Knowledge of Native Protein–Protein Interfaces Is Sufficient To Construct Predictive Models for the Selection of Binding Candidates

Petr Popov, Sergei Grudinin

▶ **To cite this version:**

Petr Popov, Sergei Grudinin. Knowledge of Native Protein–Protein Interfaces Is Sufficient To Construct Predictive Models for the Selection of Binding Candidates. Journal of Chemical Information and Modeling, 2015, 55 (10), pp.2242-2255. 10.1021/acs.jcim.5b00372 . hal-01229886

HAL Id: hal-01229886
https://inria.hal.science/hal-01229886

Submitted on 10 Sep 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

This document is the Accepted Manuscript version of a Published Work that appeared in final form in Journal of Chemical Information and Modeling, copyright © American Chemical Society after peer review and technical editing by the publisher. To access the final edited and published work see https://pubs.acs.org/doi/full/10.1021/acs.jcim.5b00372.

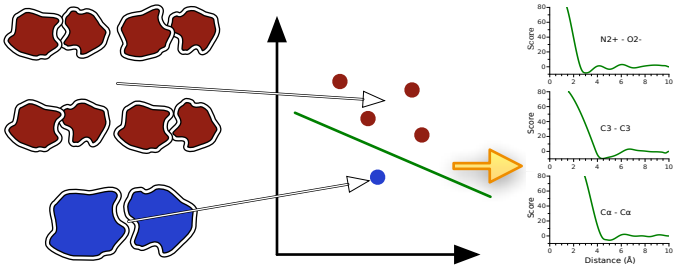
Knowledge of Native Protein-Protein Interfaces is Sufficient to Construct Predictive Models for the Selection of Binding Candidates

Petr Popov and Sergei Grudinin*

Univ. Grenoble Alpes, LJK, CNRS, LJK, Inria, F-38000 Grenoble, France

E-mail: Sergei.Grudinin@inria.fr

Abstract



Selection of putative binding poses is a challenging part of virtual screening for protein-protein interactions. Predictive models to filter out binding candidates with the highest binding affinities comprise scoring functions that assign a score to each binding pose. Existing scoring functions are typically deduced collecting statistical information about interfaces of native conformations of protein complexes along with interfaces of a large generated set of non-native conformations. However, the obtained scoring functions become biased toward the method used to generate the non-native conformations, i.e. they may not recognize near-native interfaces generated with a different method.

Present study demonstrates that knowledge of only native protein-protein interfaces is sufficient to construct well-discriminative predictive models for the selection of binding candidates. Here, we introduce a new scoring method that comprises a knowledge-based potential called *KSENIA* deduced from the structural information about the native interfaces of 844 crystallographic protein-protein complexes. We derive *KSENIA* using convex optimization with a training set composed of native protein complexes and their near-native conformations that are obtained using deformations along the low-frequency normal modes. As a result, our knowledge-based potential has only marginal bias toward a method to generate putative binding poses. Furthermore, *KSENIA* is smooth by construction, which allows to use it along with a rigid-body optimization to refine the binding poses. Using several test benchmarks we demonstrate that our method discriminates well native and near-native conformations of protein complexes from the non-native ones. Our methodology can be easily adapted to the recognition of other types of molecular interactions, such as protein-ligand, protein-RNA, etc. *KSENIA* will be made publicly available as a part of the SAMSON software platform at `https://team.inria.fr/nano-d/software`.

1 Introduction

Protein-protein interactions play crucial role in the human interactome, orchestrating most of the signaling network processes. Abrupt changes in protein-protein interactions lead to various kind of diseases, which makes protein structure prediction an important challenge in rational drug design. However, generally it is very difficult to experimentally obtain structures of protein complexes, thus computational molecular docking techniques are often used nowadays for protein-protein structure prediction. Typically, molecular docking as an integral part of the drug discovery process involves the scoring stage, where one selects the best putative binding candidates from the set of binding poses by assigning the score or the energy value E to each candidate. The scoring stage incorporates sophisticated scoring functions,[1] which are obtained with the empirical force-fields or using information derived from experimentally obtained structures of protein complexes. The latter type of scoring functions belongs to the family of the *knowledge-based* or *statistical* scoring functions. The majority of modern knowledge-based scoring functions for the protein-protein interactions are developed following the observation that the distances between the atoms in experimentally determined structures follow the Boltzmann distribution.[2] More precisely, using ideas from statistical theory of liquids, effective potentials between atoms are extracted using the inverse Boltzmann relation, $E_{ij}(r) = -k_B T \log(P_{ij}(r)/Z)$, where k_B is the Boltzmann constant, $P_{ij}(r)$ denotes the probability to find two atoms of certain types i and j at a distance r, and Z denotes the probability distribution in the reference state. The latter is the thermodynamic equilibrium state of the protein when all interactions between the atoms are set to zero. The score of a protein conformation is then given as a sum of the effective potentials between all pairs of atoms. Although this concept is old and originates from the work of Tanaka and Scheraga,[3] Miyazawa and Jernigan,[4] and Sippl,[5–7] it is still under debates.[8–11] Particularly, the computation of the reference state is a challenging problem.[12] Although some assumptions were made to ease the expression of the reference state for protein monomers,[5,13–15] to deduce scoring functions for the protein-protein docking, one usually computes the reference state based on a large set of generated non-native conformations of protein complexes (decoys).[16,17] Another type of statistical potentials is constructed using the discriminative machine learning, specifically, the linear programming approach.[18–23] The basic idea behind this approach is to solve a system of inequalities that demand the energy of the native conformation to be lower than the energy of all the decoy conformations for a particular complex, $E(P^{\text{native}}) - E(P_i^{\text{decoy}}) < 0,\ \forall P_i^{\text{decoy}} \in \mathbf{P}^{\text{decoy}}$. Although this approach circumvents the computation of the reference state, its success critically depends on the chosen set of decoy conformations $\mathbf{P}^{\text{decoy}}$. Thereby, the obtained statistical potential depends on the sampling algorithm used to generate the decoy conformations and, generally, might not distinguish the native structures equally well from decoys obtained by another sampling algorithm.

In this study we discovered that knowledge of only native protein-protein interfaces is sufficient to construct well-discriminative predictive models for the selection of putative binding candidates. Namely, we introduce a new scoring method that comprises a knowledge-based potential called *KSENIA* deduced from the structural information about the native interfaces of 844 crystallographic protein-protein complexes. As a result, our approach does not require neither the computation of the reference state nor the ensemble of non-native complexes. Thus, it can have only a marginal bias toward a method to generate putative binding poses. To the best of our knowledge, this is the first investigation of the knowledge-based potential that needs no information derived from non-native protein-protein interfaces. More precisely, we use convex optimization to train the knowledge-based potential on sets of near-native conformations with the average root mean square deviation (RMSD) between monomers of 1 Å. These are composed using the deformations along the directions of low-frequency normal

modes computed at the native conformations. We demonstrate that the obtained knowledge-based potential is capable to distinguish the native and near-native protein-protein interactions from the non-native ones. Given that rigid-body minimization refinement improves the scoring performance,[24] we also implement a rigid-body optimization protocol using the derived knowledge-based potential. Finally, we verify the robustness of our method on several protein-protein docking benchmarks.

2 Theoretical Basis

We consider N native protein-protein complex conformations P_i^{native}, $i = 1..N$. For each protein complex i we generate D decoys, P_{ij}^{decoy}, $j = 1..D$, where the first index runs over different protein complexes and the second index runs over generated decoys. Then we find a linear *scoring functional* F, defined for all possible complexes, such that for each native complex i and its decoy j the following inequality holds:

$$F(P_i^{\text{native}}) < F(P_{ij}^{\text{decoy}}) \tag{1}$$

We express the scoring functional which fulfills these assumptions in the following form:

$$F(P) = \sum_{k=1}^{M} \sum_{l=k}^{M} \int_0^{r_{\max}} n^{kl}(r) U^{kl}(r)\, dr, \tag{2}$$

where $n^{kl}(r)$ is the *number density* of atom pairs at a distance r between two atoms of types k and l (kl-pair), with one atom located in the larger protein (receptor), and the other atom located in the smaller protein (ligand). Here, M is the total number of different atom types. We used $M = 20$ atom types provided by Huang and Zou,[16] which were defined by the classification of all heavy atoms in standard amino acids according to their element symbol, aromaticity, hybridization, and polarity. The functions $U^{kl}(r)$ are unknown *scoring potentials*, which we determine below.The number density $n^{kl}(r)$ is computed as a sum over all kl-pairs in a given protein complex via:

$$n^{kl}(r) = \sum_{ij} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(r-r_{ij})^2}{2\sigma^2}} \tag{3}$$

Here, each kl-pair at a distance r_{ij} is represented by a Gaussian centered at r_{ij} with the standard deviation of σ, which takes into account possible inaccuracies and thermal fluctuations in the protein structure. In our work we chose $\sigma = 0.4$ Å, since this value demonstrated the best results in the holdout cross-validation tests[25] (see Section 2 of Supporting Information for more details). We considered only atom pairs at distances below the threshold distance $r_{\max} = 10$ Å. Using Eq. (3), we can rewrite the scoring functional F (see Eq. (2)) as the sum over all kl-pairs of atoms i and j at a distance r_{ij}:

$$F(P) = \sum_{ij} \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^{r_{\max}} e^{-\frac{(r-r_{ij})^2}{2\sigma^2}} U^{kl}(r)\, dr = \sum_{ij} \Upsilon^{kl}(r_{ij}) \tag{4}$$

We will refer to the functions

$$\Upsilon^{kl}(r) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^{r_{\max}} e^{-\frac{(x-r)^2}{2\sigma^2}} U^{kl}(x)\, dx, \tag{5}$$

which are the Gauss transform of the scoring potentials $U^{kl}(x)$, as to the *scoring functions*.

In order to determine unknown scoring potentials $U^{kl}(r)$ (see Eq. (2)), we decompose them along with the number densities $n^{kl}(r)$ in a polynomial basis:

$$\begin{aligned} U^{kl}(r) &= \sum_q w_q^{kl} \psi_q(r), \quad r \in [0; r_{\max}] \\ n^{kl}(r) &= \sum_q x_q^{kl} \psi_q(r), \quad r \in [0; r_{\max}], \end{aligned} \tag{6}$$

where $\psi_q(r)$ are orthogonal basis functions on the interval $[r_1; r_2]$, and w_q^{kl} with x_q^{kl} are the expansion coefficients of $U^{kl}(r)$ and $n^{kl}(r)$, respectively. Here, we use a set of shifted rectangular functions as the basis.[26] Given this, the scoring functional F (see Eq. (2)) can be expanded up

to the order Q as:

$$F(P) \approx \sum_{k=1}^{M}\sum_{l=k}^{M}\sum_{q}^{Q} w_q^{kl} x_q^{kl} = (\mathbf{w}\cdot\mathbf{x}), \quad \mathbf{w},\mathbf{x} \in \mathbb{R}^{Q\times M\times (M+1)/2}, \tag{7}$$

where we use the order of expansion $Q = 40$. We will refer to the vector $\mathbf{w}$ as to the *scoring vector* and to the vector $\mathbf{x}$ as to the *structure vector*. Then, we can rewrite the set of inequalities (1) as a *soft-margin quadratic optimization problem*:[27]

$$\begin{array}{l}\text{Minimize (in } \mathbf{w},\, b_i,\, \xi_{ij}\text{): } \frac{1}{2}\mathbf{w}\cdot\mathbf{w} + \sum_{ij} C_{ij}\xi_{ij} \\ \text{Subject to:} \\ y_{ij}\left[\mathbf{w}\cdot\mathbf{x}_{ij} - b_i\right] - 1 + \xi_{ij} \geq 0,\ i = 1..N,\ j = 0..D \\ \xi_{ij} \geq 0\end{array} \tag{8}$$

Here, index i runs over different protein complexes and index j runs over different conformations of the i-th protein complex. Particularly, protein conformations with $j = 0$ are native with the corresponding constants $y_{i0} = +1$ and protein conformations with $j = 1..D$ are the decoys with the corresponding constants $y_{ij} = -1$. Parameters C_{ij} can be regarded as regularization parameters, which control the importance of different structure vectors. We found the optimal values of C_{ij} parameters using the holdout cross-validation procedure[25] (see Section 2 of Supporting Information). The scoring vector $\mathbf{w}$, the offset vector $\mathbf{b}$ and the slack variables ξ_{ij} are the parameters to be optimized. The size of the optimization problem is determined by the dimensionality of the structure and the scoring vectors, which is equal to $Q\times M\times(M+1)/2 = 8400$, and by the size of the training set, $N = 844$ and $D = 225$. The latter is composed based only on local information about the native interfaces of protein-protein complexes and no other information is used (see Section 3.3). We solve problem (8) in its *dual form* using the *block sequential minimal optimization* (BSMO) algorithm as explained elsewhere.[26] Finally, given the solution of problem (8), i.e. the scoring vector $\mathbf{w}$, one may restore the scoring potentials $U^{kl}(r)$ (see Eq. (6)), the scoring functions $\Upsilon^{kl}(r)$ (see Eq. (5)), and compute the score of a protein complex according to Eq. (4).

3 Materials and Methods

3.1 Rigid-Body Minimization

The scoring functions $\Upsilon^{kl}(r)$ (see Eq. (5)) are smooth by construction. This fact allows to use these functions for the structure optimization. More accurately, for a given kl-pair of atoms at a distance r_{ij}, the negative gradient $-\nabla\Upsilon^{kl}(r_{ij})$ could be regarded as the force with which one atom acts on the other atom. Thus, one may use the set of derived functions $\Upsilon^{kl}(r)$ to optimize a particular conformation of a protein complex until a local minimum is reached, provided $\nabla\Upsilon^{kl}(r_{ij}) = 0$ for each pair of atoms. Since special calibration is required to retain structure integrity of a complex, a more relevant structure optimization would be the *rigid-body* optimization, where instead of force minimization over each pair of atoms, one minimizes the net force and the net torque acting on each monomer. The rigid-body optimization with functions $\Upsilon^{kl}(r)$ could be useful as a refinement step to process docking predictions. It was shown that rigid-body refinement could improve docking predictions dramatically.[24] In contrast to our scoring functions $\Upsilon^{kl}(r)$, most of modern statistical potentials are not differentiable.[14,28–30] Thereby, to perform structure optimization with such potentials, one either uses a smooth interpolation of potentials, or employs various derivative-free optimization strategies, e.g. Nelder-Mead[31] or Powell[32] methods and their modifications, where the convergence rate is much slower compared to first- or higher- order optimization strategies. Following this idea, we implemented the local rigid-body minimization protocol to explore whether such an optimization improves scoring capabilities of *KSENIA* . General work-flow for the local rigid-body minimization is listed in Table 1.

3.2 Normal Modes

Let us consider a system of N particles with $3N$ degrees of freedom near the equilibrium state

Table 1: The rigid-body minimization work-flow.

1	Set initial parameters for the structure optimization.
2	Compute the score U_k of the current conformation and the descent direction $\mathbf{d}_k$ in the rigid-body space.
3	Find an appropriate step size α and make a step toward the descent direction: $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha \mathbf{d}_k$.
4	Repeat steps 2-3 until desired tolerance or maximum number of iterations is achieved.
5	Take the last computed score as the final score of the optimized conformation.

$\mathbf{x_0}$. The potential energy of the system can be approximated as a quadratic form:

$$U(x_1, x_2, \ldots, x_{3N}) = U(\mathbf{x_0}) + \frac{1}{2}\sum_{i=1}^{3N}\sum_{j=1}^{3N} F_{ij} x_i x_j, \tag{9}$$

where elements of the matrix of the quadratic form $F_{ij} = \left(\frac{\delta^2 U}{\delta x_i \delta x_j}\right)_{\mathbf{x_0}}$ are the force constants at the equilibrium state $\mathbf{x_0}$. There exist a different set of coordinates y_i, where both the kinetic K and the potential U energies have the *diagonal form* and thus the Newton's equations of motion are uncoupled. This means that the solution for the equations of motion for each coordinate can be obtained separately. These coordinates y_i are called the *normal coordinates*, and the corresponding energy terms have the following form:

$$\begin{aligned} U(y_1, y_2, \ldots, y_{3N}) = &\quad U(\mathbf{x_0}) + \frac{1}{2}\sum_{i=1}^{3N} \lambda_i y_i^2, \\ K(y_1, y_2, \ldots, y_{3N}) = &\quad \frac{1}{2}\sum_{i=1}^{3N} \dot{y}_i^2 \end{aligned} \tag{10}$$

The transition matrix between the two coordinate bases is obtained via diagonalization of matrix $\mathbf{M}^{-\frac{1}{2}}\mathbf{F}\mathbf{M}^{-\frac{1}{2}} = \mathbf{L}\mathbf{D}\mathbf{L}^{\mathbf{T}}$:

$$U - U(\mathbf{x_0}) \equiv \frac{1}{2}\mathbf{x}^{\mathbf{T}}\mathbf{F}\mathbf{x} = \frac{1}{2}\mathbf{x}^{\mathbf{T}}\mathbf{M}^{\frac{1}{2}}\mathbf{L}\mathbf{D}\mathbf{L}^{\mathbf{T}}\mathbf{M}^{\frac{1}{2}}\mathbf{x} = \frac{1}{2}\mathbf{y}^{\mathbf{T}}\mathbf{D}\mathbf{y}, \tag{11}$$

where $\mathbf{M}$ is the diagonal mass matrix, i.e. $M_{ij} = m_i \delta_{ij}$. Thus, the connection between the two coordinate systems is given as a linear transformation

$$\mathbf{x} = \mathbf{M}^{-\frac{1}{2}}\mathbf{L}\mathbf{y} \tag{12}$$

Normal coordinates provide a convenient way to describe molecular fluctuations of a system near the equilibrium state. Particularly, the evolution of the system in the normal basis is the superposition of the independent harmonic oscillations along each normal coordinate y_i. Such oscillations are called normal modes[33] and are expressed as:

$$y_i(t) = A_i \cos(\omega_i t + \delta_i), \tag{13}$$

where $\omega_i \equiv \sqrt{D_{ii}}$ and δ_i correspond to the frequency and the phase of the i-th mode, respectively. The factor $A_i = \sqrt{2k_BT}/\omega_i$ is the amplitude of the fluctuation. Given the transition matrix $\mathbf{L}$ between the two bases (see Eq. (12)), oscillations in the Cartesian basis can be written as:

$$x_k(t) = L_{ki}(A_i \cos(\omega_i t + \delta_i))/\sqrt{m_k} \tag{14}$$

Thus, all atoms in a molecule for a given mode i oscillate with the same frequency and phase. However, the amplitude of the fluctuation of the Cartesian coordinate x_k, corresponding to the oscillation of the mode y_i, is different for each coordinate k and is defined by the i-th column of the transition matrix $\mathbf{L}$:

$$\begin{aligned} \langle x_k^2 \rangle_i &= L_{ki}^2 A_i^2 \langle \cos(\omega_i t + \delta_i)^2 \rangle / m_k \\ &= \frac{1}{2m_k} L_{ki}^2 A_i^2 = L_{ki}^2 \frac{k_B T}{m_k \omega_i^2} \end{aligned} \tag{15}$$

When all the modes are active, the amplitude of the fluctuation of the Cartesian coordinate x_k reads:

$$\langle x_k^2 \rangle = \frac{k_B T}{m_k} \sum_i \frac{L_{ki}^2}{\omega_i^2} \tag{16}$$

We use this theoretical framework to construct the training set of protein-protein complexes. A deeper discussion of normal modes analysis and its applications in structural biology can be found elsewhere.[33–37]

3.3 Training Set

3.3.1 Native Complexes

We used the training database of 851 non-redundant protein-protein complex structures prepared by Huang and Zou.[16] This database contains protein-protein complexes extracted from the PDB[38] and includes 655 homodimers and 196 heterodimers. We updated three PDB structures from the original training database: 2Q33 supersedes 1N98, 2ZOY supersedes 1V7B, and 3KKJ supersedes 1YVV. The training database contains only crystal dimeric structures determined by X-ray crystallography at resolution better than 2.5 Å. Each chain of the dimeric structure has at least 10 amino acids, and the number of interacting residue pairs, as defined as having at least 1 heavy atom within 4.5 Å, is at least 30. Each protein-protein interface consists only of 20 standard types of amino acids. No homologous complexes were included in the training database. Two protein complexes were regarded as homologues if the sequence identity between receptor-receptor pairs and between ligand-ligand pairs was $> 70\%$. Finally, Huang and Zou[16] manually inspected the training database and left only those structures that had no artifacts of crystallization.

3.3.2 Near-native Decoys

To exclude any possible bias to computational methods and potentials for the generation of putative binding poses, we do not use standard methods for the generation of non-native decoys, but instead we construct our training set using structural information about only protein complexes in their native conformations. For the initial set of 844 native protein complexes we generated near-native conformations, i.e. conformations within RMSD = 3 Å, for each native complex as follows. First, given the coordinate vector $\mathbf{X}^{\text{native}}$ of each monomer in a protein complex, we computed its ten lowest-frequency normal modes. Then, we formed fifteen near-native conformations for each monomer using the linear combinations of these modes :

$$\hat{\mathbf{X}} = \mathbf{X}^{\text{native}} + \sqrt{k_B T}\mathbf{M}^{-\frac{1}{2}} \sum_{i=1}^{10} r_i \frac{\mathbf{L}_i}{\omega_i}, \qquad (17)$$

where $\sqrt{k_B T}$ is the temperature factor, $\mathbf{M}$ is the diagonal mass matrix, i.e. $M_{kl} = m_k \delta_{kl}$, r_i is the random weight for each mode ranging from -1 to 1, $\mathbf{L}_i$ is the i-th column of the transition matrix between the Cartesian and the normal mode bases, and ω_i is the frequency of the i-th mode. The temperature factor $\sqrt{k_B T}$ affects the amplitude of the deformation, hence, too large temperatures cause a monomer to deform significantly breaking the covalent bonds. We tested several values of the temperature factor and found the optimal value of $\sqrt{k_B T}$ to be 10 $\text{kJ}^{\frac{1}{2}}$ (see Section 2 of Supporting Information). To ensure the absence of non-relevant conformations, we measured the RMSD between the native and the generated conformations. Indeed, the average RMSD is equal to 1.02 Å, which means that the deformations with the given temperature factor keep all generated conformations non-disrupted. At the last step, we combined conformations $\hat{\mathbf{X}}$ of two monomers representing one protein complex, resulted in $15 \times 15 = 225$ near-native conformations. To summarize, the composed training set to derive the knowledge-based potential contains 844 assemblies, where each assembly consists of one native protein complex and 225 generated near-native conformations.

We used the MMTK library[39] to perform the normal mode analysis for protein molecules and the OPLS-UA force-field[40] to compute the force constants (see Eq. (9)). Since normal modes are defined for the equilibrium state of the system, we minimized each monomer of a dimer in a vacuum using 50 steps of the steepest descent algorithm with the relative energy tolerance of $1e-3$ and the cut-off distance for all non-bonded interactions of 5 Å. We chose such a relatively small number of minimization steps in order to not significantly deform the X-Ray

structure of a monomer. Indeed, the RMSD between the initial and the minimized monomer structures did not exceed 0.5 Å. Given each monomer near the equilibrium state, we used the Fourier subspace for the reduced-basis normal modes computations.[41] We picked up ten first low-frequency modes from the Fourier basis to generate different local deformations of the protein complexes. We should note that we excluded the first six modes that correspond to the rigid-body motion.

Finally, we want to stress that *all* generated conformations represent *near-native* protein structures. Indeed, we use directions along the slowest normal modes to locally deform the monomers, however, the orientations of the monomers with respect to each other are fixed. Since all the monomer conformations differ only slightly from the native monomers (the average RMSD is 1.02 Å), the interaction interfaces of all generated complexes undergo moderate changes keeping the major part of the native contacts. To conclude, we composed the training set based only on local information about the native interfaces and no other information was used. In the Results section we demonstrate the knowledge-based potential for protein-protein interactions derived using this training set.

3.4 Test Benchmarks

Here, we describe the composed benchmarks to test and validate the *KSENIA* potential. For the accurate validation it is very important to ensure that the test and the training sets do not overlap. The first test benchmark consists of the complexes from the training set, however with different binding interfaces. The other benchmarks are built using the interfaces from the protein-protein docking benchmark v. 2, 3 and 4.[42] We ran the *align* program from the FASTA2 package[43] in order to calculate the number of the homologous interfaces between the training set and the protein-protein docking benchmark. There are no intersections between the two sets with the sequence similarity greater than 70%, there is one pair (1LEW–2OZA) with the similarity of 61 % and 12 pairs (1AVW–1AVX, 1BIS– 2B4J, 1CSO– 3SGQ, 1E96– 1E96, 1KLJ– 1FAK, 1MCV– 1FLE, 1SBW– 2UUY, 1SCJ– 2SNI, 1SFI– 2UUY, 1UJZ– 7CEI, 1ZBX– 1ZHI, 2NGR– 1GRN) with the similarity in between 50% and 60%. For each of the benchmarks, we evaluate the *success rate* of our method with respect to other tested methods, which is defined as the percentage of protein complexes for which docking predictions of a certain quality are ranked at the top positions.

3.4.1 Hex Test Benchmark

For the first test, we constructed a rigid-body benchmark starting from the native structures in the training set. More precisely, to generate decoys we used the Hex rigid-body docking program.[44,45] For the Hex input, we used polar Fourier shape expansions to polynomial order of 31, the real-space angular search step of 7.5°, the radial search range of 40 Å with a translation step of 2.5 Å and the subsequent sub-step of 1.25 Å. We ran Hex for each native complex in the training set and clustered the docking solutions with a threshold of 8 Å. Top 200 docking predictions were added to the test benchmark in addition to the native complexes, resulting in $201 \times 844 = 169,644$ protein complexes. Finally, we evaluated the success rate of the Hex scoring function on the constructed benchmark according to the quality of the docking poses. Here we define the quality according to the value of RMSD of the backbone atoms of the ligand (L_{RMSD}) after the receptors in the native and the decoy conformations have been optimally superimposed (see Table 2). To do so, we used the fast open-source RigidRMSD library[46] that computes RMSDs given spatial transforms of the docking poses.

Table 2: Quality with respect to the value of L_{RMSD}.

Quality	L_{RMSD}(Å)
1	$L_{\mathrm{RMSD}} \leq 1$
2	$1 < L_{\mathrm{RMSD}} \leq 5$
3	$5 < L_{\mathrm{RMSD}} \leq 10$

3.4.2 Zdock Test Benchmark

For the second test benchmark we used the protein-protein docking benchmark v3.0 composed by Hwang *et al.*, which consists of 124 non-redundant protein-protein complexes.[47] Then, we employed Zdock v.3.0.1 rigid-body docking software,[48] which uses a grid-based representation of two proteins and a three-dimensional fast Fourier transform to explore the search space of rigid-body docking positions. We used the bound conformation of each monomer in the benchmark for the Zdock input, randomly set initial protein orientations and used the default parameters for the docking predictions. Finally, we chose 2000 best generated rigid-body docking poses according to the Zdock v.3.0.1. scoring function for each complex. Thus, the second test benchmark consists of $124 \times 2,000 = 248,000$ protein complexes.

To evaluate the success rate of this scoring function on the constructed benchmark, we use the Critical Assessment of PRediction of Interactions (CAPRI) criterion[49] for a correct prediction (Table 3). This is a more sophisticated criterion compared to the one used above. More precisely, in addition to the ligand-RMSD, it involves the fraction of native contacts in the docking prediction f_{nat}, and the interface RMSD, I_{RMSD}. The f_{nat} parameter is the ratio of the number of native residue-residue contacts in the predicted complex to the number of residue-residue contacts in the crystal structure. A pair of residues from different monomers are considered to be in contact if they are within 5 Å from each other. The I_{RMSD} parameter is the RMSD of the interface region between the predicted and native structures after optimal superimposition of the backbone atoms of the interface residues. A residue is considered as the interface residue if any atom of this residue is within 10 Å from the other partner.

3.4.3 ItScoreTest Benchmark

Following Huang et al., we generated the ItScore test set using 91 protein complexes and Zdock v.2.1 docking program as it is described in the original ItScore paper.[16] Overall, this test set comprises 2000 decoys together with the native structure per each of 91 protein complexes and we will refer to this set as to the ItScore test benchmark. We evaluated the success rates on this set using the CAPRI prediction quality criterion (Table 3).

3.4.4 Rosetta Test Benchmark

Gray *et al.* generated the Rosetta benchmark using 54 complexes from the protein-protein docking benchmark version 0.0[50] in both the bound and the unbound conformations. For each complex, the authors generated 1,000 bound and 1,000 unbound decoys following the flexible docking protocol, which is a part of the RosettaDock suite.[51] The first step in the protocol is the random translation and rotation of one of the proteins constituting the complex. Afterwards, the side chains are optimized simultaneously with the rigid-body displacement of the protein. Finally, the full-atom minimization is performed to refine the conformation of the complex. We calculated the success rate of the RosettaDock protocol using the same quality criterion as in CAPRI[49] (Table 3). Both the bound and the unbound Rosetta benchmarks consist of $54 \times 1,000 = 54,000$ protein complexes.

3.4.5 SwarmDock Test Benchmark

Finally, we tested our scoring function on the unbound decoy set prepared by Moal et al.[1] and generously provided by Mieczyslaw Torchala from Biomolecular modelling group of the Francis Crick Institute. The SwarmDock decoy set was generated using the SwarmDock docking server[52] with initial structures in the unbound state taken from the protein-protein docking benchmark v.4.0.[42] In total, the decoy set consists of about 500 conformations per each of 176 protein complexes, and there is at least one correct prediction of at least acceptable quality (according to the CAPRI criterion, Table 3) for 122 complexes. Using this benchmark, Moal et al. compared performance of 115 various scoring functions,[1] including finite- and coarse- grained docking potentials, their

Table 3: CAPRI criterion to estimate the quality of docking predictions.

Quality	Condition
1	$f_{\text{nat}} \geq 0.5$ and ($L_{\text{RMSD}} \leq 1.0$ or $I_{\text{RMSD}} \leq 1.0$)
2	($0.3 \leq f_{\text{nat}} < 0.5$) and ($L_{\text{RMSD}} \leq 5.0$ or $I_{\text{RMSD}} \leq 2.0$) or ($f_{\text{nat}} \geq 0.5$ and $L_{\text{RMSD}} > 1.0$ and $I_{\text{RMSD}} > 1.0$)
3	($0.1 \leq f_{\text{nat}} < 0.3$ and ($L_{\text{RMSD}} \leq 10.0$ or $I_{\text{RMSD}} \leq 4.0$) or ($f_{\text{nat}} \geq 0.3$ and $L_{\text{RMSD}} > 5.0$ and $I_{\text{RMSD}} > 2.0$)

constitutive terms, molecular mechanics energy functions, and protein folding potentials. This comparison provides a good reference point of predictive capabilities for different scoring functions.[53] To be consistent with the assessment pipeline of Moal et al., in this test we did not use the refinement procedure and calculated the success rates of *KSENIA* solely based on the initial scores of the decoys. Following Moal et al., for the test we used only complexes from the protein-protein docking benchmark "v.4.0 update" in order to exclude bias in performance of scoring functions trained on the previous versions of this benchmark. This update consists of 52 protein-protein complexes non-homologous to the previous version of the benchmark.

3.5 Computational Considerations

Here we briefly describe the computational details of the potential derivation and the scoring procedure with the *KSENIA* potential. The potential is derived off-line by solving problem (8) in its *dual form* using the BSMO algorithm as explained elsewhere.[26] This procedure is iterative and takes minutes to hours depending on the desired convergence, the number of available CPU cores and the amount of memory. Then, for practical applications, as it is described in more detail in Supporting Information, we use a cubic spline interpolation of the obtained potential. More precisely, for each of 210 types of interactions, the spline coefficients are pre-computed off-line at forty equidistant points.[54] To compute the value of the scoring function for a protein complex, we construct the list of intractions using the linked-cell neighbourlist algorithm. Then, for each pair of interactions from the list, we evaluate the value of the potential and, if necessary, its derivative from the spline interpolation.

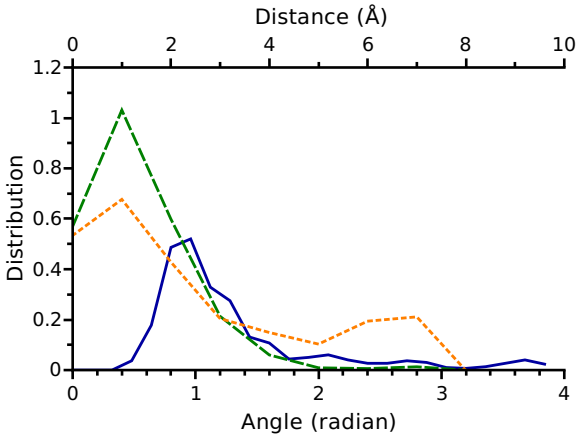


Figure 1: Distribution of the interval lengths in the four-dimensional manifold where the partial derivatives of the scoring functional are the constant-sign functions. These distributions are computed using the native structures in the training set. Blue, solid: interval length for the d-coordinate, which is the distance between the centers of mass of two monomers. Green, dashed: interval length for the α-coordinate, which is the angle of rotation of the ligand about the axis connecting the centers of mass. Orange, dotted: interval length for the β- and γ- coordinates, which are the angles of rotation about two other orthogonal axes.

4 Results

4.1 Scoring Functional

Figure 2 presents three derived scoring functions (dashed) for different atom pairs. As one can see, at short separation distances the scoring functions tend to zero. This is the arti-

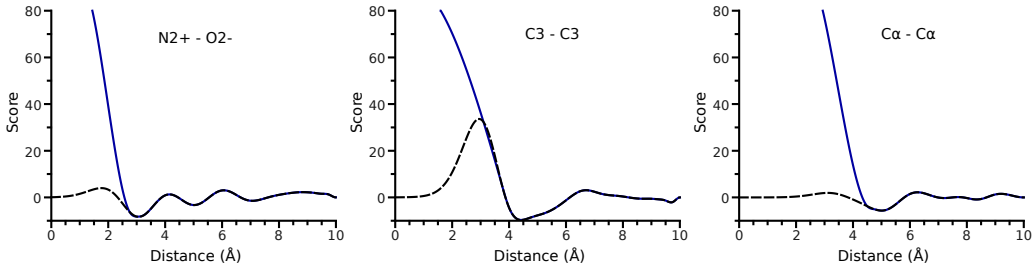


Figure 2: Examples of the derived distance-dependent scoring functions between atoms of types N2+ − O2- , C3 − C3 and $C_\alpha - C_\alpha$, respectively. Here, N2+ are guanidine nitrogens with two hydrogens, O2- are oxygens in carboxyl groups, C3 are aliphatic carbons bonded to carbons or hydrogens only and C_α are the backbone C_α atoms. Black, dashed: initially derived scoring functions without taking into account the absence of statistics at short distances. Blue, solid: re-defined scoring functions that take into account the absence of statistics at short distances.

fact of the training set, and is mainly caused by the absence of observations of atom pairs at distances close to zero. However, we want our scoring functions to be able to penalize conformations in which steric clashes between the monomers are present. Thus, we re-define the scoring functions at short distances to form artificial potential barriers (see Section 1 of Supporting Information). The initial scoring functions along with the modified scoring functions are shown in Figure 2. We refer to the latter as to *KSENIA*, which stands for ***K****nowledge-based* ***S****coring function* ***E****mploying only* ***N****ative* ***I****nterfaces* .

The scoring functional F (see Eq. (4)) of a particular protein complex P is computed as the sum of separate scores for each pair of atoms within the cutoff distance $r_{\max}$. Thus, F, as a function of $3 \times (N_A + N_B)$ variables, where N_A and N_B are the numbers of atoms in molecules A and B respectively, is not identically zero only in the conformational volume where at least one pair of atoms is within $r_{\max}$ distance. Since *KSENIA* typically possesses several maxima and minima (see Fig. 2), F is likely to be a rugged function in this volume.[55] However, we want to demonstrate that since our scoring functions were derived from the local deformations of the native conformations, the scoring functional F is smooth at least in the neighborhood of the native conformation. To show this, we explored the behavior of the scoring functional F in the four-dimensional manifold of the $3 \times (N_A + N_B)$ conformational space. Namely, given two monomers, one of which is fixed, we consider four coordinates corresponding to the rigid-body degrees of freedom: the distance d between the centers of mass of the two monomers, the rotation of the free molecule about the axis connecting the centers of mass by an angle α , and two rotations about two other orthogonal axes by angles β and γ. Then, starting from the native conformation of the complex $(d_0, \alpha_0, \beta_0, \gamma_0)$, we calculate partial derivatives in the vicinity of this conformation. More precisely, we sample the first partial derivative $\dfrac{\delta F(d, \alpha, \beta, \gamma)}{\delta e}$ at points $\{e_0 \pm \epsilon, e_0 \pm 2\epsilon, e_0 \pm 3\epsilon, \dots\}$, where $e \in \{d, \alpha, \beta, \gamma\}$, and ϵ is a sufficiently small positive value. At the point where the partial derivative changes its sign, we can not expect a gradient-based local minimization algorithm to find the nearest local minimum to the point $(d_0, \alpha_0, \beta_0, \gamma_0)$. Thus, one can characterize the smoothness of the scoring functional F at the point $(d_0, \alpha_0, \beta_0, \gamma_0)$ by four intervals $(e_0 - m\epsilon$, $e_0 + n\epsilon)$, where the partial derivative is a constant-sign function. Figure 1 shows the distribution of such interval lengths over the native conformations in the training set. The most probable size of the smooth region around the native conformation is 2.2 Å, 0.42 rad, 0.22 rad, 0.22 rad in four degrees of freedom, respectively. Practically, it means that the rigid-body minimization started from an arbitrary point within this region is expected to optimize the conformation corresponding to this point toward the

conformation corresponding to the local minimum of this region, assuming that F is convex in the neighborhood of the native conformation.

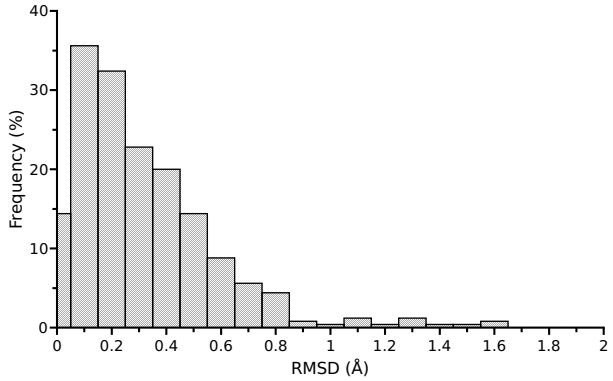


Figure 3: Histogram representing the distributions of the RMSDs between the native and minimized conformations in the training set using the rigid-body minimization protocol.

Finally, it remains to prove that the point representing the native conformation in the four-dimensional manifold lies close to the local minimum. To demonstrate this, we measured the RMSD between the native conformation and the conformation obtained after the rigid-body minimization with the *KSENIA* potential starting from the native conformation. Figure 3 shows the distribution of such RMSDs in the training set. As it could be seen, the minimized and native structures are very similar and the corresponding RMSD does not exceed 2 Å. Moreover, the most probable RMSD between the two conformations is 0.1 Å.

To summarize, we demonstrated that the scoring functional F is a smooth function in the vicinity of the native conformation. Hence, the rigid-body minimization is expected to improve predictions if started at an arbitrary point in this vicinity. Below, we provide numerical experiments that demonstrate the practical importance of the rigid-body minimization with *KSENIA*.

4.2 Performance on the Test Benchmarks

The aim of any scoring function is to differentiate the native and near-native conformations of protein complexes from the non-native ones. In this section we demonstrate that observing only the native protein complexes is sufficient to build a powerful and well-discriminative knowledge-based potential. Using six different protein-protein benchmarks described in Section 3.4, we evaluate the *success rate* of our method as described above. For each benchmark we also provide success rates of the widely-used scoring functions of Hex,[44] Zdock,[48] Rosetta,[51] ItScore-PP[16] and those tested by Moal et al.[1] as the reference.

4.2.1 Hex Test Benchmark

In the first test, we used the Hex test benchmark (see Section 3.4.1). Although the training set and this benchmark share the same native structures, their decoys are very different. More precisely, for the training, we generated *local deformations* at the protein-protein interfaces for all native complexes using directions along the low-frequency normal modes. On the other hand, to generate decoys for the test benchmark, we performed the exhaustive search in the six-dimensional space of rigid-body motions. Consequently, many different interfaces for each native complex are present. Furthermore, owing to the clustering of spatially close docking predictions, there are no similar interfaces in the test benchmark. Thus, the goal of the first test is to demonstrate that employing only local information about the native interfaces is sufficient to derive a well-discriminative scoring function.

We ranked all docking poses in the training set according to the values of the initial scoring functions and the values of *KSENIA*. Figure 4 presents the corresponding success rates for the top predictions. Clearly, the derived scoring functions predict the native interfaces very well, providing the success rates of more than 90% for the top one predictions. To explore if our scoring functions can distinguish correct interfaces (generated by Hex with quality-one, -two or -three) from the non-native ones, we removed the native structures from the test benchmark, leaving only predictions with non-zero rotational part of the spatial transform.

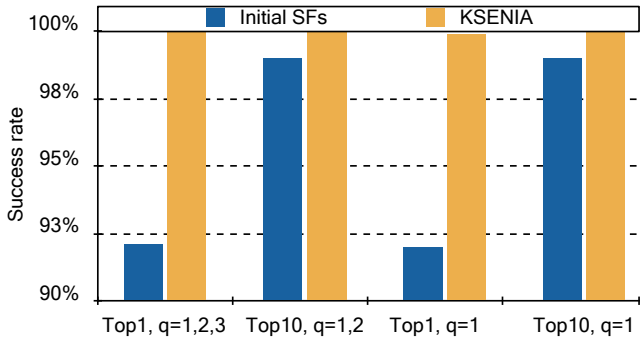


Figure 4: Performance of the scoring functions on the Hex test benchmark. Success rates of the initial scoring functions (Initial SFs) are depicted with the blue rectangles. Success rates of *KSENIA* are depicted with the yellow rectangles. Top$N$ value is defined as the percentage of protein complexes for which at least one of the docking prediction with the corresponding quality q is found within the first N docking poses. The quality of predictions q is evaluated according to the value of L_{RMSD} (see Table 2).

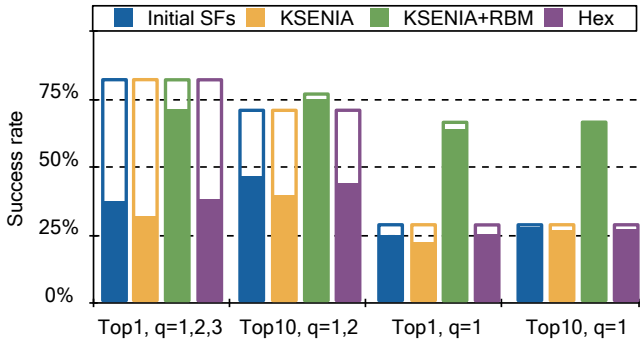


Figure 5: Performance of the scoring functions on the reduced Hex test benchmark. Success rates of the initial scoring functions (Initial SFs) are depicted with the solid blue rectangles. Success rates of *KSENIA* are depicted with the solid yellow rectangles. Success rates of *KSENIA* along with the rigid-body minimization (KSENIA+RBM) are depicted with the solid green rectangles. Success rates of the Hex scoring function are depicted with the solid purple rectangles. Hollow rectangles of the corresponding color represent the maximum achievable success rates. Top$N$ value is defined as the percentage of protein complexes for which at least one of the docking prediction with the corresponding quality q is found within the first N docking poses. The quality of predictions q is evaluated according to the value of L_{RMSD} (see Table 2).

We will refer to the obtained set as to the *reduced* Hex test benchmark. Figure 5 shows recomputed success rates for the top predictions (solid rectangles). In this figure, we also list the maximum success rates of the scoring functions (hollow rectangles) as the percentage of protein complexes for which Hex could predict poses of the corresponding quality. From Figure 5 one can see that the derived scoring functions provide a similar success rate as the Hex scoring function, which is solely based on the shape-complementarity term. However, the initial scoring functions slightly out-perform *KSENIA* on the reduced Hex test benchmark. Presumably this is because we lose some information when re-defining potentials at short distances (see Section 1 of Supporting Information). Nonetheless, *KSENIA* is dedicated to be used with the local rigid-body minimization for the refinement of the docking predictions. Thereby, at the next step we used the rigid-body minimization protocol (see Section 3.1 and Table 1) to optimize the docking poses. Then, we ranked the optimized docking predictions according to the values of *KSENIA* and re-evaluated the success rates (Figure 5, green solid rectangles). We found that the rigid-body minimization dramatically improves the scoring results. In particular, the rigid-body minimization increased the total number of quality-one poses, improving near-natives poses towards natives and rising the maximum success rate from 28% to 66%. Moreover, the corresponding success rates are more than twice better compared to both the success rates of Hex and the success rates of scoring without the refinement procedure. Such a significant improvement can be explained by the fact that the initial Hex predictions have moderate steric clashes and thus are not very suitable for the scoring with *KSENIA* without subsequent docking pose optimization. To summarize, we demonstrated that employing structural information of only native interfaces, it is possible to distinguish near-native conformations of protein complexes from the non-native decoys. We have also shown that it is possible to refine docking predictions using a smooth knowledge-based statistical potential with a rigid-body minimization

procedure, which improves the quality of the predictions and the overall performance of the scoring method. Below, we further investigate the capability of our approach on more complicated test benchmarks.

4.2.2 Zdock Test Benchmark

For the Zdock benchmark set (see Section 3.4.2) we applied the rigid-body minimization protocol with *KSENIA*, as in the previous section, ranked the poses and compared the success rates against Zdock v.3.0.1 scoring function, which includes the shape-complementarity term, the electrostatic term and the desolvation term. Figure 6 shows results obtained on this benchmark. Our approach shows around three times better success rate for the top one quality-one, -two or -three predictions. We observed that similarly to the Hex test benchmark, the rigid-body minimization improves the results of scoring significantly, indicating that the refinement procedure is very crucial for *KSENIA*. We should also note, however, that for eight complexes in the benchmark, the rigid-body minimization deteriorated several quality-one predictions to quality -two or -three. Thus, the maximum number for the top one quality-one predictions is reduced from 97% to 91%. Nonetheless, our method demonstrates around seven times better success rate for the top one predictions with the highest quality compared to the Zdock v.3.0.1 scoring function.

4.2.3 Comparison with the ItScore-PP potential

The results of the previous sections show that *KSENIA* outperforms simple empirical scoring functions that include shape, electrostatic, and desolvation terms. This indicates that native protein complexes contain themselves all necessary information to derive robust knowledge-based scoring function. In order to support this claim further, we compared the performance of the *KSENIA* potential with the performance of the ItScore-PP potential.[16] The ItScore-PP potential is a distance-dependent knowledge-based scoring function obtained using an iterative technique that avoids the explicit estimation of the reference state probabilities. ItScore-PP was derived using the same atom typization and the same set of native complexes for its training as our scoring function. Furthermore, the authors also used ItScore-PP in combination with the refinement procedure when optimizing the docking candidates. Thus, it is very interesting to compare the scoring power of ItScore-PP and *KSENIA*.

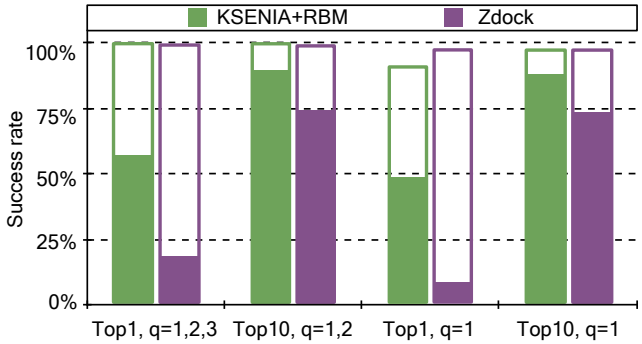


Figure 6: Performance of the scoring functions on the Zdock test benchmark. Success rates of *KSENIA* along with the rigid-body minimization (KSENIA+RBM) are depicted with the solid green rectangles. Success rates of the Zdock scoring function are depicted with the solid purple rectangles. Hollow rectangles of the corresponding color represent the maximum achievable success rates. Top$N$ value is defined as the percentage of protein complexes for which at least one of the docking prediction with the corresponding quality q is found within the first N docking poses. The quality of predictions q is evaluated according to the CAPRI criterion (see Table 3).

Similarly to the previous tests and following the scoring procedure used by ItScore-PP, we ran *KSENIA* in combination with the rigid-body refinement, ranked the predictions with respect to the score and evaluated the success rates according to the CAPRI criterion. The obtained results are presented in Figure 7. The success rates of the Zdock v.2.1 and Zdock v.2.3 scoring functions are adapted from Huang et al.[16] and plotted for the reference comparison. The former scoring function evaluates only the shape complementarity term, the latter also accounts for electrostatic and desolvation effects. As one can see, our scoring function slightly outperforms the ItScore-PP potential, provid-

ing 100% success rate for the top 10 predictions of acceptable quality. It is very important to note that, except for nine complexes, there is neither repetition nor highly homologous complexes in the test set compared to the training set.[16] Hence, there is no bias due to the overlap between the training and the test sets and the true success rates are represented.

We should note that here we did not verify the performance of *KSENIA* on the protein-protein unbound benchmark generated with the Zdock software. More precisely, after the rigid-body docking applied to the monomers in the unbound conformations, side-chains of the interface residues are, generally, in non-optimal conformations, which might be crucial for *KSENIA*. Instead, we verified the performance of *KSENIA* on the Rosetta bound and unbound test benchmarks and also using the SwarmDock test set, where side-chain conformations are optimized.

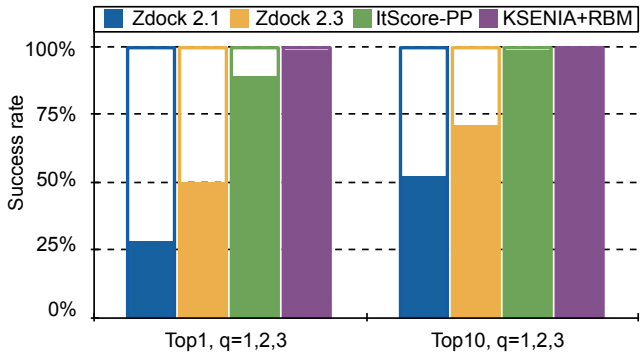


Figure 7: Performance of the scoring functions on the ItScore test benchmark. Success rates of *KSENIA* along with the rigid-body minimization (KSENIA+RBM) are depicted with the solid purple rectangle. Success rates of the ItScore-PP scoring function are depicted with the solid green rectangle. Success rates of the Zdock v.2.1 and v.2.3 scoring functions are depicted with the solid blue and yellow rectangles, respectively. Results for the ItScore-PP potential as well as for the Zdock v.2.1 and v.2.3 scoring functions are adapted from Huang et al.[16] Top$N$ value is defined as the percentage of protein complexes for which at least one of the docking prediction with at least the acceptable quality is found within the first N docking poses. The quality of predictions is evaluated according to the CAPRI criterion (see Table 3).

4.2.4 Rosetta Test Benchmark

Comparison of the performance of the Rosetta's scoring function against our rigid-body minimization with *KSENIA* is presented in Figure 8 for both the bound and the unbound benchmarks. As one can see, although Rosetta

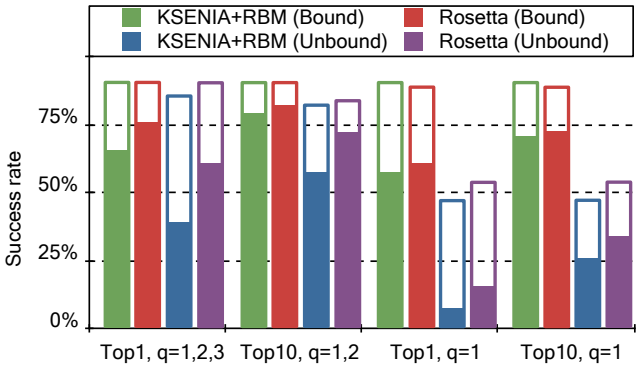


Figure 8: Performance of the scoring functions on the Rosetta bound and unbound test benchmarks. Success rates of *KSENIA* along with the rigid-body minimization (KSENIA+RBM) are depicted with the solid green and the solid blue rectangles for the Rosetta bound and unbound test benchmarks, respectively. Success rates of the Rosetta scoring function are depicted with the solid red and the solid purple rectangles for the Rosetta bound and unbound test benchmarks, respectively. Hollow rectangles of the corresponding color represent the maximum achievable success rates. Top$N$ value is defined as the percentage of protein complexes for which at least one of the docking prediction with the corresponding quality q is found within the first N docking poses. The quality of predictions q is evaluated according to the CAPRI criterion (see Table 3).

itself performs slightly better, our approach still demonstrates very good results despite the complexity of these benchmarks. Indeed, the native contacts for all the complexes in the benchmark are disturbed owing to the side-chain re-packing or homologous replacement, for example. In addition, our scoring method does not take into consideration the individual scores of the monomers. In particular, it does not penalize rare rotameric states of the side-chains, which are present in the benchmark. Nonetheless, using only distance distributions between the atoms in different monomers at their native and near-native

states, our knowledge-based potential is capable to rank quality-one poses at the top position for around 60 % of cases for the Rosetta bound benchmark, and to rank quality-one, -two or -three poses at the top position for around 45 % of cases for the Rosetta unbound benchmark.

4.2.5 SwarmDock Test Benchmark

Moal et al. compared the performance of 115 various scoring functions[1] on a decoy set was generated using the SwarmDock docking server.[52,56] The corresponding success rates of the best 40 scoring functions provided by Moal et al.[1] along with the success rates of *KSENIA* are presented in Figure 9.

As one can see, the *KSENIA* potential performs relatively well compared to the rest of the assessed potentials, being among best 40 potentials out of 115. The detailed performance of each scoring function on the entire protein-protein docking benchmark v.4.0 is listed in Table S1 from Supporting Information. While the problem of rigid-body pair-wise docking is considered to be solved,[57] the flexible docking problem still remains to be a challenge. Thus, it is interesting to see the success rates of scoring functions on the medium and difficult cases of protein-protein docking benchmark v.4.0. Here, we used the full benchmark, otherwise there are too few complexes for the test. We evaluated the corresponding performance of *KSENIA* and compared it with the success rates of other scoring functions provided by Moal et al.[1] Remarkably, *KSENIA* performs very competitively, resulted in 6 correctly predicted complexes, and only six scoring function out of 115 performed better (see Figure 10).

4.3 Crystallographic Symmetry Mates as Docking Predictions

We observed that in several cases non-native decoys replace near-native predictions at the top positions after the rigid-body minimization applied. As the result, the success rate becomes less than it could be, since the near-native predictions get a lower rank. For example, Table 4 lists scores before and after the rigid-body minimization applied to the protein complex 1ZC6 from the Hex test benchmark.

In terms of the ligand-RMSD, the decoy structure significantly differs from the native one: $L_{\mathrm{RMSD}} > 60$ Å. However, we found that the interface formed by the decoy monomers is similar to the one of the crystal-packing interfaces that are observed in the crystal structure. Typically, only one of the interfaces presented in the crystal is considered to be the native interface, and other crystal-packing interfaces or crystal contacts are considered to be the artifacts of crystallization (Fig. 11). However, distinguishing between the native interface and the crystal contacts is a challenging problem, since both are formed following the same physical principles.[58,59] For the case of homodimer 1ZC6, L_{RMSD} between the decoy and the complex forming the crystal contact is about 2.8 Å. We found these observations to be the additional evidence of the prediction capability of *KSENIA*.

4.3.1 Performance in CAPRI

To conclude the Results section we briefly overview performance of our team in CAPRI[60] rounds 26, 27 and 30, where the *KSENIA* scoring function was used. First, we used the Hex software[45] in order to generate preliminary docking poses. Then, we refined the poses using the rigid-body minimization algorithm in combination with the *KSENIA* potential. Additionally, the SCWRL4 package[61] was used at each iteration of the rigid-body minimization in order to optimize side-chain conformations. Finally, the best ten candidates were selected as the submission models for CAPRI. Figure 12 presents correct predictions for protein-protein CAPRI targets of Rounds 26-27 obtained with the described docking pipeline.

For the Targets 53-54 there were no unbound structure of one of the monomers and thus the homology modeling with the I-TASSER server[62] was used in order to generate initial docking models. For Target 53 our docking pipeline succeeded to provide one acceptable-quality prediction among ten top-ranked mod-

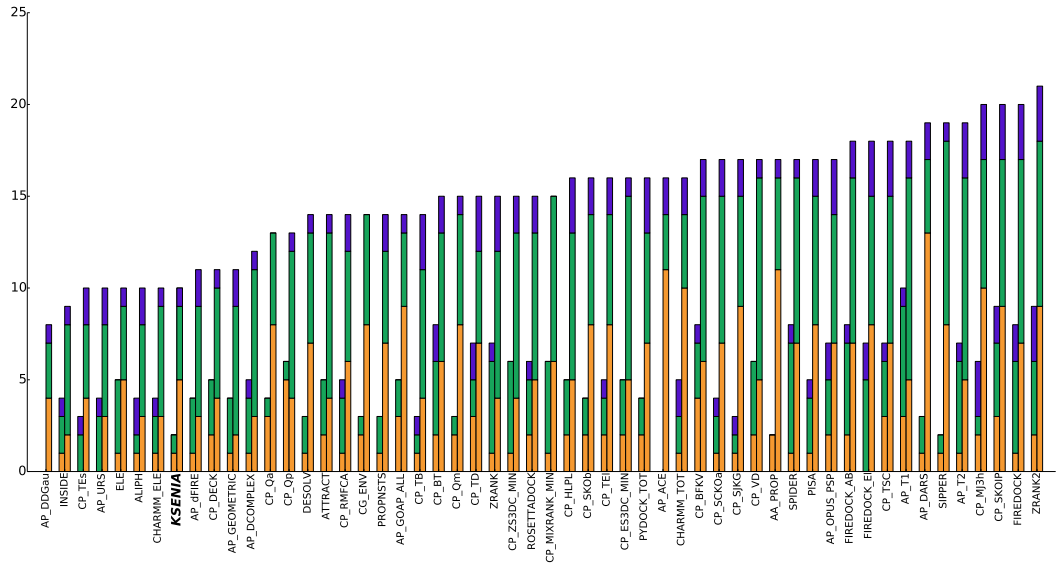


Figure 9: The success rates of the best 40 scoring functions adapted from Moal et al.[1] along with the success rates of the *KSENIA* potential on the protein-protein docking benchmark "v.4.0 update".The number of complexes for which an acceptable or better solution could be found in the top 1 and top 10 solutions was calculated for each scoring function. Acceptable quality solutions are shown in orange, medium quality solutions in green, and high quality solutions in blue for the two measures (top 1 left, top 10 middle). The functions are ordered by the top 10 success rate. We kept the same labels for the scoring function as in Moal et al.[1]

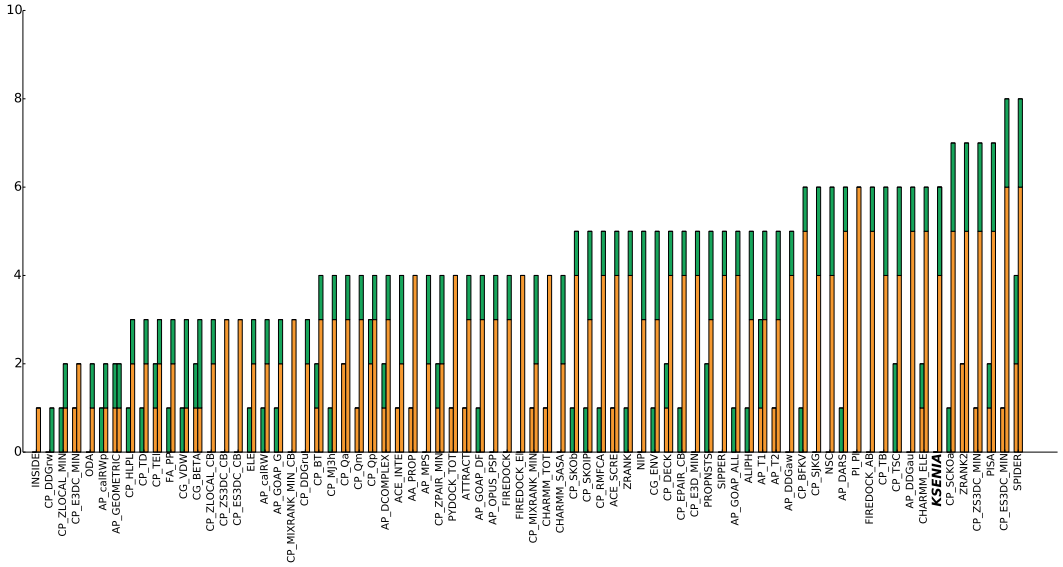


Figure 10: The success rates of the best 40 scoring functions adapted from Moal et al.[1] along with the success rates of the *KSENIA* potential on the medium and difficult cases of the protein-protein docking benchmark v.4.0. The number of complexes for which an acceptable or better solution could be found in the top 1 and top 10 solutions was calculated for each scoring function. Acceptable quality solutions are shown in orange and medium quality solutions in green for the two measures (top 1 left, top 10 middle). The functions are ordered by top 10 success rate. We kept the same labels for the scoring function as in Moal et al.[1]

Table 4: Scores for the native and one of the decoy structures before and after the rigid-body minimization.

1ZC6	Score	Score after the rigid-body minimization
U_{decoy}	-1594.740	-3036.307 (rank 1)
U_{native}	-1810.758 (rank 1)	-2144.868

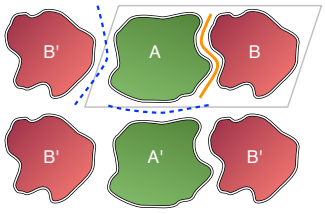


Figure 11: Schematic representation of the native interface (orange, solid) and crystal contacts (blue, dashed). The unit cell is depicted as the gray parallelogram encompassing monomers A and B, which form the native interface.

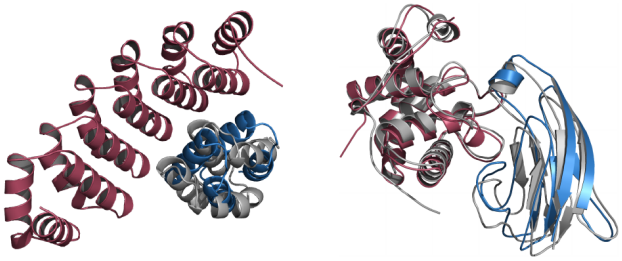


Figure 12: The native and predicted structures of the protein-protein complexes for CAPRI Targets. Left: native structure of Target 53 (grey) and acceptable-quality model produced by the docking pipeline (the two monomers are coloured in red and blue, respectively). Right: native structure of Target 58 (grey) and medium-quality model produced by the docking pipeline (the two monomers are coloured in red and blue, respectively).

els. However, there were no successful predictions for Target 54, probably due to the large difference between the true structure and the homologue model (only 4 teams out of 42 succeeded to produce acceptable-quality predictions). For Target 58 we obtained one medium-quality prediction and only four other teams out of 22 succeeded to produce predictions of the same quality. CAPRI Targets 55 and 56 were aimed to test methods for evaluating the effect of point mutations on protein-protein interaction affinity. Predictors were provided with the comprehensive datasets on the effects of every point mutant of two designed protein binders of influenza hemagglutinin[63] . Generally, point mutations stabilized the protein folds and some of them also provided effect on the binding of the complex. It turned out to be very difficult to predict the effect of the mutations following physics-based principles. As a result, only machine-learning methods provided statistically significant correlation between the predicted values and the measured K_d constants. Particularly, we obtained a good correlation between our score and the binding affinity for point mutations corresponding to four residues lying on the interface between the two proteins and failed otherwise[63] . CAPRI Round 30 was launched in collaboration with the Critical Assessment of Structure Predictions of proteins (CASP).[64] Overall, 42 interfaces in 25 targets were designated to CAPRI, comprising 19 protein dimers and 6 tetramers. In this round, we obtained correct predictions for 11 targets, including 10 predictions of medium quality.

4.4 Discussion

Reference state-based statistical methods require a large set of *false-positive* examples of protein complexes, i.e. non-native conformations, in order to compute the reference state.

Linear and quadratic programming approaches when training scoring functions also use a set of generated *false-positive* examples in order to construct the system of inequalities (1). It is a common practice in protein-protein docking to select as *false-positive* examples those decoys that possess the best score according to some well-accepted scoring function.[16,17,21] On the contrary, we have selected *false-negative* examples purely based on the structure of protein complexes in their native conformations. More precisely, our decoy sets were generated in such a way that the average RMSD between the corresponding monomers in the decoys and in the native structures is about 1 Å, keeping the relative orientation of the monomers fixed. Nonetheless, despite our training set does not contain non-native conformations with large RMSDs with respect to the native structures, we are able to reconstruct the atom-atom distance-dependent scoring functions (see Eq. (5)). As we have shown above, the obtained potentials demonstrate surprisingly good results on different protein-protein docking benchmarks. We would like to emphasize that all the benchmarks mostly consist of non-native decoys that have large RMSDs with respect to the native structures. Thus, our results strongly suggest that the native protein complexes themselves contain all necessary structural information to build well-discriminative potentials that recognize native and near-native protein-protein conformations.

Regarding the disadvantages of the proposed methodology, i.e. derivation of the *KSENIA* potential, we can point out two aspects. First, current statistic observations do not take into account conformations of individual monomers. This means that, in principle, we can imagine a situation when two very unrealistic structures of two monomers (all atomic coordinates inside each monomer are the same, for example) result in a good score of the complex. To circumvent this problem, one may either collect extra geometric information, such as triplet, quadruplet, etc. distributions of atoms in the complex, or additionally score individual monomers. Second, in our training set there are no statistics at short separation distances between the monomers inside a complex. Thus, as a result, we need to define potential barriers at short distances for the proper behaviour of the obtained scoring functions.

We would also like to stress that even though the *KSENIA* potential is derived using local perturbations of the native protein structures, it generally has no bias toward a method to generate docking predictions. This is because for the construction of the training set we did not use any standard docking prediction method. Thus, the rigid-body minimization is very important for the success of the proposed scoring methodology. Namely, the minimization is required to resolve steric clashes that often appear in docking predictions produced by various methods, particularly those that use a grid search without subsequent refinement of the docking predictions. For example, Zdock and Hex use a soft shape complementarity potential, which permits moderate overlaps between the monomers in a complex. Generally, we believe that structure optimization should be the inevitable step of a general scoring procedure when one has no information about docking predictions to score.

Our method does not, in principle, require external packages, potentials, or algorithms neither to generate the training set, nor to formulate and solve the optimization problem. In the present study, to generate the local deformations, we computed low-frequency normal modes using the MMTK package with a united-atom force-field.[39] However, the normal modes can be computed in a simpler way using, e.g. the elastic-network model,[65] the gaussian network model,[66] the rotation-translation of blocks method,[67] etc. Thus, methodology presented in this paper can be easily adapted to the recognition of other types of molecular interactions, such as protein-ligand, protein-RNA, etc., provided that the atom types assignment is modified appropriately.

5 Conclusions

Present study demonstrates that knowledge of only native protein-protein interfaces is suffi-

cient to construct well-discriminative predictive models for the selection of binding candidates. Namely, we introduced a new scoring method that comprises a knowledge-based potential called *KSENIA* deduced from the structural information about the native interfaces of 844 crystallographic protein-protein complexes. The knowledge-based potential relies on the information obtained thanks to the deformations of these interfaces computed along the low-frequency normal modes. As a result, in contrast to existing scoring functions, our potential does not require neither the computation of the reference state nor the ensemble of non-native complexes. Thus, it can have only a marginal bias toward a method to generate putative binding poses. Moreover, *KSENIA* is smooth by construction, which allows to use it along with the gradient-based rigid-body minimization. Particularly, we showed that the rigid-body optimization of the docking poses improves the scoring stage of molecular docking. Using several test benchmarks we demonstrated that our method out-performs the Hex scoring function, which is based on the shape complementarity between the monomers in a complex, and the Zdock scoring function, which also includes the electrostatic and desolvation terms. We have also demonstrated that the *KSENIA* potential slightly outperforms the ItScore-PP potential, which is the atomic distance-dependent scoring function derived using the same set of native complexes. We find remarkable that the native protein complexes themselves contain all necessary information to derive a successful and well-discriminative knowledge-based potential. Although our method performs slightly worse on the Rosetta test benchmark and SwarmDock test benchmark compared to the more sophisticated scoring functions, we believe that further improvements of *KSENIA*, e.g. accounting for the integrity of monomers, rotamer optimization, etc., will eliminate this disadvantage.

Methodology presented in this paper can be easily adapted to the recognition of other types of molecular interactions, such as protein-ligand, protein-RNA, etc. Finally, we want to note that we have successfully used *KSENIA* in the CAPRI protein docking experiment starting from Round 26.[60] We will make *KSENIA* publicly available as a part of the SAMSON software platform developed in our group at `https://team.inria.fr/nano-d/software`.

Acknowledgements The authors acknowledge Mieczyslaw Torchala from Biomolecular modelling group of the Francis Crick Institute for the provided SwarmDock test benchmark.

Funding Sources This work was supported by the Agence Nationale de la Recherche (ANR-2010-JCJC-0206-01, ANR-11-MONU-006-01).

Associated Content Supporting Information Available: Description of the artificial barriers, the cross-validation procedure, and the performance of scoring functions including the *KSENIA* one on the SwarmDock benchmark. This material is available free of charge via the Internet at `http://pubs.acs.org`.

References

(1) Moal, I. H.; Torchala, M.; Bates, P. A.; Fernández-Recio, J. The Scoring of Poses in Protein-Protein Docking: Current Capabilities and Future Directions. *BMC Bioinf.* **2013**, *14*, 286.

(2) Finkelstein, A.; Badretdinov, A.; Gutin, A. Why Do Protein Architectures Have Boltzmann-Like Statistics? *Proteins: Struct., Funct., Bioinf.* **2004**, *23*, 142–150.

(3) Tanaka, S.; Scheraga, H. A. Medium- and Long-Range Interaction Parameters between Amino Acids for Predicting Three-Dimensional Structures of Proteins. *Macromolecules* **1976**, *9*, 945–950.

(4) Miyazawa, S.; Jernigan, R. Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation. *Macromolecules* **1985**, *18*, 534–552.

(5) Sippl, M. Calculation of Conformational Ensembles from Potentials of Mean Force: an Approach to the Knowledge-Based Prediction of Local Structures in Globular Proteins. *J. Mol. Biol.* **1990**, *213*, 859–883.

(6) Sippl, M.; Ortner, M.; Jaritz, M.; Lackner, P.; Flöckner, H. Helmholtz Free Energies of Atom Pair Interactions in Proteins. *Folding Des.* **1996**, *1*, 289–298.

(7) Koppensteiner, W.; Sippl, M. Knowledge-Based Potentials–Back to the Roots. *Biochemistry* **1998**, *63*, 247–252.

(8) Thomas, P.; Dill, K. Statistical Potentials Extracted from Protein Structures: How Accurate Are They? *J. Mol. Biol.* **1996**, *257*, 457–469.

(9) Ben-Naim, A. Statistical Potentials Extracted from Protein Structures: Are These Meaningful Potentials? *J. Chem. Phys.* **1997**, *107*, 3698–3706.

(10) Liu, S.; Zhang, C.; Zhou, H.; Zhou, Y. A Physical Reference State Unifies the Structure-Derived Potential of Mean Force for Protein Folding and Binding. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 93–101.

(11) Deng, H.; Jia, Y.; Wei, Y.; Zhang, Y. What Is the Best Reference State for Designing Statistical Atomic Potentials in Protein Structure Prediction? *Proteins: Struct., Funct., Bioinf.* **2012**, *80*, 2311–2322.

(12) Leelananda, S. P.; Feng, Y.; Gniewek, P.; Kloczkowski, A.; Jernigan, R. L. In *Multiscale Approaches to Protein Modeling*; Kolinski, A., Ed.; Springer New York, 2011; pp 127–157.

(13) Lu, H.; Skolnick, J. A Distance-Dependent Atomic Knowledge-Based Potential for Improved Protein Structure Selection. *Proteins: Struct., Funct., Bioinf.* **2001**, *44*, 223–232.

(14) Samudrala, R.; Moult, J. An All-Atom Distance-Dependent Conditional Probability Discriminatory Function for Protein Structure Prediction. *J. Mol. Biol.* **1998**, *275*, 895–916.

(15) Zhou, H.; Zhou, Y. Distance-Scaled, Finite Ideal-Gas Reference State Improves Structure-Derived Potentials of Mean Force for Structure Selection and Stability Prediction. *Protein Sci.* **2002**, *11*, 2714–2726.

(16) Huang, S.-Y.; Zou, X. An Iterative Knowledge-Based Scoring Function for Protein–Protein Recognition. *Proteins: Struct., Funct., Bioinf.* **2008**, *72*, 557–579.

(17) Chuang, G.-Y.; Kozakov, D.; Brenke, R.; Comeau, S. R.; Vajda, S. Dars (Decoys As The Reference State) Potentials for Protein-Protein Docking. *Biophys. J.* **2008**, *95*, 4217–4227.

(18) Maiorov, V. N.; Grippen, G. M. Contact Potential that Recognizes the Correct Folding of Globular Proteins. *J. Mol. Biol.* **1992**, *227*, 876–888.

(19) Qiu, J.; Elber, R. Atomically Detailed Potentials to Recognize Native and Approximate Protein Structures. *Proteins: Struct., Funct., Bioinf.* **2005**, *61*, 44–55.

(20) Rajgaria, R.; McAllister, S.; Floudas, C. A Novel High Resolution Cα–Cα Distance Dependent Force Field Based on a High Quality Decoy Set. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 726–741.

(21) Tobi, D.; Bahar, I. Optimal Design of Protein Docking Potentials: Efficiency and Limitations. *Proteins: Struct., Funct., Bioinf.* **2006**, *62*, 970–981.

(22) Ravikant, D.; Elber, R. PIE–Efficient Filters and Coarse Grained Potentials for Unbound Protein–Protein Docking. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 400–419.

(23) Chae, M.-H.; Krull, F.; Lorenzen, S.; Knapp, E.-W. Predicting Protein Complex Geometries with a Neural Network. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1026–1039.

(24) Mirzaei, H.; Beglov, D.; Paschalidis, I. C.; Vajda, S.; Vakili, P.; Kozakov, D. Rigid Body Energy Minimization on Manifolds for Molecular Docking. *J. Chem. Theory Comput.* **2012**, *8*, 4374–4380.

(25) Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. International Joint Conference on Artificial Intelligence. 1995; pp 1137–1145.

(26) Derevyanko, G.; Grudinin, S. Convex-PP: Predicting Protein–Protein Interactions Using Polynomial Expansions and Convex Optimisation. *Unpublished.*

(27) Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press, New York, 2004.

(28) Huang, S.-Y.; Yan, C.; Grinter, S. Z.; Chang, S.; Jiang, L.; Zou, X. Inclusion of the Orientational Entropic Effect and Low-Resolution Experimental Information for Protein–Protein Docking in Critical Assessment of PRedicted Interactions (CAPRI). *Proteins: Struct., Funct., Bioinf.* **2013**, *81*, 2183–2191.

(29) Shen, M.-y.; Sali, A. Statistical Potential for Assessment and Prediction of Protein Structures. *Protein Sci.* **2006**, *15*, 2507–2524.

(30) Zhang, C.; Liu, S.; Zhu, Q.; Zhou, Y. A Knowledge-Based Energy Function for Protein-Ligand, Protein-Protein, and Protein-DNA Complexes. *J. Med. Chem.* **2005**, *48*, 2325–2335.

(31) Nelder, J. A.; Mead, R. A Simplex Method for Function Minimization. *Computer Journal* **1965**, *7*, 308–313.

(32) Powell, M. J. An Efficient Method for Finding the Minimum of a Function of Several Variables Without Calculating Derivatives. *Computer Journal* **1964**, *7*, 155–162.

(33) Wilson, E. B. *Molecular Vibrations: the Theory of Infrared and Raman Vibrational Spectra*; Dover Publications, New York, 1955.

(34) Brooks, B.; Karplus, M. Harmonic Dynamics of Proteins: Normal Modes and Fluctuations in Bovine Pancreatic Trypsin Inhibitor. *Proc. Natl. Acad. Sci. U. S. A.* **1983**, *80*, 6571–6575.

(35) May, A.; Zacharias, M. Energy Minimization in Low-Frequency Normal Modes to Efficiently Allow for Global Flexibility During Systematic Protein-Protein Docking. *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 794–809.

(36) Venkatraman, V.; Ritchie, D. W. Flexible Protein Docking Refinement Using Pose-Dependent Normal Mode Analysis. *Pro-*

teins: Struct., Funct., Bioinf. **2012**, *80*, 2262–2274.

(37) Cavasotto, C. N.; Kovacs, J. A.; Abagyan, R. A. Representing Receptor Flexibility in Ligand Docking Through Relevant Normal Modes. *J. Am. Chem. Soc.* **2005**, *127*, 9632–9640.

(38) Berman, H. M.; Battistuz, T.; Bhat, T. N.; Bluhm, W. F.; Bourne, P. E.; Burkhardt, K.; Feng, Z.; Gilliland, G. L.; Iype, L.; Jain, S. The Protein Data Bank. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 899–907.

(39) Hinsen, K. The Molecular Modeling Toolkit: a New Approach to Molecular Simulations. *J. Comput. Chem.* **2000**, *21*, 79–85.

(40) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(41) Hinsen, K. Analysis of Domain Motions by Approximate Normal Mode Calculations. *Proteins: Struct., Funct., Genet.* **1998**, *33*, 417–429.

(42) Hwang, H.; Vreven, T.; Janin, J.; Weng, Z. Protein–Protein Docking Benchmark Version 4.0. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 3111–3114.

(43) Pearson, W. R.; Lipman, D. J. Improved Tools for Biological Sequence Comparison. *Proc. Natl. Acad. Sci. U. S. A.* **1988**, *85*, 2444–2448.

(44) Ritchie, D. W.; Kozakov, D.; Vajda, S. Accelerating and Focusing Protein–Protein Docking Correlations Using Multi-Dimensional Rotational FFT Generating Functions. *Bioinformatics* **2008**, *24*, 1865–1873.

(45) Ritchie, D. W.; Venkatraman, V. Ultra-Fast FFT Protein Docking on Graphics Processors. *Bioinformatics* **2010**, *26*, 2398–2405.

(46) Popov, P.; Grudinin, S. Rapid Determination of RMSDs Corresponding to Macromolecular Rigid Body Motions. *J. Comput. Chem.* **2014**, *35*, 950–956.

(47) Hwang, H.; Pierce, B.; Mintseris, J.; Janin, J.; Weng, Z. Protein–Protein Docking Benchmark Version 3.0. *Proteins: Struct., Funct., Bioinf.* **2008**, *73*, 705–709.

(48) Pierce, B. G.; Hourai, Y.; Weng, Z. Accelerating Protein Docking In ZDOCK Using an Advanced 3D Convolution Library. *PloS One* **2011**, *6*, e24657.

(49) Méndez, R.; Leplae, R.; De Maria, L.; Wodak, S. J. Assessment of Blind Predictions of Protein–Protein Interactions: Current Status of Docking Methods. *Proteins: Struct., Funct., Bioinf.* **2003**, *52*, 51–67.

(50) Chen, R.; Weng, Z. Docking Unbound Proteins Using Shape Complementarity, Desolvation, and Electrostatics. *Proteins: Struct., Funct., Bioinf.* **2002**, *47*, 281–294.

(51) Gray, J.; Moughon, S.; Wang, C.; Schueler-Furman, O.; Kuhlman, B.; Rohl, C.; Baker, D. Protein–Protein Docking with Simultaneous Optimization of Rigid-Body Displacement and Side-Chain Conformations. *J. Mol. Biol.* **2003**, *331*, 281–300.

(52) Torchala, M.; Bates, P. A. *Protein Structure Prediction*; Methods Mol. Biol. (N. Y., NY, U. S.); Springer, 2014; Vol. 1137; pp 181–197.

(53) Moal, I. H.; Moretti, R.; Baker, D.; Fernández-Recio, J. Scoring Functions for Protein-Protein Interactions. *Curr. Opin. Struct. Biol.* **2013**, *23*, 862–867.

(54) Press, W. H. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*; Cambridge University Press, New York, 2007.

(55) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. The Energy Landscapes and Motions of Proteins. *Science* **1991**, *254*, 1598–1603.

(56) Torchala, M.; Moal, I. H.; Chaleil, R. A.; Fernandez-Recio, J.; Bates, P. A. SwarmDock: a Server for Flexible Protein–Protein Docking. *Bioinformatics* **2013**, *29*, 807–809.

(57) Vakser, I. A. Protein-Protein Docking: From Interaction to Interactome. *Biophys. J.* **2014**, *107*, 1785–1793.

(58) Kobe, B.; Guncar, G.; Buchholz, R.; Huber, T.; Maco, B.; Cowieson, N.; Martin, J. L.; Marfori, M.; Forwood, J. K. Crystallography and Protein-Protein Interactions: Biological Interfaces and Crystal Contacts. *Biochem. Soc. Trans.* **2008**, *36*, 1438–1441.

(59) Krissinel, E. Crystal Contacts as Nature's Docking Solutions. *J. Comput. Chem.* **2010**, *31*, 133–143.

(60) Janin, J.; Wodak, S. J.; Lensink, M. F.; Velankar, S. Assessing Structural Predictions of Protein–Protein Recognition: The CAPRI Experiment. *Rev. Comput. Chem.* **2015**, *28*, 137–173.

(61) Krivov, G. G.; Shapovalov, M. V.; Dunbrack, R. L. Improved Prediction of Protein Side-Chain Conformations with SCWRL4. *Proteins: Struct., Funct., Bioinf.* **2009**, *77*, 778–795.

(62) Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y. The I-TASSER Suite: Protein Structure and Function Prediction. *Nat. Methods* **2015**, *12*, 7–8.

(63) Moretti, R.; others., Community-Wide Evaluation of Methods for Predicting the Effect of Mutations on Protein–Protein Interactions. *Proteins: Struct., Funct., Bioinf.* **2013**, *81*, 1980–1987.

(64) Moult, J.; Fidelis, K.; Kryshtafovych, A.; Schwede, T.; Tramontano, A. Critical Assessment of Methods of Protein Structure Prediction (CASP)–Round X. *Proteins: Struct., Funct., Bioinf.* **2014**, *82*, 1–6.

(65) Tirion, M. M. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys. Rev. Lett.* **1996**, *77*, 1905.

(66) Bahar, I.; Atilgan, A. R.; Erman, B. Direct Evaluation of Thermal Fluctuations in Proteins Using a Single-Parameter Harmonic Potential. *Folding Des.* **1997**, *2*, 173–81.

(67) Tama, F.; Gadea, F. X.; Marques, O.; Sanejouand, Y. H. Building-Block Approach for Determining Low-Frequency Normal Modes of Macromolecules. *Proteins: Struct., Funct., Bioinf.* **2000**, *41*, 1–7.