranking performance with a Spearman's ρ around 0.6 for training, benchmark test, and CAPRI test sets. Furthermore, the confidence scores $P(\mathcal{X}_i|U_K)$ and $P(\text{iRMSD}(\hat{\boldsymbol{x}}_i, \boldsymbol{x}^*) \leqslant 4)$ further improved ranking. In particular, the unconditional probability for a prediction to be near-native achieved around 0.70 in ρ even for the benchmark and the CAPRI test sets. Note that this probability, a confidence score on the prediction's near-nativeness, was derived from the posterior distribution of $\boldsymbol{x}^*$; thus it uses both enthalpic and entropic contributions.

For binary assessment on classifying the nativeness of predictions, the test set is split into **a**, 26 pairs with known K_d values and **b**, 100 with predicted ones. From Table 2 we conclude that the MM-GBSW energy model performed close to random (AUROC close to 0.5) and the random forest energy model using the same features drastically improved AUROC to around 0.8 and AUPRC 0.54 ~ 0.62 across sets. Since AUROC is uninformative for highly imbalanced data (for instance, near-native predictions are 14% over all data sets), we focus on AUPRC. The next three probabilities from our BAL's confidence scores improved the AUPRC to nearly 0.80 for the training and above 0.60 for test sets. The additional uncertainty in K_d prediction from test sets **b** and CAPRI did not noticeably impact the performance compared to test set **a**.

3.6 Overall Docking Performance

We summarize our docking results (BAL predictions $\boldsymbol{x}_i$ ranked by confidence scores on their nativeness $P(\text{iRMSD}(\hat{\boldsymbol{x}}_i, \boldsymbol{x}^*) \leqslant 4)$) in Table 3; and compare them to the ZDOCK starting results (ranked by cluster size roughly reflecting entropy)and the PSO refinement results (using the same energy model as BAL and ranked by the energy model). We use N_K to denote the number of targets with at least one near-native predictions in top K; and F_K the fraction of such targets among all in a given set (training, benchmark test, or CAPRI test set). Compared to the ZDOCK starting results and PSO refinements, BAL has improved the portion of acceptable targets with top-3 predictions from 23% and 26%, respectively, to 32% for the benchmark test set. Similar improvements were found for the CAPRI set. The portion for top 10 from BAL reached 40% compared to ZDOCK's 33% over the benchmark test set. Note that BAL only refined top-10 starting results from ZDOCK thus this improvement was purely from optimization (no ranking effect).

We further visualize the test-set performance along with iRMSD estimation and UQ-derived confidence scores $P(\text{iRMSD}(\hat{\boldsymbol{x}}_i, \boldsymbol{x}^*) \leqslant 4))$. Fig. 6A shows the actual versus the estimated iRMSD of the top-1 prediction for each test target; and Fig. 6B shows the actual iRMSD versus the confidence score of each such prediction. These predictions are colored according to the quality (iRMSD)of the starting structure from ZDOCK. Predicted iRMSD values were positively correlated with actual values and rarely above 4 Å for near-native predictions. High confidence scores were almost exclusive to good predictions with low iRMSD values whereas low confidence scores corresponded to a mixture of qualities.

3.7 Case Studies

To examine the contributions of method components, such as energy model, quality estimation (iRMSD), optimization, and uncertainty quantification, we chose one successful and two failed cases for detailed analysis.

In the case of success (PDB: 1FFW, ZDOCK model 6), BAL improved prediction quality (iRMSD) from 3.4 Å to 2.2 Å and predicted a very close iRMSD of 2.5 Å. The 90% confidence interval (C. I.) in iRMSD, being [1.51 Å, 3.15 Å], contains the actual iRMSD of 2.2 Å. We project the 12D posterior $p(\boldsymbol{x}^*|\mathscr{D}(t))$ at the end of iteration t onto a 1D distribution in iRMSD$(\boldsymbol{x}^*, \widehat{\boldsymbol{x}}^{(t)})$ where $\widehat{\boldsymbol{x}}^{(t)}$ denotes the prediction at the end of iteration t. Fig. 7 shows that the predicted and the actual iRMSD values were both contained in the 90% confidence interval through iterations and both converged to the peak of the narrower posterior as iterations progressed.

In a failed case (PDB: 1QFW IM:AB; ZDOCK model 6), BAL did improve docking quality by reducing actual iRMSD by 0.23 Å compared to the starting ZDOCK model. The predicted and the actual iRMSD of iterative predictions wer e also close, as seen in Fig. 8. However, uncertainty quantification failed as the 90% confide nce interval in iRMSD didn't encompass the predicted or the actual iRMSD. This failure is attributed to the energy model, as the lowest-energy prediction $\widehat{\boldsymbol{x}}^{(t)}$ was often far from the most-probable conformation (where the peak of the posterior is) and the search drifted toward a non-native funnel. We also note that predictions were of worse quality (larger iRMSD) over iterations.

In another failed case (PDB: 2UUY; ZDOCK model 8), BAL failed to improve docking quality compared to the starting structure (iRMSD increased from 3.74 Å to 3.85Å). A close look at Fig. 9 suggests that the predictions actually improved over iterations (judging from the real iRMSD in red lines). However, even though the posteriors became narrower over iterations and predicted iRMSD values were close to the peaks of the posteriors, the actual iRMSD values were way off the predicted; and even outside the 90% confidence interval ([1.41 Å, 2.45 Å]). iRMSD prediction (quality estimation) is thus a major reason behind this failure.

3.8 Energy Landscapes and Association Pathways

We lastly investigate energy landscapes during BAL sampling. Our kriging regressor $\widehat{f}(\boldsymbol{x})$ is an unbiased estimator and works even better than the noisy observations from the random-forest energy model $y(\boldsymbol{x})$. Energy landscapes are visualized for 37 near-native regi ons for the benchmark test set (Fig. S6 in the Supplementary Material) and for the CAPRI set (Fig. S7) (non-rigid cases only). Two examples are shown in Fig. 10, depicting a (multiple) funnel-like energy landscape with a clear association paths from the starting to the end or the native complex along gradient descents. Three more supplemental videos are provided to visualize the BAL sampling trajectories using protein structures.

4 Conclusions

We present the first uncertainty quantification (UQ) study for protein docking. This is accomplished by a rigorous Bayesian framework that actively samples a noisy and expensive black-box function (i.e., collecting data $\mathcal{D}$) while updating a posterior distribution $p(\boldsymbol{x}^*|\mathcal{D})$ directly over the unknown global optimum $\boldsymbol{x}^*$. The iterative feedback between Thompson sampling and posterior updating is linked by a Boltzmann distribution with adaptive annealing schedule and non-parametric kriging regressor. The inverse uncertainty quantification on the location of the global optimum can easily forward-propagate for the uncertainty quantification of any quality of interest as a function of the global optimum, including the interface RMSD that measures dissimilarity between protein-docking solutions and native structures.

We demonstrate the superb performances of Bayesian active learning (BAL) on a protein docking benchmark set as well as a CAPRI set full of homology docking. Compared to the starting points from initial rigid docking as well as the refinement from PSO, BAL shows significant improvement, accomplishing a top-3 near-native prediction for about one-third of the benchmark and CAPRI sets. Its UQ results achieve tight uncertainty intervals whose radius is 25% of iRMSD with a 85% confidence level attested by empirical results. Moreover, its estimated probability of a prediction being near-native achieves an AUROC over 0.93 and AUPRC over 0.60 (more than 4 times over random classification).

Besides the optimization and UQ algorithms, other contributions specific to protein docking build on and advance the state of the art, especially those studies addressing the challenges of modeling conformational changes,[4,16,29,37] constructing energy models,[3] investigating advanced sampling strategies,[4,5,38] and studying these coupled factors altogether.[3,7,39] We for the first time represent the conformational space for protein docking as a flat Euclidean space spanned by complex normal modes blending flexible- and rigid-body motions and anticipating protein conformational changes, a homogeneous and isotropic space friendly to high-dimension optimization. We also construct a funnel-like energy model using machine learning to associate binding energies of encounter complexes sampled in docking with their iRMSD. These innovations also contribute to the excellent performances of BAL; and lead to direct visualization of binding energy funnels and protein association pathways in conformational degrees of freedom. Looking ahead, there is still much room toward addressing aforementioned challenges for *ab initio*, flexible protein docking that can be both fast and accurate.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Dr. Thom Vreven and Dr. Zhiping Weng for the ZDOCK rigid-docking decoys for the benchmark set, Haoran Chen for helping on cNMA, and Yuanfei Sun for proofreading the manuscript. We also thank anonymous reviewers for insightful comments. Part of the CPU time was provided by the Texas A&M High Performance Research Computing.

Funding

This work was supported by the National Institutes of Health (R35GM124952) and the National Science Foundation (CCF-1546278).

References

(1). Mosca R; Céol A; Aloy P Interactome3D: adding structural details to protein networks. Nat. Methods 2013, 10, 47. [PubMed: 23399932]

(2). Smith GR; Sternberg MJ Prediction of protein–protein interactions by docking methods. Curr. Opin. Struct. Biol. 2002, 12, 28–35. [PubMed: 11839486]

(3). Gray JJ; Moughon S; Wang C; Schueler-Furman O; Kuhlman B; Rohl CA; Baker D Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J. Mol. Biol. 2003, 331, 281–299. [PubMed: 12875852]

(4). Moal IH; Bates PA SwarmDock and the Use of Normal Modes in Protein-Protein Docking. Int. J. Mol. Sci. 2010, 11, 3623–3648. [PubMed: 21152290]

(5). Shen Y Improved flexible refinement of protein docking in CAPRI rounds 22–27. Proteins: Struct., Funct., Bioinf. 2013, 81, 2129–2136.

(6). Jiménez-García B; Roel-Touris J; Romero-Durana M; Vidal M; Jiménez-González D; Fernández-Recio J LightDock: a new multi-scale approach to protein–protein docking. Bioinformatics 2017, 34, 49–55.

(7). Marze NA; Roy Burman SS; Sheffler W; Gray JJ Efficient flexible backbone protein–protein docking for challenging targets. Bioinformatics 2018, 34, 3461–3469. [PubMed: 29718115]

(8). Pfeiffenberger E; Bates PA Refinement of protein-protein complexes in contact map space with metadynamics simulations. Proteins: Struct., Funct., Bioinf 2019, 87, 12–22.

(9). Rudden LSP; Degiacomi MT Protein Docking Using a Single Representation for Protein Surface, Electrostatics, and Local Dynamics. J Chem Theory Comput 2019, 15, 5135–5143. [PubMed: 31390206]

(10). Der Kiureghian A; Ditlevsen O Aleatoric or epistemic? Does it matter? Structural Safety 2009, 31, 105–112.

(11). Li W; Schaeffer RD; Otwinowski Z; Grishin NV Estimation of Uncertainties in the Global Distance Test (GDT TS) for CASP Models. PLoS One 2016, 11, e0154786. [PubMed: 27149620]

(12). Rasheed M; Clement N; Bhowmick A; Bajaj C Statistical Framework for Uncertainty Quantification in Computational Molecular Modeling. ACM Conference on Bioinformatics, Computational Biology and Biomedicine Seattle, 2016; p 146.

(13). Shahriari B; Swersky K; Wang Z; Adams RP; Freitas N. d. Taking the Human Out of the Loop: A Review of Bayesian Optimization. Proceedings of the IEEE 2016, 104, 148–175.

(14). Ortega P; Grau-Moya J; Genewein T; Balduzzi D; Braun D A nonparametric conjugate prior distribution for the maximizing argument of a noisy function. Advances in Neural Information Processing Systems. Lake Tahoe, 2012; pp 3005–3013.

(15). Shen Y; Paschalidis IC; Vakili P; Vajda S Protein docking by the underestimation of free energy funnels in the space of encounter complexes. PLoS Comput. Biol. 2008, 4, e1000191. [PubMed: 18846200]

(16). Oliwa T; Shen Y cNMA: a framework of encounter complex-based normal mode analysis to model conformational changes in protein interactions. Bioinformatics 2015, 31, i151–i160. [PubMed: 26072477]

(17). Ingber L Adaptive simulated annealing (ASA): Lessons learned. CoRR 2000, cs.MS/0001018.

(18). Chilès JP; Delfiner P Geostatistics: Modeling Spatial Uncertainty, 2nd Edition; 2012.

(19). Russo DJ; Van Roy B; Kazerouni A; Osband I; Wen Z, et al. A tutorial on thompson sampling. Fou. and Tre. in Mach. Learn. 2018, 11, 1–96.

(20). Agrawal S; Goyal N Analysis of thompson sampling for the multi-armed bandit problem. Conference on learning theory Edinburgh, 2012; pp 39–1.

(21). Chapelle O; Li L An empirical evaluation of thompson sampling. Advances in Neural Information Processing Systems. Granada, 2011; pp 2249–2257.

(22). Brooks BR; Brooks CL; Mackerell AD; Nilsson L; Petrella RJ; Roux B; Won Y; Archontis G; Bartels C; Boresch S; Caflisch A; Caves L; Cui Q; Dinner AR; Feig M; Fischer S; Gao J; Hodoscek M; Im W; Kuczera K; Lazaridis T; Ma J; Ovchinnikov V; Paci E; Pastor RW; Post CB; Pu JZ; Schaefer M; Tidor B; Venable RM; Woodcock HL; Wu X; Yang W; York DM; Karplus M CHARMM: the biomolecular simulation program. J. Comput. Chem. 2009, 30, 1545–1614. [PubMed: 19444816]

(23). Györfi L; Kohler M; Krzyzak A; Walk H A Distribution-Free Theory of Nonparametric Regression; Springer Series in Statistics; Springer-Verlag: New York, 2002.

(24). Villemonteix J; Vazquez E; Walter E An informational approach to the global optimization of expensive-to-evaluate functions. Journal of Global Optimization 2009, 44, 509.

(25). Hernández-Lobato JM; Hoffman MW; Ghahramani Z Predictive entropy search for efficient global optimization of black-box functions. Advances in Neural Information Processing Systems. Montreal, 2014; pp 918–926.

(26). Hwang H; Vreven T; Janin J; Weng Z Protein-Protein Docking Benchmark Version 4.0. Proteins 2010, 78, 3111–3114. [PubMed: 20806234]

(27). Shen Y; Vakili P; Vajda S; Paschalidis IC Optimizing noisy funnel-like functions on the Euclidean group with applications to protein docking. IEEE Conference on Decision and Control. New Orleans, 2007; pp 4545–4550.

(28). Mirzaei H; Zarbafian S; Villar E; Mottarella S; Beglov D; Vajda S; Paschalidis IC; Vakili P; Kozakov D Energy minimization on manifolds for docking flexible molecules. J. Chem. Theory Comput. 2015, 11, 1063–1076. [PubMed: 26478722]

(29). Chen H; Sun Y; Shen Y Predicting protein conformational changes for unbound and homology docking: learning from intrinsic and induced flexibility. Proteins: Struct., Funct., Bioinf. 2017, 85, 544–556.

(30). Kozakov D; Clodfelter KH; Vajda S; Camacho CJ Optimal clustering for detecting near-native conformations in protein docking. Biophys. J. 2005, 89, 867–875. [PubMed: 15908573]

(31). Vajda S; Kozakov D Convergence and combination of methods in protein–protein docking. Curr. Opin. Struct. Biol. 2009, 19, 164–170. [PubMed: 19327983]

(32). Méndez R; Leplae R; Lensink MF; Wodak SJ Assessment of CAPRI predictions in rounds 3–5 shows progress in docking procedures. Proteins: Struct., Funct., Bioinf 2005, 60, 150–169.

(33). Kastritis PL; Bonvin AMJJ Are Scoring Functions in Protein-Protein Docking Ready To Predict Interactomes? Clues from a Novel Binding Affinity Benchmark. J. Proteome Res. 2010, 9, 2216–2225. [PubMed: 20329755]

(34). Yugandhar K; Gromiha MM Protein–protein binding affinity prediction from amino acid sequence. Bioinformatics 2014, 30, 3583–3589. [PubMed: 25172924]

(35). Kennedy J; Eberhart R Particle swarm optimization. Proceedings of ICNN'95-International Conference on Neural Networks Perth, 1995; pp 1942–1948.

(36). Clerc M; Kennedy J The particle swarm-explosion, stability, and convergence in a multidimensional complex space. IEEE transactions on Evolutionary Computation 2002, 6, 58–73.

(37). Venkatraman V; Ritchie DW Flexible protein docking refinement using pose-dependent normal mode analysis. Proteins: Struct., Funct., Bioinf 2012, 80, 2262–2274.

(38). Kuroda D; Gray JJ Pushing the Backbone in Protein-Protein Docking. Structure 2016, 24, 1821–1829. [PubMed: 27568930]

(39). Moal IH; Chaleil RAG; Bates PA Flexible Protein-Protein Docking with SwarmDock. Methods Mol. Biol 2018, 1764, 413–428. [PubMed: 29605931]

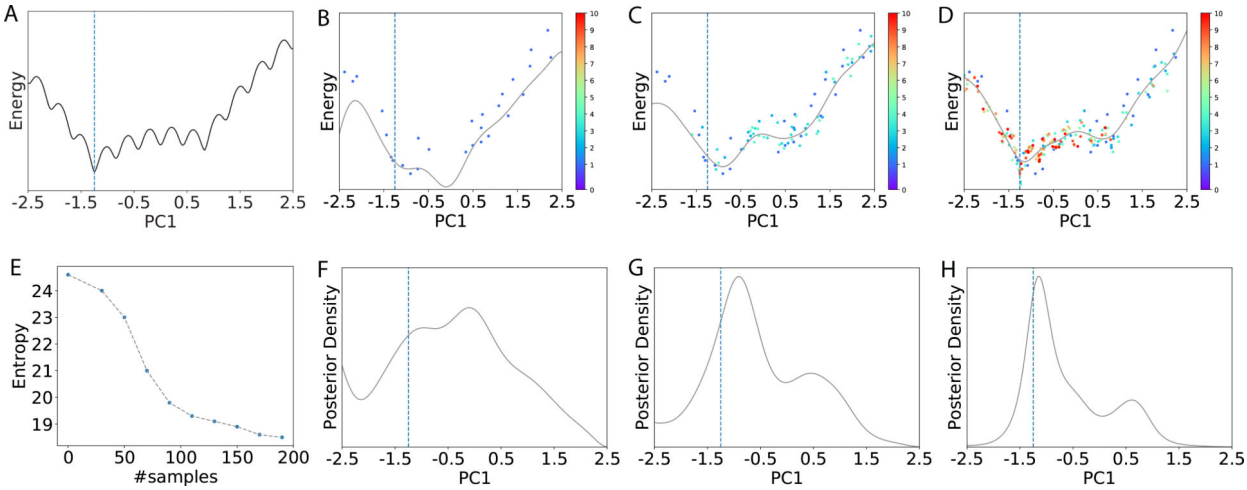


Figure 1:

Illustration of the Bayesian active learning (BAL) algorithm. (A): A typical energy landscape projected onto the first principal component (PC1) just for visualization, using all samples collected. The dashed line indicates the location of the optimal solution. (B)-(D): The samples (dots) and the kriging regressors (light curves) in the 1st, 4th and 10th iteration, respectively. Samples are colored from cold to hot for increasing iteration indices and those in the same iteration have the same color. (E): The entropy (measuring uncertainty) of the posterior reduces as the number of samples increases. Its quick drop, as the number of samples increases from 30 to 100, corresponds to a drastic change of the kriging regressor, which suggests increasing exploitation in possible function basins. After 100 samples, the entropy goes down more slowly, echoing the smaller updates of the regressor. (F)-(H): The corresponding posterior distributions for (B)-(D), respectively.

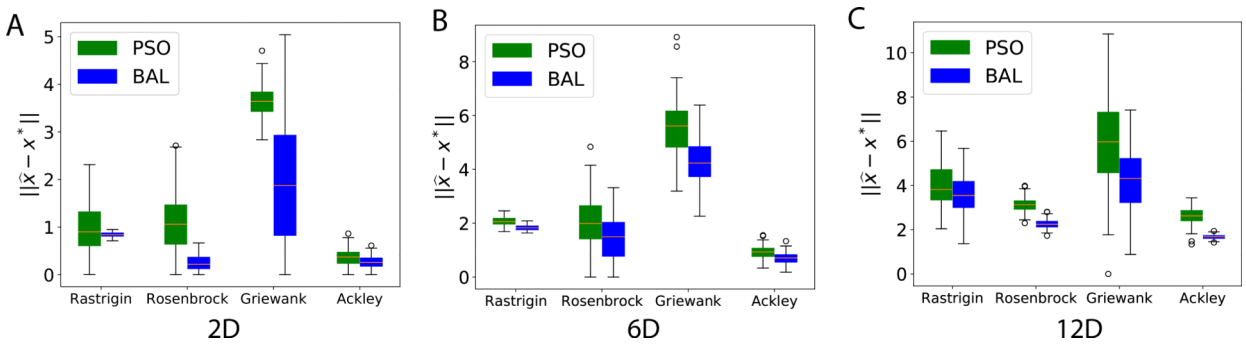


Figure 2:

Optimization performances of PSO and BAL over four non-convex test functions in various dimensions. The performance metric is $\|\hat{\boldsymbol{x}} - \boldsymbol{x}^*\|$, the distance between the predicted and the actual global optima, and the box plot is generated using 100 optimization trajectories in each case.

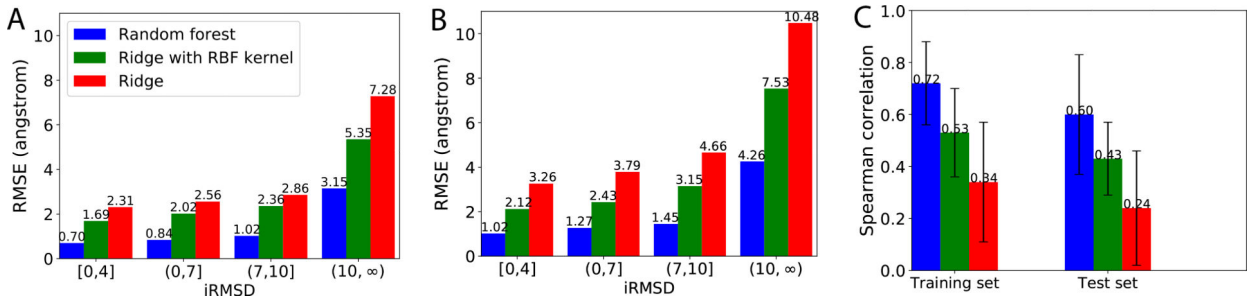


Figure 3:

The root mean square error (RMSE) between the predicted and actual iRMSD for (A): Training set and (B): Test set. (C): The Spearman's ρ between predicted $y(\boldsymbol{x})$ and the real iRMSD for the training and test sets.

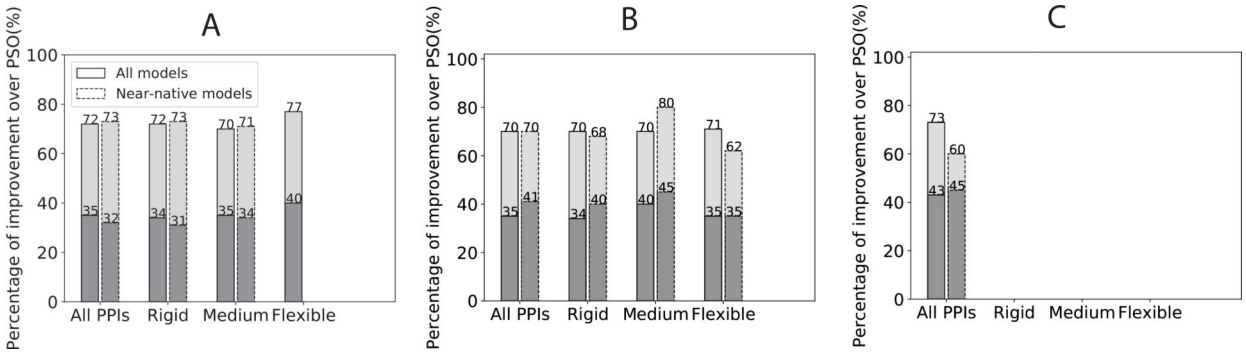


Figure 4:

The percentage of BAL predictions with iRMSD improvement against PSO for A. the training set, B. the benchmark test set, and C. the CAPRI set. The CAPRI set is not further split because it only contains 15 targets and is predominantly in the flexible category. The darker gray portions correspond to significant improvement (over 0.5 Å in iRMSD) compared to corresponding PSO predictions.

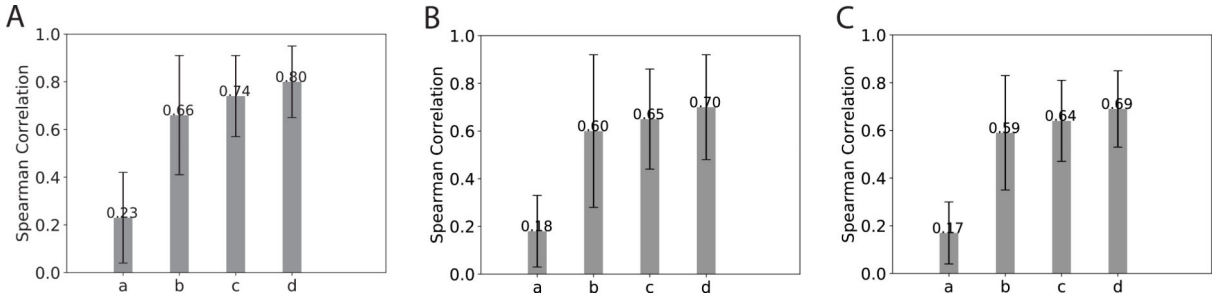


Figure 5:

Ranking performance shown in the bar plot (with error bar in black) of Spearman's ρ for (A) Training set, (B) Test set and (C) CAPRI set, respectively. Scoring function a, b, c, and d in each figure correspond to the MM-GBSW model, our random-forest energy model, and confidence scores $P(\mathcal{X}_i|U_K)$, and $P(\text{iRMSD}(\hat{\boldsymbol{x}}_i, \boldsymbol{x}^*) \leqslant 4)$, respectively.

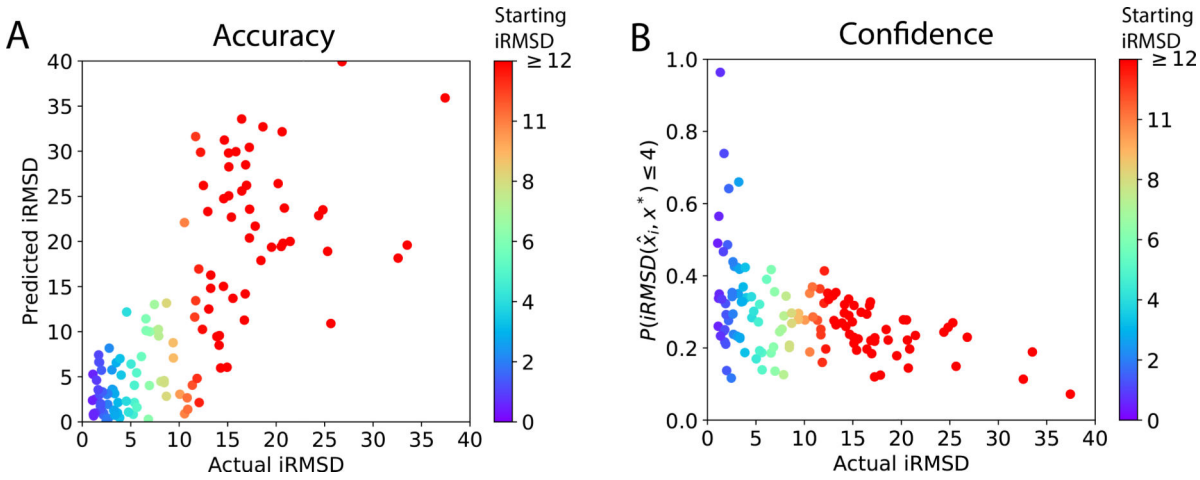


Figure 6:

Actual iRMSD of the top-1 predictions versus predicted iRMSD or confidence scores in the benchmark test set. Each point representing the prediction for a target is colored according to the starting structure's quality (iRMSD value).

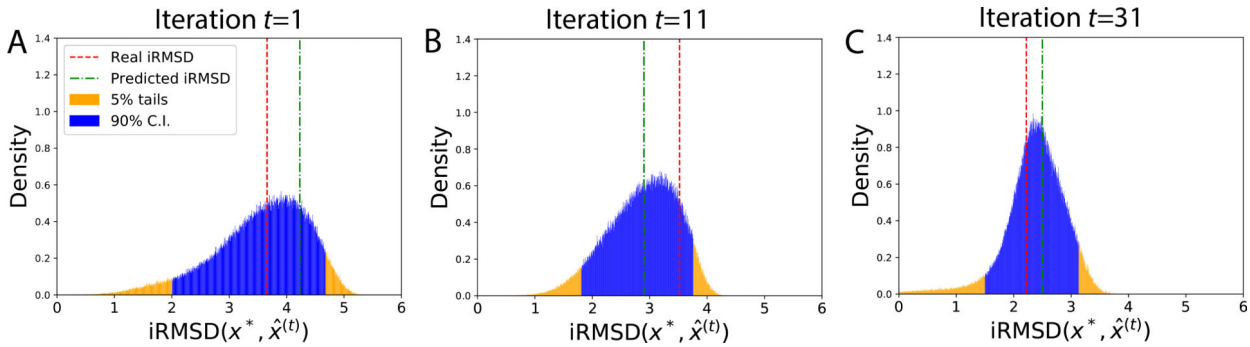


Figure 7:

A successful case (PDB: 1FFW; ZDOCK model 6): The po sterior distributions of the native structure $\boldsymbol{x}^*$ in its iRMSD to predictions $\hat{\boldsymbol{x}}^{(t)}$ over iterations t.

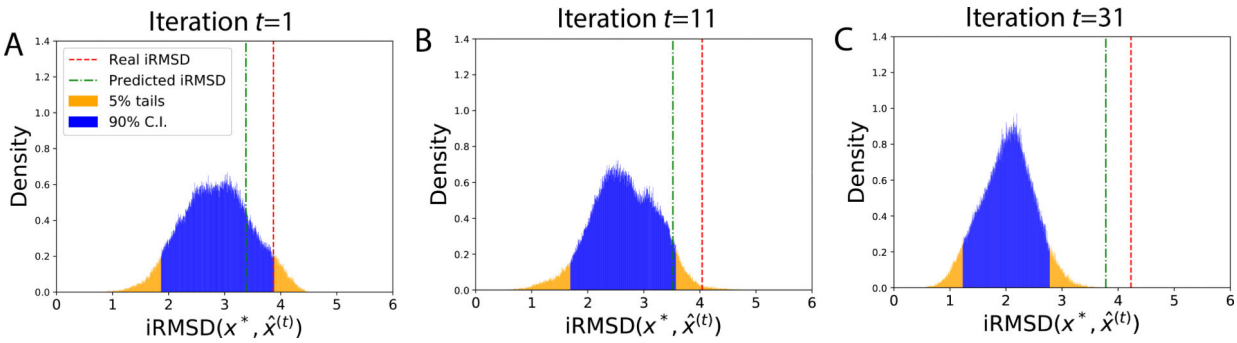


Figure 8:

A case with failure in UQ (PDB: 1QFW_IM:AB; ZDOCK model 6). The posterior distributions of the native structure in iRMSD did not encompass actual or predicted iRMSD in their 90% confidence intervals. The predictions, although better than the starting ZDOCK model, actually became worse over iterations, which is likely driven by a low-energy non-native funnel in the energy model.

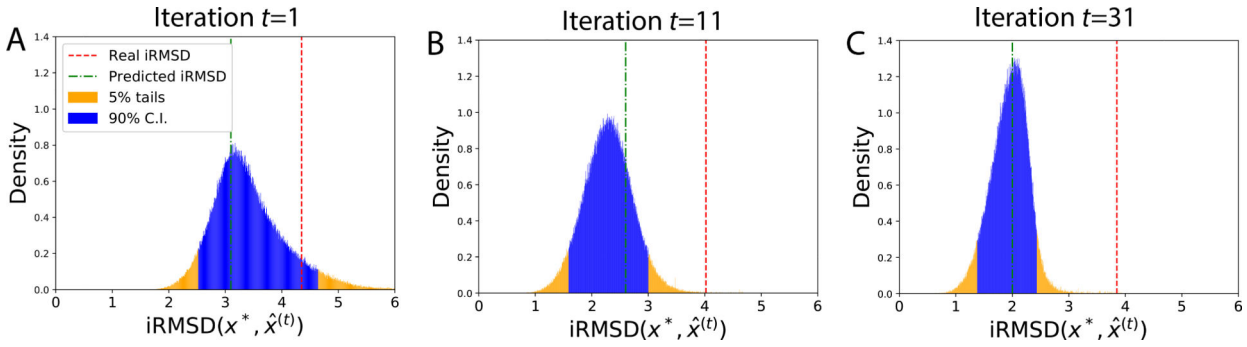


Figure 9:

A case with failure in optimization and UQ (PDB: 2UUY; ZDOCK model 8). The posterior distributions of the native structure in iRMSD did encompass predicted iRMSD, but not actual iRMSD, in their 90% confidence intervals. The predictions actually became better over iterations but did not improve over the starting ZDOCK model.

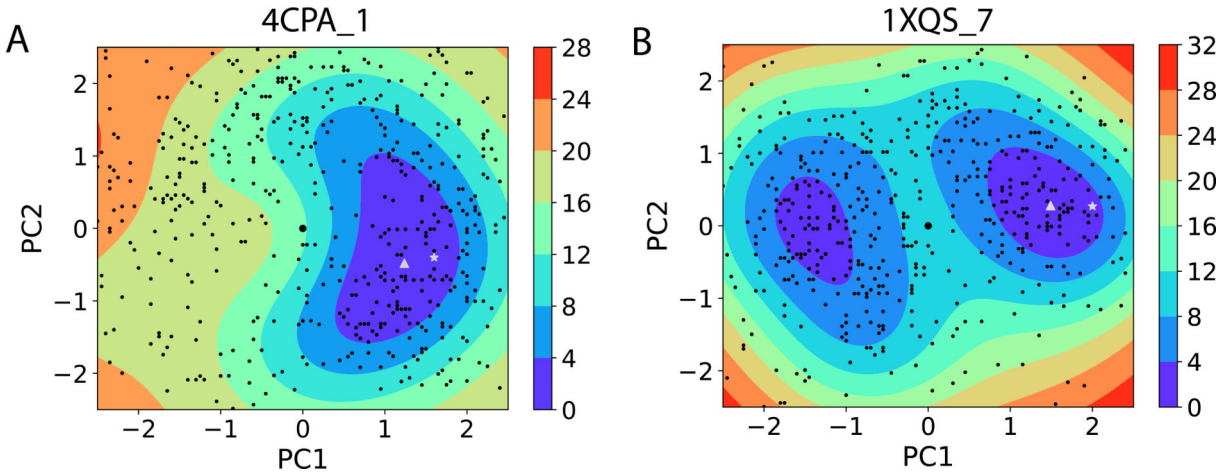


Figure 10:

The estimated energy landscapes along the first two principal components (PC) for two medium-difficulty docking cases with near-native starting ZDOCK models. The starting structures, BAL samples, lowest-energy predictions, and ground-truth native structures are represented as large black dots at the origin, black dots, white triangles, and white stars, respectively. All the color-coding energy values are in the unit of RT and relative to the lowest sample energy value within each region.

Table 1:

Uncertainty quantification performances of BAL on protein docking based on η, the relative error in iRMSD; and $\widehat{P}$, the portion of confidence intervals from 100 runs encompassing the global optima. For η, means (and standard deviations in parentheses) are reported.

Dataset	$1-\sigma$	0.99	0.95	0.90	0.85	0.80
Training	η	0.40 (0.23)	0.35 (0.18)	0.31 (0.17)	0.27 (0.15)	0.22 (0.10)
	$\widehat{P}$	0.97	0.91	0.84	0.79	0.75
Test **a**	η	0.43 (0.26)	0.39 (0.21)	0.28 (0.16)	0.25 (0.13)	0.21 (0.09)
	$\widehat{P}$	0.95	0.87	0.83	0.75	0.73
Test **b**	η	0.44 (0.22)	0.38 (0.16)	0.26 (0.14)	0.23 (0.10)	0.19 (0.08)
	$\widehat{P}$	0.91	0.84	0.80	0.74	0.71
CAPRI	η	0.43 (0.20)	0.35 (0.13)	0.27 (0.11)	0.22 (0.10)	0.20 (0.09)
	$\widehat{P}$	0.91	0.85	0.81	0.74	0.70

Table 2:

Binary assessment of our 5 scoring functions on a prediction $\hat{\boldsymbol{x}}_i$ being near-native or not: MM-GBSW energy model $\Delta E(\hat{\boldsymbol{x}}_i)$, random-forest energy model $y(\hat{\boldsymbol{x}}_i)$, and 3 BAL-determined probabilities that a region/cluster $\mathcal{X}_i$ is near-native given anative-containing list, the prediction $\hat{\boldsymbol{x}}_i$ in that region is near-native given a native-containing list, or $\hat{\boldsymbol{x}}_i$ is near-native.

Dataset	Assessment	$\Delta E(\hat{\boldsymbol{x}}_i)$	$y(\hat{\boldsymbol{x}}_i)$	$P(\mathcal{X}_i \mid \boldsymbol{U_K})$	$P(\text{iRMSD}(\hat{\boldsymbol{x}}_i, \boldsymbol{x}^*) \leqslant 4 \mid \boldsymbol{U_K})$	$P(\text{iRMSD}(\hat{\boldsymbol{x}}_i, \boldsymbol{x}^*) \leqslant 4)$
Training	AUROC	0,489	0.806	0.903	0.944	0.967
	AUPRC	0.241	0.624	0.684	0.771	0.796
Test **a**	AUROC	0.460	0.810	0.892	0.929	0.939
	AUPRC	0.199	0.550	0.592	0.613	0.634
Test **b**	AUROC	0.490	0.789	0.847	0.898	0.927
	AUPRC	0.203	0.540	0.571	0.609	0.615
CAPRI	AUROC	0.491	0.771	0.844	0.893	0.919
	AUPRC	0.214	0.561	0.600	0.610	0.614

Table 3:

Summary of docking results measured by the number and the portion of targets in each set that have an acceptable near-native top-1, 3, 5, or 10 prediction.

	ZDOCK (Starting Point)				PSO				BAL			
Dataset (size)	N_1 (F_1)	N_3 (F_3)	N_5 (F_5)	N_{10} (F_{10})	N_1 (F_1)	N_3 (F_3)	N_5 (F_5)	N_{10} (F_{10})	N_1 (F_1)	N_3 (F_3)	N_5 (F_5)	N_{10} (F_{10})
Training (50)	6 (12%)	11 (22%)	13 (26%)	17 (34%)	9 (18%)	15 (30%)	17 (34%)	20 (40%)	14 (28%)	19 (38%)	20 (40%)	22 (44%)
Test (126)	20 (16%)	29 (23%)	33 (26%)	41 (33%)	26 (21%)	33 (26%)	39 (31%)	45 (36%)	32 (25%)	40 (32%)	42 (33%)	50 (40%)
CAPRI (15)	2(13%)	2 (13%)	3 (20%)	4 (27%)	3 (20%)	3 (20%)	4 (27%)	5 (33%)	3 (20%)	4 (27%)	5 (33%)	5 (33%)