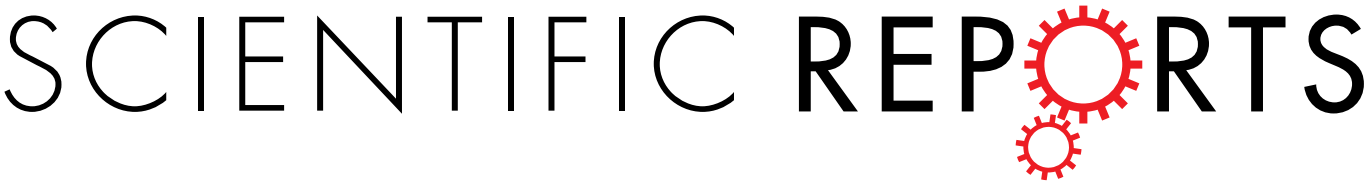

OPEN

Protein docking refinement by convex underestimation in the low-dimensional subspace of encounter complexes

Received: 18 September 2017
Accepted: 21 March 2018
Published online: 12 April 2018

Shahrooz Zarbafian[2], Mohammad Moghadasi[1], Athar Roshandelpoor[1], Feng Nan[1], Keyong Li[1], Pirooz Vakili[1,2], Sandor Vajda[3], Dima Kozakov[5] & Ioannis Ch. Paschalidis[1,3,4,6]

We propose a novel stochastic global optimization algorithm with applications to the refinement stage of protein docking prediction methods. Our approach can process conformations sampled from multiple clusters, each roughly corresponding to a different binding energy funnel. These clusters are obtained using a density-based clustering method. In each cluster, we identify a smooth "permissive" subspace which avoids high-energy barriers and then underestimate the binding energy function using general convex polynomials in this subspace. We use the underestimator to bias sampling towards its global minimum. Sampling and subspace underestimation are repeated several times and the conformations sampled at the last iteration form a refined ensemble. We report computational results on a comprehensive benchmark of 224 protein complexes, establishing that our refined ensemble significantly improves the quality of the conformations of the original set given to the algorithm. We also devise a method to enhance the ensemble from which near-native models are selected.

Proteins are a key element of the cell and play an important role in a variety of cellular functions such as ligand binding, metabolic control, cell signaling and gene regulation. The prediction of the tertiary structure of protein complexes is known as the *protein-protein docking* problem. Several experimental techniques, primarily X-ray crystallography and nuclear magnetic resonance (NMR), are used to predict the 3-dimensional (3D) structure of macromolecular complexes, including proteins, but these methods are usually expensive, time-consuming, and may not be applicable to short-lived molecular complexes. Therefore, computational protein docking methods are very much in need and have attracted considerable attention in the last two decades.

Based on the principles of thermodynamics, the most stable state of a protein complex (the *native* conformation) occurs when its Gibbs free energy attains its minimum value. Binding involves conformational changes to the unbound state of the complex components, affecting their backbones and the side-chains. In this light, the protein docking problem can be posed as an optimization problem in which the variables are the atomic coordinates of the proteins and the objective is to minimize the binding energy of the complex. While this formulation ignores configuration-related entropy terms of free energy, these can be incorporated by post-processing low energy solutions using metrics of density and cluster size.

Despite significant progress in recent years, protein docking is still regarded as a very challenging problem in structural biology due to the complexity of the energy landscape of protein-protein or other protein[1] interactions. This complexity stems from the fact that the energy function is composed of multiple force-field energy terms (such as the Lennard-Jones potential, solvation, hydrogen bonding, electrostatics, etc.) acting in different space

[1]Division of Systems Engineering, Boston University, Boston, Massachusetts, United States of America. [2]Department of Mechanical Engineering, Boston University, Boston, Massachusetts, United States of America. [3]Department of Biomedical Engineering, Boston University, Boston, Massachusetts, United States of America. [4]Department of Electrical and Computer Engineering, Boston University, Boston, Massachusetts, United States of America. [5]Department of Applied Mathematics and Statistics and Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, New York, United States of America. [6]Present address: 8 Saint Mary's St., Boston, MA, 02215, United States of America. Correspondence and requests for materials should be addressed to S.V. (email: vajda@bu.edu) or D.K. (email: midas@laufercenter.org) or I.C.P. (email: yannisp@bu.edu)

scales and resulting in a multi-frequency behavior of the various energy terms. Therefore, the energy function exhibits multiple deep funnels and extremely many local minima over its multidimensional domain.

In response to this level of complexity, state-of-the-art docking protocols employ a two-stage approach. At the first stage, we use a simplified energy function (expressed as a correlation function) and sample on a conformational space grid an enormous number of docked receptor-ligand conformations, using Fast Fourier Transforms (FFT) for energy evaluation. To conduct this initial sampling we use the protein docking server *ClusPro 2.0* which is based on a docking program called PIPER[2]. These conformations are then sorted by their energy values, and the top few thousands with the lowest energy are retained for further processing. At the second stage of docking protocols, we seek to *refine* low energy conformations by moving off-grid and utilizing more elaborate energy functions. Our work in this paper focuses on this *refinement stage*[3]. One of the distinguishing features of our work is that it *does not* assume any prior knowledge about the native structure. In fact, the input to our algorithm is the output of the PIPER docking software, which consists of the lowest energy globally sampled conformations. We evaluate how well these initial conformations are refined by considering the number of good quality solutions in the refined ensemble.

The *refinement problem* we outlined, inherits the complex structure of the binding energy landscape. Approaches that have been considered almost invariably involve efficient sampling and methods that attempt to "smooth" the energy function. A successful strategy is to use *Monte Carlo*-based sampling[4]. An alternative method that resamples around low-energy PIPER structures has also been proposed[5]. A host of methods seek to leverage the *funnel-like* shape of the energy function[6–8]. In fact, similar strategies have been used in protein folding[9–12]. The binding energy funnel is restricted to a neighborhood of the native complex[13] and there is a free energy gradient toward the native state. However, the funnel is rough, giving rise to many local minima[14] that correspond to encounter complexes, some of which may be visited along a particular association pathway[15,16].

Underestimation

An early algorithm designed for protein folding, the *Convex Global Underestimator* (*CGU)* method[17], introduced the idea of using an approximation of the envelope spanned by the local minima of the energy function in the form of *convex canonical quadratic* underestimators. CGU, however, used a restricted class of underestimators[18], limiting its effectiveness. The *Semi-Definite programming-based Underestimation (SDU)* method[18,19] uses the same approach as CGU but it considers the class of "general" convex quadratic functions to underestimate, in addition to introducing an exploration strategy biased by the underestimator.

In this paper we propose a number of generalizations to SDU. First, and following our earlier preliminary work[20], we consider the more general class of *convex polynomial functions* for underestimation. Polynomial functions are more flexible than quadratic functions used in the aforementioned methods[17–19] and can more tightly approximate a funnel.

A second generalization is the ability to handle multiple local funnels in the original cluster presented for refinement. This is important because by deriving a single underestimator (as in Nan *et al.*[20]), we will tend to "average" a complex energy landscape and produce a minimum of the underestimator that may not correspond to a low-energy funnel basin. We resolve this issue by establishing an effective exploration procedure using density-based clustering as follows. First, we run a density-based clustering algorithm on the set of (PIPER) structures which are the inputs to the refinement protocol. This phase eliminates outliers and low-density regions of the conformational space, resulting in multiple sub-clusters whose size is greater than a pre-specified threshold. Then, we construct one underestimator per sub-cluster which allows us to approximate and explore each sub-cluster separately. Finally, we combine all the sampled conformations from all clusters, and pick the low-energy conformations as the output of the refinement protocol.

Dimensionality Reduction

An important question in underestimation is to determine the appropriate multi-dimensional space in which underestimation takes place. Our experience has suggested that for many complexes, underestimation in the entire 6D space of conformational variables (translations and rotations of the ligand with respect to the receptor) may not be effective and can produce underestimators whose minimum is outside the range of the cluster. This is due to "singularities" of the energy landscape resulting in energy being very steep along some directions and flat along others.

Realizing this, the original SDU[18,19] removes the center-to-center distance of receptor and ligand from the 6D parameterization of the space because this dimension does not exhibit any significant variation over the ensemble of input samples, which implies a very narrow energy funnel along this dimension. These initial attempts led us to a more fundamental re-assessment of the space in which underestimation must take place. In our previous work[21], we discovered that the near-native cluster in protein-protein complexes exhibits reduced dimensionality, suggesting that *proteins associate along preferred pathways*, similar to sliding of a protein along DNA in the process of protein-DNA recognition. We extracted the landscape features via *Principal Component Analysis (PCA)* using two distinct energy functions, one derived from PIPER sampling[2] and the other using RosettaDock[4]. In both cases, we found that most of the variability (more than 75%) in the cluster can be explained by 3 (and sometimes 2) eigenvectors, suggesting that the energy landscape consists of a *permissive subspace* spanned by the 2 or 3 eigenvectors with the largest eigenvalues and a *restrictive landscape* spanned by the remaining eigenvectors. Figure 1 illustrates the landscape of the 2YVJ complex. It plots the distributions of Interface RMSD (root mean square deviation of interface atoms from the native) in Å and energy values based on structures generated by PIPER along the 5 eigenvectors produced by PCA, plotted from top to bottom in decreasing corresponding eigenvalue. Dark blue diamonds indicate low energy data points used for the PCA. Notice how the variability of the data points decreases from top (very wide) to bottom (very narrow).

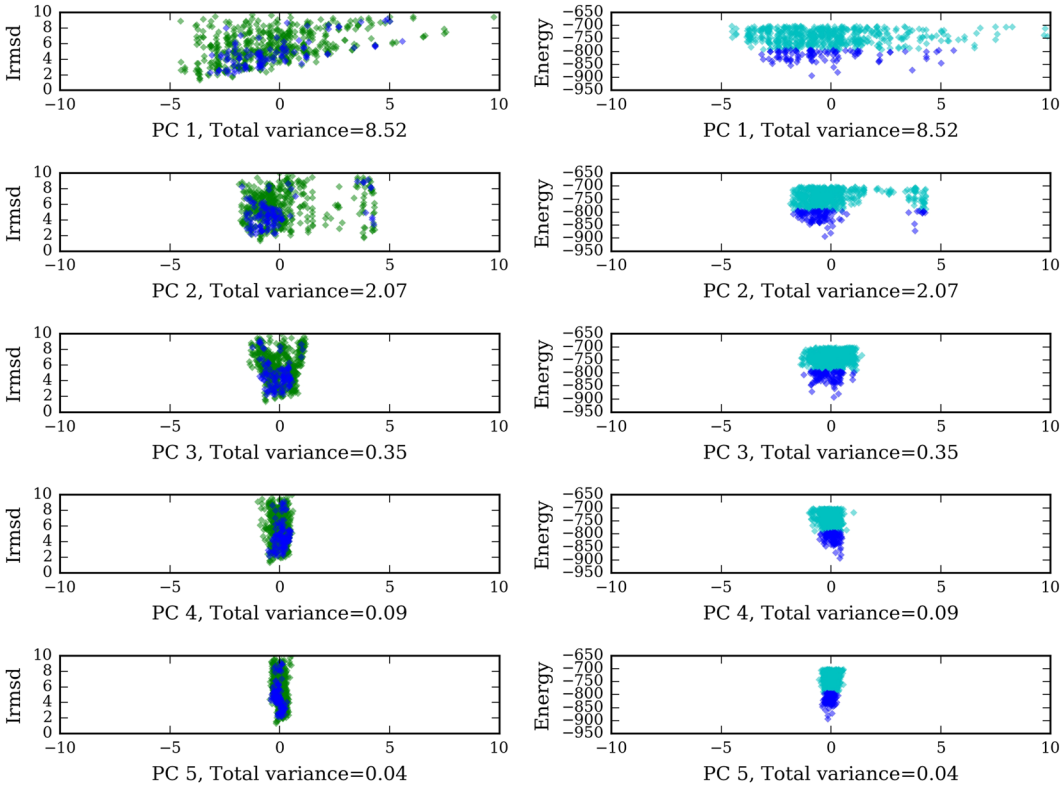

Figure 1. The near-native energy landscape of the 2YVJ complex using ClusPro conformations.

This behavior has a deep biophysical explanation. Docking is initially driven by a diffusive search governed by Brownian motion, which brings the two molecules close. The encounter complex can be thought of as an ensemble of conformations in which the two molecules can rotationally diffuse along each other, or participate in a series of "microcollisions" that properly align the reactive groups. The second step of association consists of conformational rearrangements leading to the native complex. While it has been generally recognized that association proceeds through a transition state, little was known of the encounter complex structures and configurations, as their populations are low, their lifetimes are short, and they are difficult to trap. In our earlier work[21], we have used results from the application of NMR Paramagnetic Relaxation Enhancement (PRE), a technique that is extremely sensitive to the presence of lowly populated states in the fast exchange regime[22–24]. Our results indicate that the PRE profiles obtained experimentally are consistent with the presence of the encounter complexes that our landscape dimensionality analysis revealed.

In this paper we use this insight to propose a new *stochastic global optimization* method we call *Subspace Semi-Definite programming-based Underestimation (SSDU)*. SSDU is based on SDU with all the generalizations we introduced earlier. The most fundamental difference however, is that underestimation takes place only in the permissive conformational subspace found by PCA. This has the effect of avoiding high-energy barriers and evaluating the energy function only at non-singular points. Since the (typically) 3D permissive subspace contains encounter complexes, the *sequence of permissive subspaces* SSDU generates amounts to a characterization of a *smooth preferred association pathway*. Put differently, these subspaces correspond to a decreasing sequence of *energy plateaux* paving a smoother way of descending to the native state.

The remainder of the paper is organized as follows. We start by presenting the SSDU algorithm (Methods). The computational results on a large benchmark set of protein structures are presented in the "Results and Discussion" Section. We conclude with some final remarks.

Notation. Vectors will be denoted using lower case bold letters and matrices by upper case bold letters. For economy of space we write $\mathbf{v} = (v_1, \ldots, v_n)$ for $\mathbf{v} \in \mathbb{R}^n$. Prime denotes transpose. For a matrix $\mathbf{P}$, $\mathbf{P} \succeq 0$ indicates positive semi-definiteness.

Methods

Dimensionality Reduction. A receptor-ligand conformation can be parameterized by a 6D vector $\psi = (\boldsymbol{\rho}, \boldsymbol{y})$, where $\boldsymbol{\rho} = (r, a, b) \in \mathbb{R}^3$ represents the translation vector from ligand center to receptor center and $\mathbf{y} = (y_1, y_2, y_3) \in \mathbb{R}^3$ is a parameterization of the rotation of the ligand with respect to the three axes. The space of conformations ψ is a nonlinear manifold and the parameterization of the rotations corresponds to a projection from a (flat) tangent space (in which $\boldsymbol{y}$ is defined) to the manifold itself, projecting straight lines on the tangent space map onto geodesics of the manifold. We refer the reader to the Supplement and related papers[25,26] for a more detailed discussion of these spaces.

In the translation vector $\boldsymbol{\rho}$, r is the length of the vector and a,b indicate the so called exponential coordinates of the azimuth and zenith angles of $\boldsymbol{\rho}$ (see Supplement). Let us denote by $f\colon \mathbb{R}^6 \to \mathbb{R}$ the energy function of a conformation parameterized by $\boldsymbol{\psi} = (r, a, b, y_1, y_2, y_3)$.

As we mentioned earlier, in low-energy clusters where conformations are well-packed, there is no significant variation in the center-to-center distance r between a ligand and the receptor, and this variable can be easily optimized separately once all other variables are determined. Thus, we remove r from $\boldsymbol{\psi}$ and minimize f with respect to the remaining variables $\mathbf{x} = (a, b, y_1, y_2, y_3) \in \mathbb{R}^5$.

We already discussed in the previous section that the region of the space in the neighborhood of the native state is composed of high energy barriers that prevent the ligand to move in one or two directions[21], giving rise to a *restrictive subspace* spanned by these directions. Orthogonal to the restrictive subspace we have a *permissive subspace* where the energy is much smoother. To identify the restrictive and permissive subspaces, we apply PCA and convert the 5D parameterization of the conformational space ($\boldsymbol{x}$) into linearly uncorrelated variables called *principal components* using an orthogonal transformation. This transformation seeks to find a set of principal components with the following property: the first principal component accounts for the largest possible variability in the data, and each succeeding component has the highest variance amongst all possible components which are orthogonal to the preceding components.

Suppose now we have obtained a sample of K local minima of f in the $\boldsymbol{x}$-space together with the their corresponding energy values: $(\boldsymbol{x}^{(i)}, f^{(i)} = f(\boldsymbol{x}^{(i)}))$, $i = 1, \ldots, K$. We perform PCA and let $\boldsymbol{z}^{(i)}$ be the ith sample point (local minimum) expressed in the basis of the principal coordinates. Our earlier work[21] shows that in most protein-protein complexes, the first 3 PCA eigenvalues are significantly larger than the other 2 eigenvalues. Thus, we can take the first 3 principal coordinates $\{z_1, z_2, z_3\}$ to form the permissive subspace, while the remaining 2 coordinates $\{z_4, z_5\}$ form the restrictive subspace we wish to eliminate. We denote the new coordinates of the i th sample point in the 3D permissive subspace by $\phi^{(i)} = (z_1^{(i)}, z_2^{(i)}, z_3^{(i)}) \in \mathbb{R}^3$.

Next, we aim at minimizing the energy function f by constructing a semidefinite underestimator over the samples $\boldsymbol{\phi}^{(i)}$, $i = 1, \ldots, K$, in the permissive landscape.

Underestimation. As discussed in the previous section, our method is based on finding convex underestimators which can be regarded as an approximation of the envelope spanned by the local minima of the binding energy function. In an effective underestimation, the minimum of the convex underestimator will be an approximation of the global minimum of the funnel-like binding energy function. Therefore, we can bias further sampling towards the underestimator's minimum. Below, we first explain how the convex underestimator can be calculated, then, in the next subsection, we focus on how to bias sampling towards to the underestimator's minimum point.

Following our earlier preliminary work[20], we consider the class of general convex polynomial underestimators. Let $U(\boldsymbol{\phi})$ be a degree $2d$ polynomial and $\phi \in \mathbb{R}^n$, where $n = 3$ in the case of seeking an underestimation in the 3D permissive subspace. In general, it is hard to show whether a general polynomial function is convex or not (except for the special case of quadratic underestimators where $2d = 2$). It has been shown that even verifying the convexity of a degree-4 polynomial is an intractable problem[27].

Instead, we will use a computationally tractable relaxation for convexity, called *SOS-convexity*[28]. The main idea is to verify whether a quadratic function constructed from the Hessian matrix of $U(\cdot)$ (the matrix of 2nd partial derivatives) is a Sum-of-Squares (see Supplement). We can then formulate the problem of finding a convex polynomial underestimator of the sample points $(\boldsymbol{\phi}^{(i)}, i = 1, \ldots, K)$ as the following problem:

$$\begin{aligned} \min_{U(\cdot)} \quad & \sum_{i=1}^{K} [f^{(i)} - U(\phi^{(i)})] \\ \text{s.t.} \quad & f^{(i)} \geq U(\phi^{(i)}), \qquad \forall\, i, \\ & U(\,\cdot\,) \text{ is SOS-convex}, \end{aligned} \tag{1}$$

where the optimization is over the coefficients of the polynomial $U(\cdot)$. This problem can be reformulated as a standard semi-definite programming (SDP) problem (see Supplement). We use the CSDP solver[29] to solve this SDP problem. Solving (1) provides us with the optimal coefficients of the polynomial convex function $U(\boldsymbol{\phi})$ that can be regarded as a tight underestimator of the K local minima $(\boldsymbol{\phi}^{(i)}, i = 1, \ldots, K)$.

Sampling. Let $\phi^* \in \mathbb{R}^3$ be the global minimum of the convex underestimator obtained from the solution of (1). We will use it to sample more conformations in the vicinity of $\boldsymbol{\phi}^*$. If the underestimation step succeeds in capturing the shape of the free energy function, then the sampling step will help us generate more conformations in the vicinity of the global minimum of the energy function.

First, we generate $\overline{K}$ random samples $\mathbf{s}^{(l)} \in \mathbb{R}^5$, $l = 1, \ldots, \overline{K}$, where each random dimension $s_i^{(l)}$ has a uniform distribution in the range of $(-0.5\beta\sigma_i, 0.5\beta\sigma_i)$, $i = 1, \ldots, 5$, where β is a constant and σ_i is the i th PCA eigenvalue (i th diagonal element of $\boldsymbol{\Sigma}$ in Eq. (S.1) of the Supplement), where $\sigma_1 \geq \ldots \geq \sigma_5$. Then, we construct the 5D global minimum $\boldsymbol{z}^*$ by appending an approximation of z_4^*, z_5^* to $\boldsymbol{\phi}^*$. As discussed earlier, the last two principal coordinates z_4, z_5 have small variation over the samples; therefore we can consider their sample mean as a good approximation, i.e., $z_i^* = (1/K)\sum_{j=1}^{K} z_i^{(j)}$, $i = 4, 5$, and set $\mathbf{z}^* = (\phi^*, z_4^*, z_5^*)$.

Next, we generate the new sample points in the vicinity of the underestimator's global minimum by randomly perturbing $\boldsymbol{z}^*$ along each principal coordinate by $\boldsymbol{s}^{(l)}$. We transform these new sample points from the principal coordinates to the original coordinates, obtaining $\tilde{\mathbf{x}}^{(l)}$. (Specifically, $\tilde{\mathbf{x}}^{(l)} = \mathbf{W}(\mathbf{z}^* + \mathbf{s}^{(l)}) + \overline{\mathbf{x}}$ where $\boldsymbol{W}$ is the matrix

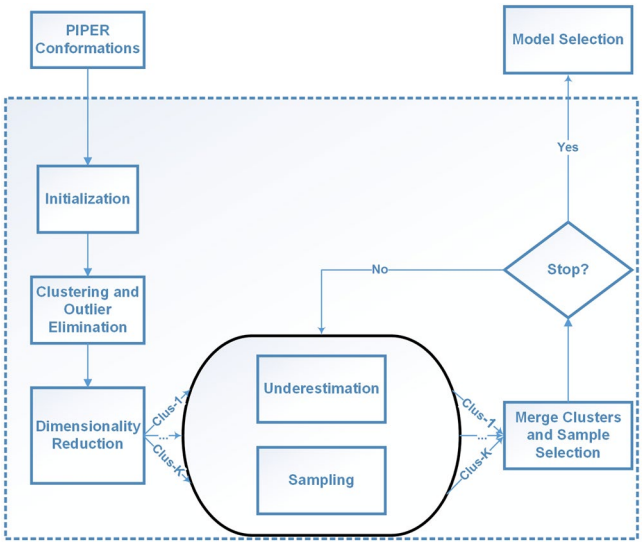

Figure 2. The flowchart of the SSDU procedure.

defined by Eq. (S.1) of the Supplement and $\bar{\mathbf{x}}$ is the mean of the K local minima expressed as vectors in the $\boldsymbol{x}$-space).

The sampling range of random samples $\boldsymbol{s}^{(l)}$ at each dimension i is proportional to the variance σ_i to guarantee an effective coverage of the conformational space which preserves the sample distribution. Furthermore, in order to construct the 6D conformational parameterization of these generated sample points, we need to append the sample mean of the center-to-center distance r, i.e., $\bar{r} = (1/K)\sum_{i=1}^{K} r^{(i)}$, which results in the new sample conformation $\tilde{\psi}^{(l)} = (\bar{r}, \tilde{\mathbf{x}}^{(l)}) \in \mathbb{R}^6$.

Clustering and Outlier Elimination. As we have discussed in the Introduction, the input conformations we wish to refine may span several energy funnels. To separate these funnels before underestimation, we perform *clustering and outlier elimination*. The idea is simply to cluster the input conformations with respect to a distance measure (we use the Euclidean distance). To that end, we employ a *density-based* clustering method called *Density-Based Spatial Clustering of Applications with Noise (DBSCAN)*[30]. Given a set of sample points in the conformational space, DBSCAN groups the points which are closely packed together in a dense region and eliminates the outlier points sitting in the low-density regions. In this scheme, the dense regions are defined as the *clusters*, which are separated by the low-density regions. DBSCAN requires two input parameters: (*i*) ε, the distance threshold which is defined as the maximum distance of two sample points to be considered as neighbors, and (*ii*) N_{min}, the minimum number of points required to form a cluster. The second parameter N_{min} ensures that all clusters found by DBSCAN will contain at least N_{min} points, and the algorithm will automatically eliminate outliers located in low-density regions.

In the case of having multiple local minima in the neighborhood of the native structure, the clustering phase will tend to group the conformations around each local minimum in a separate cluster. In the sequel, we explain how we use these clusters to handle the situations in which most of the underestimation-based refinement methods with a single underestimator[18–20] may fail to locate the global minimum of the energy function in the near-native region.

SSDU Algorithm. We have now described all key steps of the SSDU algorithm. The entire algorithm is outlined below in Algorithm 3. We note that the algorithm explores separately the potential multiple sub-clusters discovered by DBSCAN. Using the sampling approach we outlined, we sample K conformations in each sub-cluster. We then merge all these conformations and pick the lowest energy conformations. We can iterate over the steps of SSDU until meeting the stopping criteria. The retained conformations can be regarded as the SSDU outputs. Figure 2 shows a flowchart of the SSDU procedure demonstrating the process of refining the initial PIPER sample conformations to produce the ensemble of refined structures.

Local Minimization. All the presented sampling approaches use a common local minimization subroutine. Its main role is to account for flexibility of side chains during the search. We have explored and optimized this protocol in our previous work[31]. It consists of the following steps. We first run a *side-chain positioning (SCP)* algorithm[31,32] that solves a relaxed formulation of a combinatorial optimization problem in order to repack the amino acid residues at the interface of the receptor-ligand complex. Then we run a *rigid-body energy minimization* algorithm[25] which locally minimizes the position and orientation of the ligand with respect to the receptor.

Energy Function. Our choice of energy function is a high-accuracy docking energy potential that can be calculated as a weighted sum of a number of force-field and knowledge-based energy terms[4,33,34]. Following our earlier work[5,31], we consider the following energy terms to find the interaction free energy value:

$$E = w_{VDW}E_{VDW} + w_{SOL}E_{SOL} + w_{COUL}E_{COUL} + w_{HB}E_{HB} + w_{DARS}E_{DARS} + w_{RP}E_{RP},$$

where E_{VDW} is the Lennard-Jones potential, E_{SOL} is an implicit solvation term[35], E_{COUL} is the Coulomb potential, E_{HB} is a knowledge-based hydrogen bonding term[36], and E_{DARS} is a structure-based intermolecular potential that is derived from the non-redundant database of native protein-protein complexes which uses a novel *DARS (Decoys as Reference State)*[37] reference set. The last term, E_{RP}, is a statistical energy term associated with a set of rotamers selected from the backbone-dependent rotamer library[38]. The weight set of the energy function is adopted according to the selections in Gray *et al.*[4].

Validation Dataset and Input Preparation. We validated our algorithm on a comprehensive benchmark of 230 protein complexes consisting of *Enzymes*, *Antibodies and Other types*[39].

Other types of complexes exhibit multiple deep funnels in the vicinity of the native structure which makes them particularly difficult cases for protein docking refinement, whereas enzyme interactions are usually driven by shape complementarity, making them relatively easier cases. In fact, considering a wide spectrum of docking test cases in terms of difficulty, enables us to examine the performance gain compared to the ClusPro server in different scenarios. Moreover, other types of complexes present an opportunity to evaluate the effect of the density based clustering component built into SSDU where fitting multiple underestimators seems inevitable. Input preparation consists of two steps: (1) running global FFT sampling using PIPER; and (2) filtering the conformations to retain the top 1000 and 1500 for enzymes/antibodies and other types, respectively. These top energy conformations are supplied as the input to the SSDU algorithm.

Data availability. The complexes we considered are part of a standard docking benchmark publicly available[40]; the structures are available through the protein data bank[41]. The code we developed is available upon reasonable request to the authors and a version will soon be released at a public depository.

Results and Discussion

In this section, we compare the SSDU-produced ensemble with the corresponding input ensemble produced by ClusPro. We use ClusPro as a baseline for comparison because it has been established to perform comparably well to other methods[42]. In fact, ClusPro has ranked first multiple times among automated servers in the rounds of the *Critical Assessment of Prediction of Interactions* (CAPRI) community-wide experiment in the years 2009, 2013 and 2016. Furthermore, we have access to the ClusPro source code and can appropriately adjust its output for the purposes of our refinement experiments. In what follows, we consider both the number of near-native conformations in each ensemble and the implications in selecting a near-native conformation out of the refined ensemble without knowing the native structure.

The results are based on the following parameter settings: $K = 1000$ indicates the number of conformations for enzymes and antibodies and $K = 1500$ for other types of complexes, provided as the input to SSDU, $\varepsilon = 1.0$ and $N_{min} = 100$ are the parameters used in DBSCAN (Step 2 of Alg. 3), $\eta = 0.3$ (Step 6 of Alg. 3), and a maximum number of iterations equal to 3 is used for SSDU termination.

Algorithm 1. SSDU Algorithm.

1: **Initialization:** Starting from K sample points in conformational space $\mathscr{S}$, perform local minimization to obtain K distinct local minima $\psi^{(1)}, \ldots, \psi^{(K)}$ of $f(\cdot)$.
2: **Clustering and Outlier Elimination:** Run DBSCAN over the input sample points to split the dataset into several clusters. Let n be the number of clusters found by DBSCAN and $\{\mathscr{C}_1, \ldots, \mathscr{C}_n\}$ the corresponding clusters.
3: **Dimensionality Reduction:** For each sample point i reduce $\psi^{(i)} \in \mathbb{R}^6$ to $\mathbf{x}^{(i)} \in \mathbb{R}^5$, then transform $\mathbf{x}^{(i)}$ to $\phi^{(i)} \in \mathbb{R}^3$ using PCA.
4: **Exploration:** For each cluster $\mathscr{C}_i$, $i = 1, \ldots, n$,

- **Underestimation:** Solve the SDP in (1) to obtain the convex polynomial underestimator $U_i(\phi)$. Set the predictive point ϕ_i^* to be the minimizer of $U_i(\phi)$. Transform ϕ_i^* to $\mathbf{z}_i^*$ in the 5D conformational space.
- **Sampling:** Transfer $\mathbf{z}_i^*$ from the principal coordinates into the original coordinates and generate random samples $\tilde{\mathbf{x}}_i^{(l)}$, $l = 1, \ldots, K$, for each cluster $\mathscr{C}_i$. Construct $\tilde{\psi}_i^{(l)}$ in the 6D conformational space from $\tilde{\mathbf{x}}_i^{(l)}$.

5: **Sample Selection:** Merge the output sampled conformations of all clusters and the inputs to the algorithm and select K top conformations with the lowest energy value. Let ψ^G be the conformation with minimum energy value amongst the K retained conformations.
6: **Termination:** If $\|\psi^G - \psi^*\| < \eta$ or there is no progress in reducing $f(\psi)$ or the maximum number of iterations is reached then stop; otherwise go to Step 4.

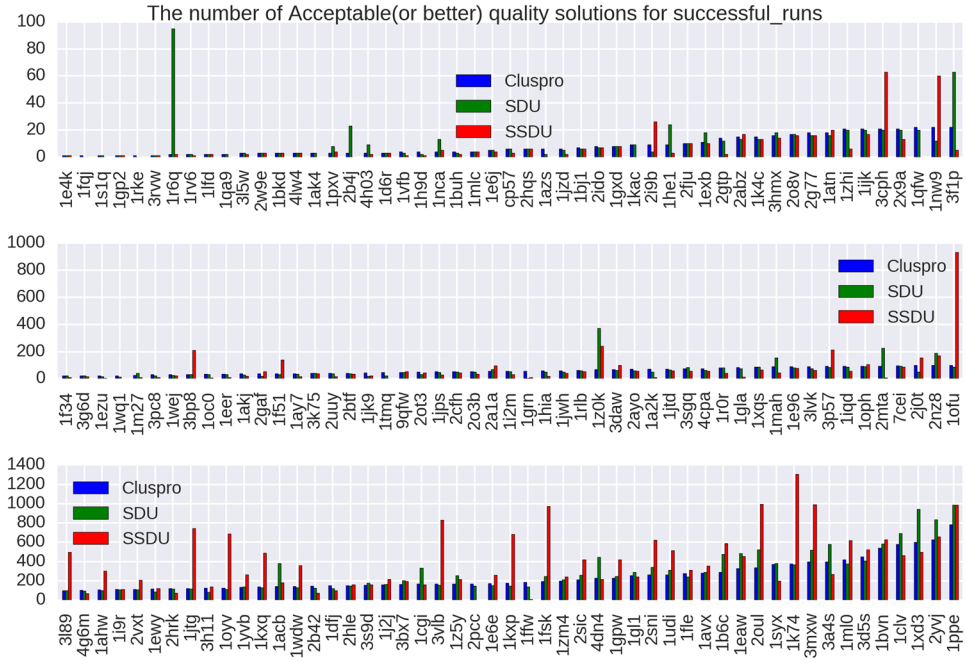

Figure 3. The *x*-axes of these plots list 156 out of 224 protein complexes that have either ClusPro or SSDU non-zero CAPRI Acceptable (or better) quality solutions. The complexes are sorted by the number of ClusPro counts and the *y*-axis shows the number of Acceptable (or better) quality solutions out of an ensemble of 1000 or 1500 conformations for enzymes/antibodies and other types, respectively, produced by ClusPro, or refined by SDU and SSDU.

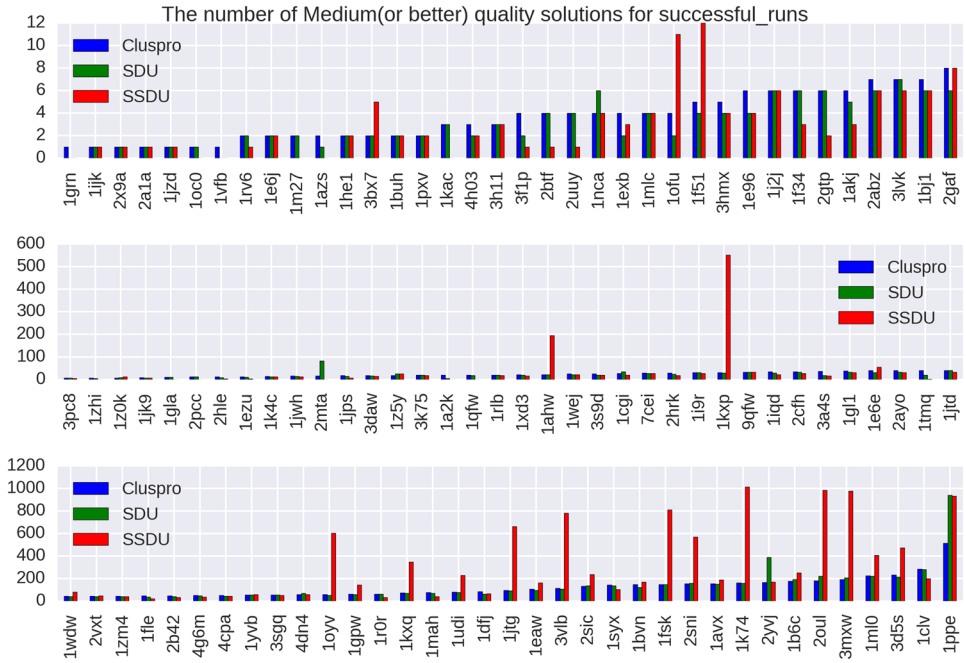

Figure 4. The *x*-axes of these plots list 110 out of 224 protein complexes that have either ClusPro or SSDU non-zero CAPRI Medium (or better) quality solutions. The complexes are sorted by the number of ClusPro counts and the *y*-axis shows the number of Medium (or better) quality solutions out of an ensemble of 1000 or 1500 conformations for enzymes/antibodies and other types, respectively, produced by ClusPro, or refined by SDU and SSDU.

Figure 5. The *x*-axes of these plots list 29 out of 224 protein complexes that have either ClusPro or SSDU non-zero CAPRI High quality solutions. The complexes are sorted by the number of ClusPro counts and the *y*-axis shows the number of High quality solutions out of an ensemble of 1000 or 1500 conformations for enzymes/antibodies and other types, respectively, produced by ClusPro, or refined by SDU and SSDU.

Benchmark		SSDU vs. ClusPro	SSDU vs. SDU
Acceptable (or better)	Average	24.62%	21.31%
	Total	53.14%	30.37%
Medium (or better)	Average	53.26%	58.25%
	Total	132.69%	112.43%
High	Average	410.71%	405.88%
	Total	424.93%	157.06%

Table 1. Percentage improvement of Acceptable (or better), Medium (or better) and High quality solutions by SSDU versus SDU and ClusPro for a benchmark of 224 complexes. Note that for each of the entries in the table, complexes with zero number of solutions for both ClusPro and SDU/SSDU are removed.

Protein Docking Refinement. To show the impact of the SSDU algorithm, we provide three different plots (Figs 3, 4 and 5) showing the number of Acceptable (or better), Medium (or better) and High quality solutions before and after SSDU. The classification of the quality of the solutions is based on metrics adopted in the CAPRI experiments[43]. These metrics are: interface RMSD (iRMSD), backbone RMSD (LRMSD) and the number of native contacts preserved (Fnat). To classify a conformation using these metrics, the program DockQ was used[44]. DockQ combines normalized values of iRMSD, LRMSD and Fnat to generate a continuous score in the range [0,1]; the higher the score, the better the quality of a solution. Specifically, a conformation of a protein complex is classified into four categories: Incorrect, Acceptable, Medium or High based on its DockQ score. Moreover, in addition to SSDU and ClusPro, the quality of the solutions produced by the SDU algorithm[19] is presented as well in order to measure the performance boost from the innovations we have introduced in this paper.

We note that *the unbound protein structures* were used to generate the input to the SSDU/SDU algorithms. The use of unbound structures is important since we want to assess docking performance in the absence of any knowledge about the native conformation. As we mentioned earlier, the inputs to SSDU/SDU are the top 1000 and 1500 energy conformations from ClusPro for enzymes/antibodies and other types, respectively. The output of SSDU/SDU has the same number of conformations as the input and contains a mixture of conformations from the input and SSDU/SDU re-sampled conformations. Specifically, the re-sampled conformations from SSDU/SDU are merged with the input conformations and then subjected to energy filtering to retain the same number of lowest energy conformations as the input. For example, if the input has 1000 conformations and SSDU density-based clustering discovers three clusters, the number of conformations after merging them with the input will be 4000 (1000 per cluster and 1000 from the input), from which the 1000 lowest energy conformations are selected as the SSDU output.

We also note that we report results on 224 out of 230 complexes in the benchmark[39]. The 6 removed complexes are 4GAM, 4GXU, 2H7V, 4FQI, 1DE4, 1N2C. These complexes were removed because one of the programs we use failed to produce a score/solution for many conformations (DockQ for the first three, SSDU for the fourth, and SDU for the last two).

As it is apparent from Figs 3, 4 and 5, SSDU substantially increases the number of acceptable or better quality solutions. The amount of improvement by SSDU compared to SDU and ClusPro is reported in Table 1. The average improvement is determined by calculating the percentage improvement for each protein complex and averaging over different complexes in the benchmark, whereas the total improvement is the percentage improvement when the number of near-native hits are aggregated over all the complexes in the benchmark.

We also noticed that SSDU tends to decrease the variability of iRMSD and total energy in a near-native cluster. We provide in the Supplement an example of landscape analysis showing this effect (for the same 2YVJ complex we showed in Fig. 1).

Post-Processing Ensemble Enrichment. We have established that SSDU generates outputs with significantly higher quality compared to the input ClusPro conformations. Next, we examine whether we can select a small number (specifically, 10) of enriched clusters from this SSDU ensemble which maintain a significant portion of the high quality conformations.

Selecting a high quality conformation remains a very challenging problem in the protein docking community. In CAPRI, participating groups test their methods in blind predictions of given target protein complexes. As mentioned before iRMSD, LRMSD and Fnat are used to categorize the predictions into Incorrect, Acceptable, Medium, and High quality. Reflecting how challenging the problem is, CAPRI allows for 10 submissions from each participating group.

ClusPro, against which we compare our results, uses clustering as a way of taking into account entropic metrics that were not included in the energy function we described earlier. Specifically, the ClusPro clustering algorithm[45] is a greedy algorithm where, at each iteration, the structure with the largest number of neighbors is identified (two conformations are considered neighbors if their pairwise iRMSD is less than a 9 Å threshold). Then, the conformation with the highest number of neighbors is labeled a *cluster center* and along with its neighbors form a cluster and removed from the ensemble. The procedure is repeated for the remaining conformations. Overall, a maximum of 30 clusters are formed where each cluster contains at least 10 members. The collection of cluster centers produced in this manner forms a putative set of high quality conformations. ClusPro selects the centers of the 10 largest clusters as its submissions to CAPRI.

We will consider whether replacing the ClusPro ensemble with the SSDU ensemble also *enriches* the top 10 selected clusters. In this work, and because SSDU is an improved sampling method, we focus solely on the question of *cluster discrimination*, that is, selecting 10 enriched clusters. The question of *conformation discrimination*, which amounts to selecting a single representative conformation from each top cluster, is outside the scope of this paper and is left open to future work.

We form SSDU clusters by clustering the conformations in the SSDU ensemble in exactly the same way as ClusPro. We rank these clusters using a ranking method we describe in the sequel. For each complex we compare two sets of clusters. The first (ClusPro) set is formed by clustering the ClusPro produced structures and ranking the clusters in decreasing cluster size. The second (SSDU) set is formed by first refining with SSDU the ClusPro ensemble, then generating (typically 30) clusters using the ClusPro clustering algorithm, and finally ranking these clusters using the method we describe next. In each case, we compute the number of Acceptable/Medium/High quality solutions among the top 3, 5 and 10 clusters.

Ranking the SSDU ensemble. We will next employ a machine learning approach for ranking the 30 clusters generated from the SSDU set. Some related work on using machine learning approaches, different than ours, for ranking has appeared in the literature[46,47]. We used several classification algorithms on this dataset: random forests, support vector machines with linear and radial kernels and logistic regression. *Random forests*[48] achieved the best performance and in the remainder of this section we will focus on this classifier. To perform the classification we characterize each cluster with a set of 9 features described below:

1. The first four consist of the average energy value of the top 25%, 50%, 75% and 100% lowest energy conformations in the cluster, respectively.
2. The 5th feature is the number of conformations (size) of the cluster.
3. The last four features consist of the average RMSD between the cluster center and the top 25%, 50%, 75% and 100% conformations, respectively, in an ordered list of cluster conformations ranked in increasing RMSD from the cluster center.

We label each cluster by evaluating the DockQ score of the cluster center: if it has Acceptable quality score (or better) it is given a label of $+1$ (positive class); otherwise a label of -1 (negative class).

The random forest classification algorithm trains a set of unpruned de-correlated classification trees using random selection of training data and random selection of variables. It classifies a new sample by taking a majority vote of all trees, which reduces through averaging the variance of the decision. To each new sample we associate a probability of the sample belonging to the positive class as follows. The new sample is classified by each tree in the random forest and ends up in some leaf node of the tree. The percentage of training samples assigned to that leaf node which belong to the positive class is used as a surrogate of the probability that the new sample belongs to the positive class. These probabilities are then averaged over all trees in the forest to compute an overall probability

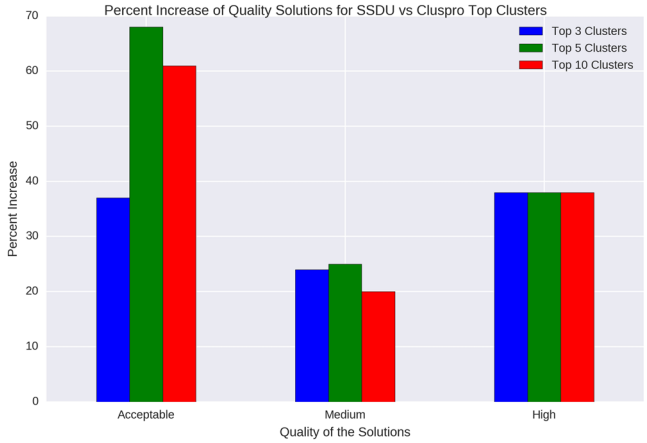

Figure 6. The percentage of increase in the number of Acceptable/Medium/High quality solutions among the top 3, 5 and 10 clusters achieved by SSDU over ClusPro.

that the sample belongs to the positive class. A classification decision can then be made by comparing that probability to a given threshold. Moreover, samples can be ranked using this probability.

We train random forest classifiers by randomly dividing the whole dataset into *non-overlapping* training and testing datasets, assigning 60% of the complexes in the training set and the remaining 40% to the test dataset. Because there are in general fewer clusters with a positive label, we oversampled those clusters in the training set so as to have a more balanced representation of positive and negative class clusters for training the random forest. The test dataset is not biased; it is selected at random from the entire benchmark and for each complex we select all its associated clusters. We evaluate classification performance through the Receiver Operating Characteristic (ROC) curve computed on the test set. The ROC plots the true positive rate (fraction of positive test samples correctly identified as positive) vs. the false positive rate (fraction of negative samples incorrectly identified as positive) as the threshold used for the classification decision changes. The Area Under the ROC Curve (AUC) is used as a prediction performance metric. An AUC of 1 represents perfect classification accuracy, whereas an AUC of 0.5 represents a naive random classifier which assigns samples to a class by flipping a coin.

We use the probability of a sample belonging to the positive class in order to rank (in decreasing order of the probability) the SSDU set of clusters. Similar to the ClusPro results, we count the number of Acceptable/Medium/High quality solutions among top 3, 5 and 10 clusters. Finally, we measure the improvement in the number of quality solutions in each of the three categories.

As we described, we processed the SSDU cluster set using non-overlapping datasets for training and testing. We repeated training and testing 15 times, each time with a different random split of the dataset, and averaged the AUC computed on the test set (out-of-sample) over the 15 runs. This yielded an average AUC for other type of complexes equal to 0.62. This value indicates adequate classification accuracy, significantly better than random selection. Figure 6 shows the amount of improvement SSDU achieves over ClusPro for different quality categories of Acceptable/Medium/High among the top 3, 5 and 10 clusters. It is apparent from these results that SSDU can noticeably enrich the top clusters among different categories of solutions quality. For instance, SSDU can improve the density of Acceptable, Medium and High quality solutions among the top 10 clusters by 61%, 20% and 38%, respectively.

Conclusions

We presented a new protein docking refinement protocol which is shown to effectively refine the quality of the solutions produced by first-stage global search methods like PIPER, which is implemented in the protein docking server ClusPro 2.0.

The SSDU algorithm we developed builds on our earlier SDU method[18,19] and works by underestimating the energy function in a set of local minima generated by local minimization methods. SSDU uses the minimum of the convex underestimator it generates to concentrate further sampling in its vicinity, assuming that this minimum resides close to the basin of the energy funnel spanned by the local minima. Four innovations introduced in this work are: (*i*) the use of our landscape analysis[21] to restrict underestimation in a lower-dimensional (typically 3D) permissive conformational subspace that avoids high-energy barriers; (*ii*) the use of density-based clustering to eliminate low-density regions and identify potential multiple high-density sub-clusters that are then separately refined by SSDU; (*iii*) the use of more flexible convex polynomial underestimators, and (*iv*) the use of a machine learning approach to effectively increase the number of Acceptable/Medium/High CAPRI quality solutions among the top clusters.

We demonstrate the effectiveness of SSDU on a comprehensive benchmark of 224 complexes containing Enzymes, Antibodies and Other Types of complexes. We show that SSDU is capable of increasing the number of quality solutions on a spectrum of different complexes in different quality categories defined by the CAPRI community-wide experiment. It was also shown that novelties introduced in this paper make SSDU superior to its predecessor SDU algorithm. Furthermore, we showed that we can further process the outputs to refine the quality of the solutions among the top clusters generated by SSDU, thereby potentially increasing the chance of picking a high quality representative from these clusters by other algorithms.

References

1. Huang, Y., Liu, S., Guo, D., Li, L. & Xiao, Y. A novel protocol for three-dimensional structure prediction of rna-protein complexes. *Scientific reports* **3** (2013).
2. Kozakov, D., Brenke, R., Comeau, S. R. & Vajda, S. PIPER: An FFT-based protein docking program with pairwise potentials. *Proteins* **65**, 392–406 (2006).
3. Heo, L., Lee, H. & Seok, C. Galaxyrefinecomplex: Refinement of protein-protein complex model structures driven by interface repacking. *Scientific reports* **6** (2016).
4. Gray, J. J. *et al*. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Molecular Biology* **331**, 281–299 (2003).
5. Mamonov, A. B. *et al*. Focused grid-based resampling for protein docking and mapping. *Journal of Computational Chemistry* **37**, 961–970, https://doi.org/10.1002/jcc.24273 (2016).
6. McCammon, J. Theory of biomolecular recognition. *Current Opinion in Structural Biology* **8**, 245–249 (1998).
7. Zhang, C., Chan, J. & DeLisi, C. Protein-protein recognition: Exploring the energy funnels near the binding sites. *Proteins* **34**, 255–267 (1999).
8. Tovchigrechko, A. & Vakser, I. How common is the funnel-like energy landscape in protein-protein interactions? *Protein Science* **10**, 1572–1583 (2001).
9. Leopold, P. E., Montal, M. & Onuchic, J. N. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc. Natl. Acad. Sci. USA* **89**, 8721–8725 (1992).
10. Bryngelson, J., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. Funnels, pathways, and the energy landscape of protein-folding - a synthesis. *Proteins–Structure, Function, and Genetics* **21**, 167–195 (1995).
11. Dill, K. Polymer principles and protein folding. *Protein Science* **8**, 1166–1180 (1999).
12. Tsai, C.-J., Kumar, S., Ma, B. & Nussinov, R. Folding funnels, binding funnels, and protein function. *Protein Sci.* **8**, 1981–1990 (1999).
13. Selzer, T. & Schreiber, G. New insights into the mechanism of protein-protein association. *Proteins – Structure, Function, and Genetics* **45**, 190–198 (2001).
14. Trosset, J.-Y. & Scheraga, H. A. Reaching the global minimum in docking simulations: A Monte Carlo energy minimization approach using Bezier splines. *PNAS* **95**, 8011–8015 (1998).
15. Camacho, C. J., Weng, Z., Vajda, S. & DeLisi, C. Free energy landscapes of encounter complexes in protein-protein association. *Biophys. J.* **76**, 1166–1178 (1999).
16. Camacho, C. J., Kimura, S. R., DeLisi, C. & Vajda, S. Kinetics of desolvation-mediated protein-protein binding. *Biophys J* **78**, 1094–1105 (2000).
17. Phillips, A., Rosen, J. & Dill, K. *From Local to* Global Optimization (P.M. *Pardalos et al. Eds*), chap. Convex Global Underestimation for Molecular Structure Prediction, 1–18 (Kluwer Academic Publishers, 2001).
18. Paschalidis, I. C., Shen, Y., Vakili, P. & Vajda, S. SDU: A semi-definite programming-based underestimation method for stochastic global optimization in protein docking. *IEEE Trans. Automat. Contr.* **52**, 664–676 (2007).
19. Shen, Y., Paschalidis, I. C., Vakili, P. & Vajda, S. Protein Docking by the Underestimation of Free Energy Funnels in the Space of Encounter Complexes. *PLoS Computational Biology* **4** (2008).
20. Nan, F. *et al* A subspace semi-definite programming-based underestimation (ssdu)method for stochastic global optimization in protein docking. In *Proceedings of the 53rd IEEE Conference on Decision and Control* (Los Angeles, California, 2014).
21. Kozakov, D. *et al*. Encounter complexes and dimensionality reduction in protein-protein association. *eLIFE* **3**, e01370 (2014). elifesciences.org/content/3/e01370/.
22. Iwahara, J. & Clore, G. M. Detecting transient intermediates in macromolecular binding by paramagnetic NMR. *Nature* **440**, 1227–1230 (2006).
23. Clore, G. M. Visualizing lowly-populated regions of the free energy landscape of macromolecular complexes by paramagnetic relaxation enhancement. *Molecular BioSystems* **4**, 1058–1069 (2008).
24. Fawzi, N. L., Doucleff, M., Suh, J.-Y. & Clore, G. M. Mechanistic details of a protein–protein association pathway revealed by paramagnetic relaxation enhancement titration measurements. *Proceedings of the National Academy of Sciences* **107**, 1379–1384 (2010).
25. Mirzaei, H. *et al*. Rigid body energy minimization on manifolds for molecular docking. *Journal of Chemical Theory and Computation* **8**, 4374–4380 (2012).
26. Mirzaei, H. *et al*. Energy minimization on manifolds for docking flexible molecules. *Journal of Chemical Theory and Computation* **11**, 1063–1076, https://doi.org/10.1021/ct500155t (2015).
27. Ahmadi, A. A., Olshevsky, A., Parrilo, P. A. & Tsitsiklis, J. N. NP-hardness of deciding convexity of quartic polynomials and related problems. *CoRR* abs/1012.1908 (2010).
28. Ahmadi, A. A. & Parrilo, P. A. A complete characterization of the gap between convexity and SOS-convexity. *SIAM Journal on Optimization* **23**, 811–833 (2013).
29. Borchers, B. CSDP, a C library for semidefinite programming. *Optimization Methods and Software* **11**, 613–623, https://doi.org/10.1080/10556789908805765 (1999).
30. Ester, M. *et al*. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD* **96**, 226–231 (1996).
31. Moghadasi, M. *et al*. The impact of side-chain packing on protein docking refinement. *Journal of Chemical Information and Modeling* **55**, 872–881, https://doi.org/10.1021/ci500380a (2015).
32. Moghadasi, M., Kozakov, D., Vakili, P., Vajda, S. & Paschalidis, I. C. A new distributed algorithm for side-chain positioning in the process of protein docking. In *Proceedings of the 52nd IEEE Conference on Decision and Control*, 739–744 (Florence, Italy, 2013).
33. Andrusier, N., Nussinov, R. & Wolfson, H. Firedock: Fast interaction refinement in molecular docking. *Proteins: Struct., Funct., Bioinf.* **69**, 139–59 (2007).
34. Pierce, B. & Weng, Z. Zrank: Reranking protein doking predictions with an optimized energy function. *Proteins: Struct., Funct., Bioinf.* **67**, 1078–86 (2007).
35. Schaefer, M. & Karplus, M. A comprehensive analytical treatment of continuum electrostatics. *J Phys Chem* **100**, 1578–1599 (1996).
36. Kortemme, T., Morozov, A. V. & Baker, D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein–protein complexes. *Journal of Molecular Biology* **326**, 1239–1259 (2003).
37. Chuang, G.-Y., Kozakov, D., Brenke, R., Comeau, S. R. & Vajda, S. DARS (Decoys As the Reference State) potentials for protein-protein docking. *Biophysical journal* **95**, 4217–27 (2008).
38. Shapovalov, M. & Dunbrack, R. Jr. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **19**, 844–858 (2011).
39. Vreven, T. *et al*. Updates to the integrated protein-protein interaction benchmarks: Docking benchmark version 5 and affinity benchmark version 2. *J. Molecular Biol.* 3031–3041 (2015).
40. Protein docking benchmark version 5 and affinity benchmark version 2. https://zlab.umassmed.edu/benchmark/ (2015).
41. Protein data bank. https://www.rcsb.org/pdb/home/home.do (2017).
42. Kozakov, D. *et al*. The ClusPro web server for protein-protein docking. *Nature Protocols* **12**, 255–278 (2017).
43. Janin, J. Assessing predictions of protein-protein interaction: The CAPRI experiment. *Protein Science* (2005).
44. Basu, S. & Wallner, B. DockQ: A quality measure for protein-protein docking models. *Plos One* (2016).

45. Kozakov, D., Clodfelter, K., Vajda, S. & Camacho, C. Optimal clustering for detecting near-native conformations in protein docking. *Biophysical Journal* **89**, 867–875 (2005).
46. Moal, I. H. *et al*. IRaPPA: information retrieval based integration of biophysical models for protein assembly selection. *Bioinformatics* **33**, 1806–1813 (2017).
47. Pfeiffenberger, E., Chaleil, R. A., Moal, I. H. & Bates, P. A. A machine learning approach for ranking clusters of docked protein-protein complexes by pairwise cluster comparison. *Proteins: Structure, Function, and Bioinformatics* **85**, 528–543 (2017).
48. Breiman, L. Random forests. *Machine learning* **45**, 5–32 (2001).

Acknowledgements

Research partially supported by the NIH/NIGMS under grants GM093147 and GM061867, by the NSF under grants CNS-1645681, CCF-1527292 and IIS-1237022, and by the ARO under grant W911NF-12-1-0390.

Author Contributions

M.M., F.N., K.L. and I.C.P. contributed to the development of the SSDU algorithm. K.L., D.K., S.V., P.V. and I.C.P. contributed to the dimensionality reduction approach on the rotational manifolds. M.M, S.Z., S.V, P.V. and D.K. contributed to the development of the local optimization protocol. S.Z. and M.M. contributed to the SSDU code and performed tests and experiments. S.Z., A.R. and I.C.P. contributed to the model selection and ranking method. P.V., S.V., D.K. and I.C.P. co-supervised the project. S.Z., M.M., A.R. and I.C.P. co-wrote the manuscript. All authors reviewed and commented on the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-23982-3.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.