Receptor-free Discrimination of Peptide Native Bound Conformations Suggests Limited Transferability of the “cen_std+score4l” Rosetta Energy Function From Proteins to Peptides

Andrea Bazzoli

Università degli Studi di Milano: Universita degli Studi di Milano https://orcid.org/0000-0003-0980-7110

Alessandro Contini (alessandro.contini@unimi.it)

Università degli Studi di Milano, Dipartimento di Scienze Farmaceutiche, Sezione di Chimica Generale e Organica “A. Marchesini”, Via Venezian, 21 20133 Milano, Italy https://orcid.org/0000-0002-4394-8956

Research article

Keywords: Peptide modeling, conformational selection, decoys, native discrimination, Rosetta, rigid-body docking

Posted Date: September 3rd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-69979/v1

Receptor-free discrimination of peptide native bound conformations suggests limited transferability of the "cen_std+score4L" Rosetta energy function from proteins to peptides

Andrea Bazzoli[1] *and Alessandro Contini*[1*]

[1]Università degli Studi di Milano, Dipartimento di Scienze Farmaceutiche, Sezione di Chimica Generale e Organica "A. Marchesini", Via Venezian, 21 20133 Milano, Italy

*To whom correspondence should be addressed. E-mail: **alessandro.contini@unimi.it**

ABSTRACT

One of the strategies of peptide–protein docking is to pregenerate an ensemble of peptide conformations in the absence of the receptor, and then dock them as rigid bodies onto its surface. Success of this strategy requires that the scoring function that drives the pregeneration step be able to discriminate in favor of conformations that resemble the native bound conformation. Here we

present a study on the discrimination of peptide native bound conformations as achieved without receptor by the "cen_std+score4L" Rosetta energy function, a low-resolution scoring function equivalent to one chosen for other tasks where the modeling of solvent effects is of special importance. The *cen_std+score4L* function was able to assign, on average, lower energies to native-like than to non-native decoy conformations for only 3 of our 18 test peptides; it also ranked one or more native-like decoys in the top 1% for only 2 peptides. However, by optimizing the weights of the energy terms that define the *cen_std+score4L* function, native discrimination improved substantially: Native-like decoys were assigned lower energies than non-native decoys for 16 peptides, with a discrimination signal larger than noise for 9 peptides, that is, 50% of the test set. And for 9 peptides, too, native-like decoys ranked in the top 1%. An ensuing energetic analysis of native-like versus non-native decoys suggests that native peptide conformations have solvation and non-local electrostatics that poorly recapitulate those of native protein conformations. Native peptide conformations are also characterized by few backbone–backbone H-bonds and by lack of compactness, presumably to optimize interaction with the receptor. Overall, this study lays groundwork for pregenerating dockable peptide conformations with Rosetta, whether the subsequent docking will be performed by Rosetta or some other software.

KEYWORDS

Peptide modeling, conformational selection, decoys, native discrimination, Rosetta, rigid-body docking

INTRODUCTION

Peptides are increasingly being used in pharmaceutics as inhibitors of protein–protein interactions (PPIs)[1]. These interactions pervade the cell, and not surprisingly a multitude of diseases are caused by deregulated PPIs[1, 2]. Because of their ample and shallow binding interfaces, PPIs have historically challenged inhibition by small molecules[3, 4]. Peptide inhibitors, on the other hand, by offering a larger contact area, could more easily achieve the necessary potency, an intuition supported by the estimate that 15–40% of PPIs are mediated by peptides[5]. Rational design of PPI inhibitors requires the structure of the PPI interface to be known in atomic detail. Presumably, however, this information is still unavailable for a large majority of the 130,000–650,000 human PPIs[6, 7], since the number of human biological assemblies in the Protein Data Bank (PDB)[8] that contain two or more protein chains is just above 25,000. Nevertheless, when the PDB does contain the structure of one or both partners of a PPI, efforts to model the binding of candidate peptide inhibitors are chiefly based on peptide–protein docking[9]. Methods of this sort usually take as input the amino acid sequence of the peptide candidate, the structure of the target protein, or "receptor", and any available information on the position of the binding site. Their goal is then to predict, often on energetic grounds, the most likely structure of the peptide–protein complex.

Peptide–protein docking is typically a two-step process in which a coarse-grained exploration of the possible binding modes is followed by high-resolution refinement of those having lower energies[10–19]. During the exploratory step, the peptide may be forced to adopt a limited number of pregenerated conformations, which are docked as rigid bodies at the binding site (local docking) or, when that is unknown, around the entire surface of the receptor (global docking). These methods assume that at least one of those conformations is sufficiently similar to the native bound conformation. In some cases, conformations are pregenerated by assemblying fragments extracted

from the PDB. The source of fragments are bound peptides[19] or even protein chains[15, 17], in accordance with the observation that peptide–protein interfaces are often structurally similar to interactions within monomeric proteins[20]. Other methods[11, 12] pregenerate only three idealized conformations (extended, α-helical, and polyproline-II) whose secondary structure was found to represent most bound peptides in the PDB[21, 22]. Their approach was inspired by the principle of "conformational selection"[23, 24] as applied to peptide binding, whereby receptors select bound peptide conformations that are at least partially stable in solution. As a more thorough implementation of this principle, the AnchorDock[13] method simulates the peptide alone in implicit solvent, and then selects one dockable conformation from one of the trajectory's low-energy clusters. While generally successful, these docking methods based on conformational selection have failed on 20–30% of test cases[11–13], a shortcoming that can be immediately ascribed to the incompleteness of their initial ensemble of conformations. Therefore, the ability to pregenerate peptide conformations that are metastable in solution—and hence likely to resemble, in part or in full, the native bound conformation—may substantially improve this class of docking methods.

A conformational generator with that ability requires a scoring function that can discriminate in favor of the native bound conformation. In general, the discriminatory power of a scoring function is assessed outside of any search method, on a cleverly constructed set of so-called "decoy" conformations, where native-like conformations are interspersed among many more, deliberately challenging, non-native conformations[25–27]. Once native-like decoys get scored consistently better than non-native decoys, the scoring function may be confidently integrated into a conformational search method. At that point, failure to produce native-like conformations will suggest a need for improvement not of the scoring function but of the sampling mechanism that is

coupled to it. However, scoring function and sampling mechanism are both dictated by model resolution. Peptide modeling at full atomic resolution achieved a fairly exhaustive exploration of metastable states when coupled to techniques of enhanced sampling[28–30]. The computational cost of full atomic resolution is nonetheless still prohibitive for what is intended to be a few-hour-long preliminary step to docking. For this reason, pregeneration of peptide conformations to cover sufficiently well the native bound conformation may only be feasible at low resolution. And that is the resolution at which to assess native discrimination by the underlying scoring function.

In the Rosetta macromolecular modeling software suite[31], a scoring, or "energy", function is a linear combination of terms that capture energetic contributions of different physical nature. This functional form is used under both of Rosetta's resolution modes, namely, the "fullatom" mode, where all atoms of the biomolecule are explicitly represented[32], and the "centroid" mode, where nonpolar hydrogens are neglected and side-chain atoms beyond C^{β} are represented by a single pseudo-atom (the centroid) located in an idealized center of mass[33]. With its low number of conformational degrees of freedom, centroid mode is to be chosen when coarse exploration of conformational space pays off better than exploitation of current solutions, such as in the early stages of *ab initio* structure prediction[34, 35]. A basic energy function in centroid mode is called *cen_std* and consists of four energy terms—*vdw*, *cbeta*, *env*, and *pair*—all of which are assigned a weight of 1.0 in the linear combination, as listed in Table 1. The table also presents five other energy terms—*hbond_lr_bb*, *hbond_sr_bb*, *rama*, *rg*, and *chainbreak*—as part of another energy function, called *score4L*. The *cen_std* and *score4L* functions were previously coupled in the low-resolution stages of two Rosetta protocols, namely, NGK[36], used for loop modeling, and FlexPepDock[10], used for peptide–protein docking. Loop modeling deals with residues that have, expectedly, higher solvent exposure than those in the protein structural databases from which most

of the *cen_std* and *score4L* terms were statistically derived[33, 37–40]. Likewise, even though FlexPepDock keeps the peptide close to the receptor, approximately 50% of the peptide surface is always facing the solvent. In light of the success of both NGK and FlexPepDock under those partially solvated conditions, it is worth taking it one step further and exploring whether the predictive power of centroid-mode Rosetta under *cen_std* and *score4L* still holds at the even higher solvent exposures of fully solvated peptides.

Here we report on the ability of the "cen_std+score4L" Rosetta energy function to discriminate native-like versus non-native bound peptide conformations in the absence of the cognate receptor. For brevity, we will refer to such peptide conformations as "native-like" and "non-native", respectively, taking the "bound" qualifier for granted. Discrimination by the default set of weights will be evaluated on a test set of 18 peptides over ensembles of ~7,000 decoy conformations per peptide. That performance will then be compared, over the same decoy ensembles, against the performance of an optimized weight set obtained by training on a separate set of 9 peptides. Weight optimization was driven by a native discrimination function that measured the ability of the Rosetta energy function to assign lower energies to native-like decoys than to non-native decoys. We actually considered only eight of the nine energy terms, discarding the *chainbreak* term, because the *chainbreak* energy was always zero for our decoys (that is, each decoy was a fully connected polymeric chain) and hence no optimized weight for this energy term would have reflected an actual improvement in native discrimination. The presence of optimized weights with a negative sign led us to investigate whether the corresponding energy terms tend to favor non-native over native-like conformations, contrary to the spirit in which such terms were originally conceived. The results of that analysis will be presented here, with special attention to the different occurrence of intramolecular hydrogen bonds in native-like versus non-native conformations.

METHODS

Benchmark peptide set

The benchmark set chosen for this study is the same set of 27 peptides that was used to benchmark the PIPER-FlexPepDock peptide–protein docking method[17]. The set, listed in Table 2, is nonredundant with respect to the receptor domain. Although no known receptor was directly used to optimize the set of weights here published, it is important that these weights should be robust with respect to the receptor, because their intended use is for generating native-like peptide conformations for effective docking onto any type of receptor.

Table 1. Rosetta energy terms whose weights were optimized in the present work

energy function	energy term	physical origin of the energy term	default weight[a]	optimized weight[b]	references
cen_std	env	solvation	1.0000	-1.0000	[[33], [38]]
	cbeta	solvation	1.0000	-0.3028	[[33], [38]]
	pair	electrostatic/disulfide interactions with sequence separation ≥ 9	1.0000	-3.5681	[[33], [38]]
	vdw	repulsive van der Waals interactions	1.0000	0.7553	[[33], [38]]
score4L	hbond_lr_bb	backbone–backbone H-bonds with sequence separation ≥ 5	1.0000	0.8295	[[32], [39], [41]]
	hbond_sr_bb	backbone–backbone H-bonds with sequence separation < 5	1.0000	-0.8092	[[32], [39], [41]]
	rama	backbone torsional angle preferences	0.1000	0.0728	[[33]]
	rg	attractive van der Waals interactions; solvation	2.0000	-0.3074	[[33], [38]]
	chainbreak[c]	polymeric chain connectivity	1.0000	-	[[42]]

[a]Default weight in the *cen_std* and *score4L* functions. [b]Optimized weight in the *cen_std+score4L* function. [c]Though contributing to the *score4L* function, the *chainbreak* term was excluded from our weight optimization.

Nonetheless, among these 27 nonredundant peptides, the energy weights were optimized on a slightly different training set than that chosen for the PIPER-FlexPepDock study. In particular, since the original training set contained only one mostly-helical peptide (2a3i), we decided to substitute a second mostly-helical peptide (2fmf) for one of the seven mostly-coiled peptides (1jwg). Given the 9 training peptides, the remaining 18 peptides constituted the test set (Table 2).

Generation of decoy conformations

For each peptide of the benchmark we generated one ensemble of decoy conformations for subsequent scoring by both the default weights and the optimized weights. Each ensemble consisted of 7,054 decoy conformations. These were divided into 54 decoys that served as starting conformations of stochastic local-search simulations, plus 7,000 decoys produced in those simulations. One starting decoy was the native conformation. Three others were schematic helical, extended, and hairpin conformations, respectively, that were meant to represent non-native conformations with regular secondary structure (Figures S1A–D). From each of these four starting decoys, 20 simulations were launched, each producing 50 decoys, for a total of 1,000 decoys per starting decoy. The remaining 50 starting decoys were aimed at sampling less regular instances of the non-native state (Figure S1E). They were generated by fragment assembly using Rosetta's FlexPepDock[10], where the receptor, required by the program, was set to be a fictitious, 12-residue receptor, and was placed hundreds of angstroms away from the initial peptide position in order to limit its effect on the assembly. Fragments were picked by FragmentPicker[43] based on the secondary structure predicted by the PSIPRED server[44]. Those 50 fragment-assembled decoys were the starting points of additional simulations that produced 60 decoys each, for a total of 3,000 decoys. All simulations were done in centroid mode[33] under combined *cen_std* and *score4L* energy terms with default weights; the *rama* weight was set to be 4-fold higher than its default value to more strongly favor naturally occurring (φ, ψ) pairs. Simulations were of the Monte Carlo[45] type: At each step a backbone dihedral angle was selected at random and then randomly perturbed, for at most 5,000 attempts per step, until the energy of the new conformation met a Boltzmann acceptance criterion, upon which the new conformation was added to the decoy set. The Boltzmann temperature was such that an energy increase of 1 (arbitrary units) was accepted with a probability of 0.02. Dihedral perturbations per attempt were of at most 20° for φ

and ψ angles; ω angles could only flip by 180° and only for proline residues, with a 0.9 probability for the *cis*-to-*trans* flip and a 0.1 probability for the *trans*-to-*cis* flip. The accumulation of dihedral perturbations over successive steps produced decoys that tended to have a progressively higher root mean square displacement (RMSD) from the starting conformation, as concluded from the visual inspection of ~20 RMSD profiles over time. The resulting decoys were quite heterogeneous in the way they differed from the native conformation (Table S1). We therefore judged that the decoy sets were sufficiently representative samples of conformational space on which to train or test the native discrimination power of the energy function. To this end, decoys were classified into "native-like" and "non-native", according to whether their C^{α}-RMSD from the native conformation was ≤ 1 Å or > 1 Å, respectively.

Table 2. Benchmark results

	peptide[b]	sequence / secondary structure[c]	top-1 RMSD (Å)[d]		top-1% RMSD (Å)[e]		native rank[f]	
			def[g]	opt[h]	def	opt	def	opt
training set[a]	1czy	PQQATDD CEECCCC	3.1	2.4	3.0	2.1	5097	1031
	1er8	HPFHLLVY CCCBCCBC	3.6	1.1	3.6	**0.8**	4815	57
	1jd5	AIAYFIPD CEEEETCC	3.7	**0.7**	3.6	**0.3**	5099	1
	2fmf	QDQVDDLLDSLGF HHHHHHHHHHHCC	2.3	6.8	**0.5**	6.4	32	1505
	1mfg	EYLGLDVPV CCCCCCEEC	4.1	5.1	3.9	**0.5**	4996	9
	2a3i	QQKSLLQQLLTE CCCCHHHHHHHC	4.4	7.7	3.5	7.1	3903	1023
	2cch	HTLKGRRLVFDN TTTTCCCCCCCC	5.4	5.5	5.3	**0.1**	5171	39
	2ds8	ALRVVK CCEECC	2.9	1.0	2.5	**0.9**	3695	66
	2hpl	DDLYG CCCCC	2.5	2.2	1.7	1.8	4707	1002
	average		3.5	3.6	3.1	2.2	4168	526
test set	1awr	HAGPIA CCCCCC	3.3	**0.8**	3.1	**0.5**	5187	1
	1eg4	NMTPYRSPPPYVP TTTTTTCCCCCCC	8.3	3.4	7.4	2.7	5289	592
	1elw	GPTIEEVD CCCCCCCC	4.3	**0.5**	4.2	**0.0**	5074	1
	1jwg	DEDLLHI CCCCCCC	4.3	**0.6**	3.4	**0.0**	4981	1
	1lvm	ENLYFQ CCEEEC	3.1	**0.2**	3.1	**0.0**	4528	1
	1ntv	NFDNPVYRKT CEETTTTCCC	4.3	3.0	4.1	2.7	4750	1093
	1nvr	ASVSA CEEEC	3.4	**0.5**	3.3	**0.2**	5020	1
	1nx1	DAIDALSSDFT HHHHHHHHHCC	**1.0**	6.0	**0.8**	5.8	1	2623
	1ou8	GAANDENY CCCCCCCC	3.4	3.6	3.3	3.3	4811	968
	1rxz	KSTQATLERWF CEEECTTTTTC	4.1	7.0	3.8	3.0	4246	1130

continued

test set	1ssh	GPPPAMPARPT CCCCCCCCCCC	7.4	**0.2**	7.2	**0.1**	5331	1
	1u00	ELPPVKIHC CCCCEECCC	4.3	4.2	4.1	**0.2**	5067	9
	1x2r	LDEETGEFL CTTTTTCCC	4.8	**0.4**	3.5	**0.3**	3461	1
	2b9h	RRNLKGLNLNLH CCTTTTCCCCCC	4.6	6.5	4.5	4.3	4976	329
	2c3i	KRRRHPSG CCCCCCCC	4.6	**0.9**	4.0	**0.3**	5125	1
	2h9m	ARTKQ TTTTC	**0.7**	2.4	**0.6**	2.2	1	767
	2o02	GLLDALDLAS THHHHHCCCC	3.4	2.0	2.7	1.2	4265	883
	3d1e	GQLGLF CBCCCC	3.4	2.3	3.0	2.1	4930	72
	average		4.0	2.5	3.7	1.6	4280	471

[a]Training and test sets are the same as published by Alam et al.[17] except for the swapping of peptides 2fmf and 1jwg. [b]PDB ID of the peptide's native conformation in complex with a cognate receptor. [c]Secondary structures computed with STRIDE[46] as published by Alam et al. [d]RMSD between the native conformation and the lowest-energy decoy. [e]Lowest RMSD between the native conformation and any decoy whose energy is in the lowest 1%. [f]Rank of the lowest-energy decoy having an RMSD ≤ 1 Å from the native conformation. [g]Default weights. [h]Optimized weights. RMSD values ≤ 1 Å are highlighted in bold; that shown to be equal to 1.0 but not in bold is really > 1 at the precision level chosen for the RMSD calculations (i.e., three decimal places). RMSD was computed over atoms C^{α}.

Native discrimination function

Drawing from our previous experience with this matter[47], we first extracted the unweighted values of the eight *cen_std+score4L* energy terms for all decoy conformations of all benchmark peptides. We then looked for a new set of weights $\boldsymbol{w}^* = (w_1^*,...,w_8^*)$ that minimized a native discrimination function f_{ntv} defined as

$$f_{ntv}(\boldsymbol{w}) = \sum_{p \in training\ set} Z_p(\boldsymbol{w}) , \qquad (1)$$

where $Z_p(\boldsymbol{w})$ is the "native discrimination z-score" for peptide p under weight set $\boldsymbol{w}$, analogous to a statistical metric that proved effective in the field of microarray data analysis[48]:

$$Z_p(\boldsymbol{w}) = \frac{\mu_{p,ntv}(\boldsymbol{w}) - \mu_{p,other}(\boldsymbol{w})}{\sigma_{p,ntv}(\boldsymbol{w}) + \sigma_{p,other}(\boldsymbol{w})}. \quad (2)$$

In this equation, considering all decoy conformations generated for peptide p, $\mu_{p,ntv}(\boldsymbol{w})$ and $\mu_{p,other}(\boldsymbol{w})$ are the average energies of native-like and non-native decoys, respectively, as obtained from weight set $\boldsymbol{w}$, while $\sigma_{p,ntv}(\boldsymbol{w})$ and $\sigma_{p,other}(\boldsymbol{w})$ are the corresponding standard deviations. As stated above, native-like decoys are defined to be those with C^{α}-RMSD $\leq$ 1Å from the native conformation, whereas non-native decoys are all other decoys. The energy of decoy d has the usual Rosetta form,

$$E(d, \boldsymbol{w}) = \sum_{i=1}^{8} w_i \, e_i(d), \quad (3)$$

where w_i is the weight of the ith energy term and $e_i(d)$ is the unweighted value of the ith energy term for decoy d. By minimizing f_{ntv}, therefore, we aimed at a weight set that tended to assign much lower energies to native-like than to non-native conformations. We also expected that the capability of the optimized weights to discriminate native-like conformations should hold in general, as the training set was chosen to be representative of various types of amino acid sequence and secondary structure (Table 2).

Rosetta version

All calculations were performed with Rosetta 2018.12.

RESULTS

Rosetta energy weight optimization

Energy weights were optimized using the Nelder–Mead simplex algorithm[49] as implemented in the GNU Scientific Library (http://www.gnu.org/software/gsl/). Because this is a local optimization algorithm, we ran it multiple times from different initial weight sets and then selected the optimized weight set that produced the lowest value of f_{ntv} over all runs. We chose ten initial weight sets: Eight of them had one weight equal to 1.0 and all other weights equal to 0.1; the other two initial weight sets were the default weights of the *cen_std* and *score4L* functions and a uniform weight set where all weights were equal to 1.0, respectively. The 10 runs converged on similar weight sets as to the sign and relative magnitude of each weight (Table S2). The final values of f_{ntv} were between –8.54 and –8.11, with clear improvement relative to the initial f_{ntv} values, all between –0.12 and +6.81 (Figure 1). The minimal value of f_{ntv}, –8.54, was obtained in two different runs. As the optimized weight sets of these two runs also showed an essentially identical native discrimination over all 18 test peptides (Figure S2), we arbitrarily chose one to report on, namely, that produced from the initial weights labeled *cbeta_10x*. Prior to testing, the set was normalized with respect to the *env* weight, an operation made possible by the invariance of f_{ntv} to multiplication of the weight set by a constant. Such normalized weights are listed in Table 1, and their improved discrimination of native-like conformations is the subject of the next section.

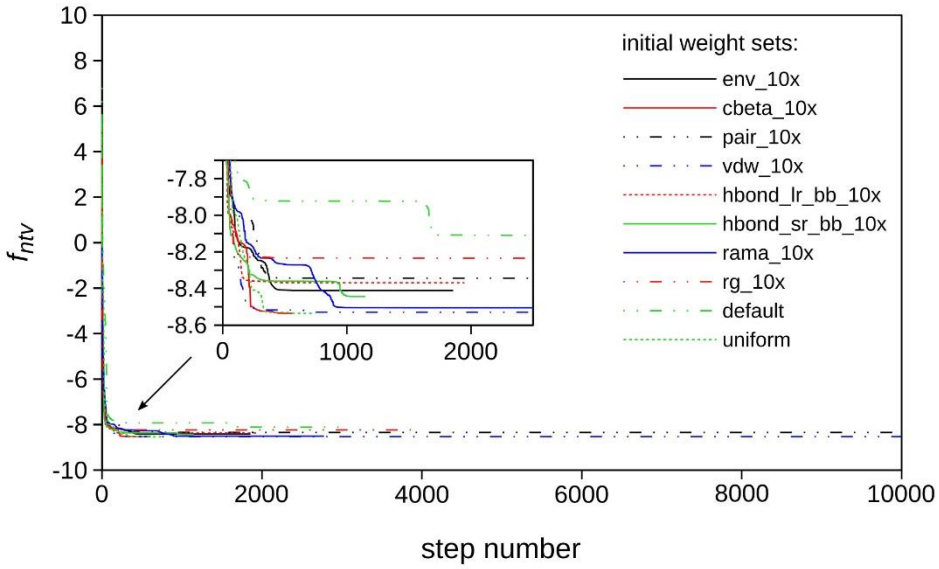

Figure 1. Convergence of the native discrimination function in the ten weight optimizations. The value of the native discrimination function (f_{ntv}) is plotted as a function of step number for the ten different runs of the Nelder–Mead optimization algorithm. Each run started from a different weight set. For each energy term *eterm*, the set of initial weights labeled *eterm_10x* assigned a weight of 1.0 to *eterm* and a weight of 0.1 to all other energy terms. The *default* initial set assigned energy terms their default weights in the *cen_std* and *score4L* energy functions. The *uniform* initial set assigned all energy terms a weight of 1.0. The optimal f_{ntv} value of –8.54 was achieved by two runs, started from weight sets *cbeta_10x* and *uniform*, respectively.

Discrimination of native-like peptide conformations

As shown in Figure 2B, native discrimination under the optimized weights improved for 8 training peptides out of 9 (triangles below the diagonal). More importantly, assessment on the 18 peptides of the test set revealed that native discrimination improved for 16 peptides (Figures 2A, 2C). Here, while default weights discriminate in favor of native-like decoys (i.e., Z_p is negative) for 3 peptides, optimized weights do so for 16 peptides. In these favorable cases, the discrimination signal is

larger than noise (i.e., $Z_p < -1$) for no peptide under default weights and for 9 peptides under optimized weights. Thus, not only did weight optimization produce an energy function that better discriminates native-like conformations; it also turned an energy function that is rarely effective at discriminating native-like conformations into one that is effective half of the time. The improvement in native discrimination is comparable to that obtained by optimizing weights on the original training set of this benchmark (Figure S3), which indicates that our weight optimization procedure was also robust to the swapping of a pair of peptides between training and test set.

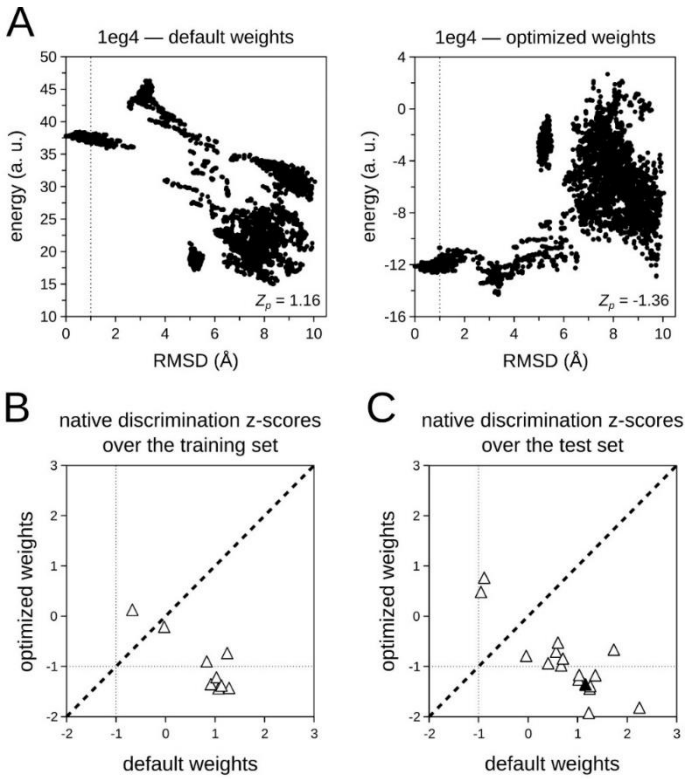

Figure 2. Weight optimization improved discrimination of native-like peptide conformations. (A) Example of the improvement in native discrimination z-score (Z_p) caused by weight optimization. For all 7,054 decoy conformations of test peptide 1eg4, energy (arbitrary units) was computed with the *cen_std+score4L* function under default (left) or optimized (right) weights. Clearly,

weight optimization turned native-like decoys (points to the left of the 1-Å RMSD line) from energetically unfavorable to favorable compared to non-native decoys (points to the right). (B–C) Native discrimination z-scores of the *cen_std+score4L* energy function under optimized versus default weights, for the 9 training peptides (B) and the 18 test peptides (C). Negative z-scores indicate discrimination in favor of native-like conformations, as defined in eq 2. If z-scores are < –1, the discrimination signal is larger than noise. The diagonal line denotes equal performance. The filled triangle in (C) represents peptide 1eg4. Here note, too, that only 17 triangles are visible because those of peptides 1eg4 and 1awr are perfectly overlapping.

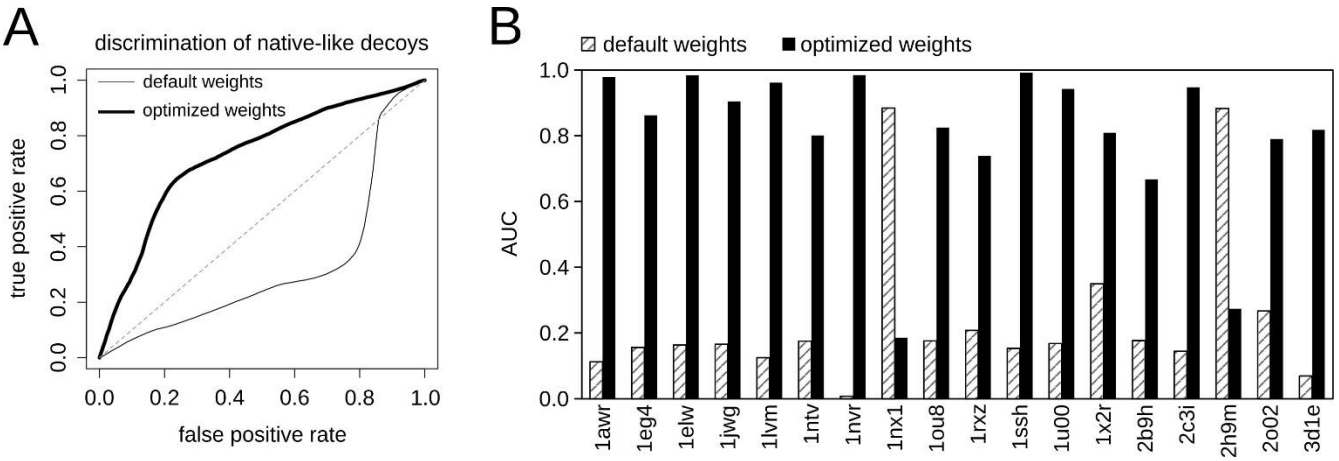

Figure 3. Optimized weights often define a better classifier of native-like decoys than default weights. (A) ROC curves for the discrimination of native-like decoys (RMSD from native ≤ 1 Å) by default and optimized weights, respectively, as plotted over all 126,972 decoys of the test peptides with discrimination thresholds given by the energy ranks ($\in \{1,\ldots,7054\}$). The diagonal line denotes random classification. The plot was made with the ROCR package[50]. (B) For each test peptide, area under the ROC curve (AUC) for the default weights versus that for the optimized weights, where both curves were plotted as in (A) over the peptide's 7,054 decoys.

Optimized weights perform better than default weights even if the discrimination of native-like versus non-native decoys is evaluated in terms of binary classification. Receiver operating characteristics (ROC) analysis over all decoys of the test set (Figure 3A) shows that decoy classification as a function of energy rank produces a larger area under the curve (AUC) with optimized weights (AUC=0.73) than with default weights (AUC=0.32). Accordingly, AUC values are 2–130-fold higher for optimized weights than for default weights on 16 of the 18 test peptides (Figure 3B and Figure S4). For these 16 peptides, which are the same 16 for which Z_p also improved (Figure 2C), classification by default weights is always worse than random (AUC $\leq$ 0.35), whereas classification by optimized weights is always better than it (AUC $\geq$ 0.67). Because the fraction of native-like decoys for these 16 peptides is between 7% and 23% (Table S1), one would conclude that the observed improvement in the discrimination of native-like decoys is independent of the precise mixture of native-like and non-native decoys. However, evidence to the contrary is provided by peptides 1nx1 and 2h9m, which have the highest fractions of native-like decoys (37% and 33%, respectively) and are the only two peptides for which performance after weight optimization became worse.

Native discrimination z-score (eq. 2) and AUC are both measures of separation between two energy distributions: that of native-like decoys versus that of non-native decoys. A related information that is also useful in practice is whether, among a small group of top-ranking (i.e., lowest-energy) decoys, at least some are native-like. In particular, for rigid-body docking tasks, this information helps one figure out what fraction of top-ranking, pregenerated peptide conformations should be destined to docking with some guarantee that native-like conformations, if any, will be represented. We therefore compared default and optimized weights in terms of three

additional metrics that verify the presence of native-like decoys among those that are top-ranking, as summarized in Table 2. Considering the test set first, the lowest energy decoy is native-like for 2 peptides under default weights and for 8 peptides under optimized weights (column "top-1 RMSD"). A similarly substantial increase, from 2 to 9 peptides, is observed when native-like decoys are sought in the lowest 1% of the energy distribution ("top-1% RMSD"). Moreover, the lowest rank of a native-like decoy is, on average, 9 times better under optimized than under default weights ("native rank"). Nevertheless, despite weight optimization there are still 6 peptides for which native-like decoys are not even ranked in the top 10%. These results are consistent with those on the training set, even though in the "top-1 RMSD" category the optimized weights outperformed the default weights only on 5 of the 9 training peptides. We therefore expect that given an ensemble of peptide conformations of which some are native-like, rescoring them with the optimized weights will make it ~4 times more likely for a native-like conformation to be included among the tiny fraction selected for rigid-body docking.

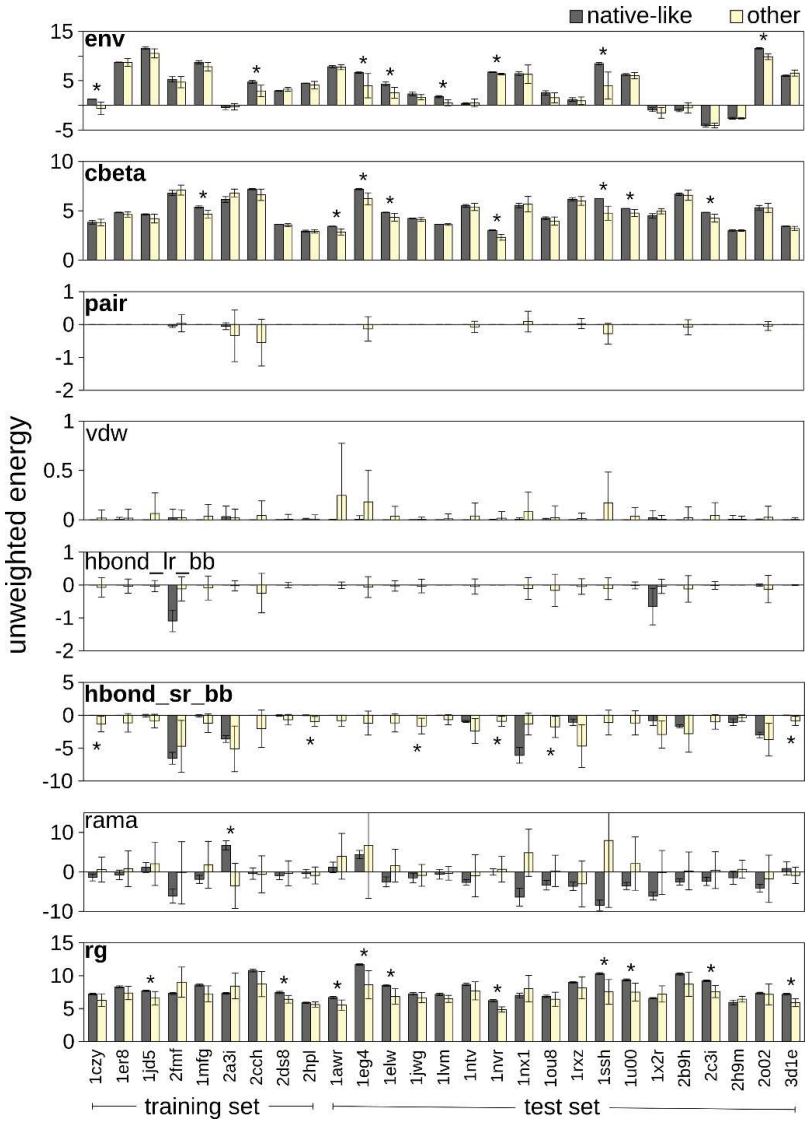

Figure 4. Negative optimized weights often correspond to unfavorable energies of native-like conformations. Average unweighted energies (arbitrary units) over native-like and non-native (other) decoy conformations of the benchmark peptides are reported for each of the eight energy terms of the *cen_std+score4L* function. Error bars represent standard deviations. Energy terms whose optimized weights are negative are printed in bold. Asterisks mark peptides for which the discrimination against native-like conformations (*ntv*) by the energy term has a signal that is larger than noise, namely, $\mu_{ntv} - \mu_{other} > \sigma_{ntv} + \sigma_{other}$, where μ_D and σ_D represent, respectively, average and standard deviation of the energy term over decoy set D.

Optimized weights reflect the relative energies of native-like versus non-native conformations

The optimized weights are listed in Table 1. For five energy terms (*env*, *cbeta*, *pair*, *hbond_sr_bb*, and *rg*) the optimized weight is negative. Because a negative weight favors conformations with higher values of its corresponding energy term, we expected those five energy terms to frequently have higher values for native-like decoys than for non-native decoys, at least in the training set. To test this hypothesis, we compared for each energy term the distribution of energy values over native-like decoys with the distribution of energy values over non-native decoys. The results for all peptides of the benchmark are illustrated in Figure 4 and summarized in Table S3. For four of the five energy terms with a negative weight (*env*, *cbeta*, *hbond_sr_bb*, and *rg*), native-like decoys have a higher average energy than non-native decoys ($\mu_{ntv} > \mu_{other}$) for 78% or more of the training set. Also, the *pair* term has $\mu_{ntv} > \mu_{other}$ for two of the three training peptides on which it is defined (i.e., peptides longer than 9 residues[33]; for shorter peptides, *pair* energies are always zero). By contrast, the *vdw* and *rama* terms have $\mu_{ntv} > \mu_{other}$ for at most 33% of the training set, in accordance with their positive weight. Unexpected results were instead obtained from the *hbond_lr_bb* term, which has a positive weight but $\mu_{ntv} > \mu_{other}$ for the majority (88%) of the training set. Optimization of this weight, however, might have been dominated by peptide 2fmf, whose value of $\mu_{ntv} - \mu_{other}$ for the *hbond_lr_bb* term is negative and has a magnitude at least 4-fold larger than all others. Similar trends of native versus non-native energy distributions were observed in the test set (Figure 4 and Table S3). Hence, if the energy terms of the *cen_std+score4L* function retained their modeling power even for peptides in solution, one would expect peptide native-like conformations to be less energetically favorable than non-native conformations in

terms of solvation (*env*, *cbeta, rg*), non-local electrostatics (*pair*, *hbond_lr_bb*, *hbond_sr_bb*), and non-local attractive van der Waals interactions (*rg*); by contrast, their repulsive van der Waals interactions (*vdw*, *rama*) and their local electrostatic and attractive van der Waals interactions (*rama*) should be more favorable.

Understanding why native-like conformations have higher H-bond energies

The energetic differences observed in Figure 4 between native-like and non-native conformations can be better understood by decomposing each energy term into its elementary energetic contributions, as specified in the term's own definition[32, 33, 39]. Such a deeper level of analysis was of especially clear interpretation in the case of the *hbond_lr_bb* and *hbond_sr_bb* terms, revealing that native-like conformations differ from non-native conformations in the number of intramolecular hydrogen bonds. The results of that analysis are presented here below.

The *hbond_lr_bb* and *hbond_sr_bb* terms evaluate the energy of backbone–backbone H-bonds at sequence separations ≥ 5 and < 5 residues, respectively[32, 41]. A peptide's unweighted H-bond energy is defined for both terms as

$$E_{Hbond} = \sum_{H,A:\, E_{HA}<0} w_H w_A E_{HA} \,, \tag{4}$$

where E_{HA}, the energy of a single H-bond, is computed with the same orientation-dependent formula used in Rosetta's *fullatom* mode[32, 39]:

$$E_{HA} = f(E_1(d_{HA}) + E_2(\theta_{AHD}) + E_3(\theta_{BAH}) + E_4(\rho, \theta_{BAH}, \chi_{BA})) \,. \tag{5}$$

In these equations, H and A denote a donor hydrogen atom and an acceptor atom, respectively; w_H and w_A are scaling factors that depend on the atom types of H and A, respectively; f is a function that smooths out the transition between favorable (< 0) and unfavorable (≥ 0) energies; d_{HA}, θ_{AHD}, θ_{BAH}, and χ_{BA} are the four geometric parameters illustrated in Figure 5A; ρ represents the orbital hybridization of the acceptor atom; and E_1–E_4 are analytic functions that were fitted in order that protein structural models generated by Rosetta optimally recapitulated crystallographic native structures in terms of distributions of H-bond features. Atoms H and A are considered to be H-bonded if and only if $E_{HA} < 0$.

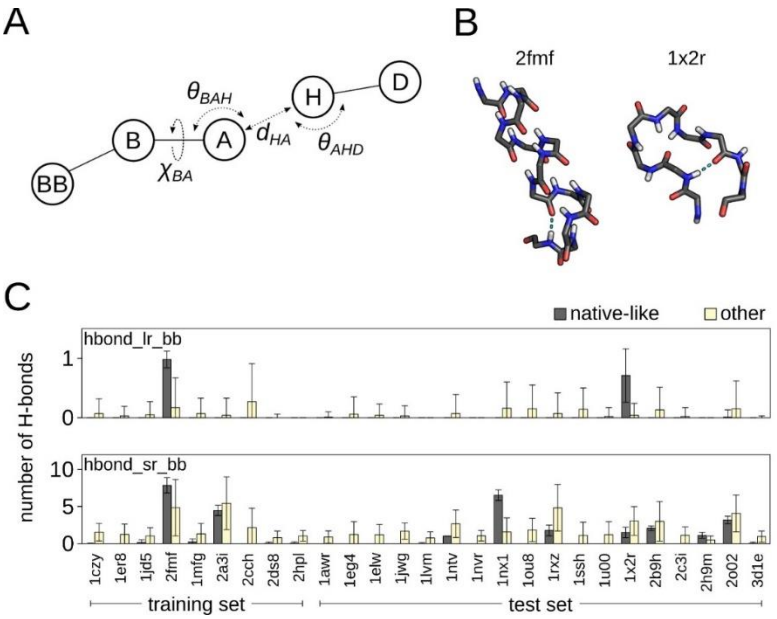

Figure 5. Backbone–backbone H-bonds are less frequent in native-like conformations than in non-native conformations. (A) Geometric parameters defining H-bond energy in Rosetta. d_{HA}: hydrogen–acceptor distance; θ_{AHD}: acceptor–hydrogen–donor angle; θ_{BAH}: acceptor's base–acceptor–hydrogen angle; χ_{BA}: acceptor type-specific dihedral angle around the acceptor's base–acceptor bond. (B) Native backbone structures of peptides 2fmf and 1x2r. These are the only two native structures in the benchmark to have a backbone–backbone H-bond with sequence separation ≥ 5 (cyan broken line). (C) Average number of backbone–backbone H-bonds at sequence

separations ≥ 5 (*hbond_lr_bb*) and < 5 (*hbond_sr_bb*) over native-like and non-native (other) decoys, respectively, for each peptide of the benchmark; error bars represent standard deviations.

hbond_lr_bb energies are, on average, higher for native-like decoys than for non-native decoys ($\mu_{ntv} > \mu_{other}$) for 88% of the training peptides (Figure 4 and Table S3), and yet the optimized weight of the *hbond_lr_bb* term, 0.8, is well into the positive range (Table 1). As mentioned above, optimization of this weight might have been dominated by the only training peptide with $\mu_{ntv} < \mu_{other}$, namely, 2fmf, for which the value of $\mu_{ntv} - \mu_{other}$ (–0.97) has at least 4-fold larger magnitude than that of any other peptide. Peptide 2fmf is the only training peptide whose native-like decoys contain any backbone–backbone H-bonds that are long-ranged (i.e., with sequence separation ≥ 5), as shown in Figure 5C. In particular, the H-bond between Phe13's NH and Leu8's O (Figure 5B) is contained in 98% of 2fmf's native-like decoys. By contrast, for all training peptides (except 2hpl, whose *hbond_lr_bb* energy is always zero by definition), there always exist non-native decoys which contain long-ranged H-bonds, but they represent only a small minority of non-native decoys (7% on average). From this we gather that optimization of the *hbond_lr_bb* weight converged so decidedly on a positive value because the Phe13–Leu8 H-bond in peptide 2fmf's native-like decoys outweighed the long-ranged H-bonds in all non-native decoys of the training set.

Extending the analysis to all 27 peptides of the benchmark, we observe that long-ranged H-bonds are absent in all native-like decoys of 24 peptides, whereas they are present in 0.1–19% of non-native decoys for 23 peptides (Figure 5C). Hence, in accordance with a previous survey of peptide–protein complexes[22], while long-ranged H-bonds should be accessible to a peptide in solution, they tend to be avoided when the peptide binds to a protein. Indeed, these intramolecular

backbone–backbone H-bonds may prevent the formation of peptide–protein interactions, both polar and hydrophobic, that are more favorable for the complex.

The tendency of backbone–backbone H-bonds to be less frequent in native-like decoys than in non-native decoys is also present, though less dramatically, at sequence separations < 5. The average number of these short-ranged H-bonds is lower in native-like decoys for 24 of 27 peptides (Figure 5C). Also, short-ranged H-bonds are present in at most 22% of native-like decoys for 18 peptides, whereas they are present in at least 42% of non-native decoys for every peptide. Such a disproportion helps explain why average *hbond_sr_bb* energies are higher for native-like decoys than for non-native decoys over 89% of the benchmark (Figure 4 and Table S3). The only peptides for which the average number of short-ranged H-bonds is higher in native-like decoys than in non-native decoys are peptides 2fmf and 1nx1, whose native conformation is mostly helical, and the 5-residue peptide 2h9m. However, the presence of short-ranged H-bonds in native-like decoys does not show any clear dependence on length or secondary structure (Table 2 and Figure 5C). A general conclusion that we can draw from these results is that native peptide conformations do not optimize the number of short-ranged H-bonds except when helical or, less likely, when very short. Nonetheless, short-ranged H-bonds are expected to be more accessible to native peptide conformations than long-ranged H-bonds.

DISCUSSION

Evidence has accumulated that peptides bind to their target proteins by "conformational selection"[51–55], motivating an approach to peptide–protein docking based on pregenerating peptide conformations that are metastable in solution[13]. In order for this pregeneration to yield a complete ensemble and yet occur in a reasonable time, the peptide may need to be represented

by a low-resolution model. The Rosetta centroid model[33] is a low-resolution model that was successfully used, among others, in the initial stages of loop modeling[36] and peptide–protein docking[10]. The low-resolution energy function chosen for both tasks is one that combines two functions known as *cen_std* and *score4L*, respectively (Table 1). Because both tasks arguably required an accurate modeling of solvent effects, we thought it worth to investigate whether this *cen_std*+*score4L* energy function suits as well the task of predicting peptide metastable conformations in solution. The present study assessed the ability of the *cen_std*+*score4L* energy function (not including the *chainbreak* term) to favor peptide native-like conformations among thousands of decoy conformations, all evaluated without the receptor. The study is therefore a first step in that ampler investigation, inspired by the conformational selection paradigm.

With default weights, the *cen_std*+*score4L* energy function was generally unable to discriminate native-like from non-native conformations. Out of 18 test peptides, it assigned lower average energies to native-like than to non-native decoys only in 3 cases (Figure 2C), and classified decoys worse than a random classifier for 16 peptides (Figure 3B). Moreover, for 16 test peptides no native-like decoy had an energy in the lowest 1% (Table 2). These results clearly discourage from using the default *cen_std*+*score4L* weights for predicting peptide metastable conformations in solution. The same results support the notion that energy functions statistically derived from native protein structures, as *cen_std*+*score4L* largely is[33, 37–40], are unlikely to do well at modeling peptides. While this may intuitively seem obvious, the above mentioned success of combining *cen_std* and *score4L* to model partly solvated regions[10, 36] spoke to the contrary, so a merit of the present work is to provide clarifying evidence in this regard.

Nonetheless, with optimized weights, the *cen_std*+*score4L* function was able to discriminate in favor of native-like decoys for 16 of the 18 test peptides, reaching effectiveness ($Z_p < -1$) in 9

cases (Figure 2C). For those 16 peptides classification became better than random, with AUC ≥ 0.67 (Figure 3B). Optimized weights also resulted in native-like decoys being ranked in the top 1% for 9 peptides, for 8 of which a native-like decoy had the lowest energy of all (Table 2). The optimized *cen_std+score4L* energy function is therefore expected to have ~50% chance of effectively discriminating native-like from non-native peptide conformations in terms of energy distributions; and it is also expected to have ~50% chance of including one or more native-like conformations, if any, among a tiny pool of lowest-energy rigid-body-docking candidates. We are aware that these estimates of performance assume a conformational distribution similar to those of our decoy sets, which, in spite of their heterogeneity (Table S1), may not constitute fully representative samples of conformational space. We also note, however, that these are quite conservative estimates, since our chosen RMSD threshold for native-likeness, 1 Å, is a strict one. If, for instance, the threshold were increased to 2 Å, the present optimized weights would rank native-like decoys in the top 1% for one additional peptide (Table 2). Impressively, the described enhancement of discriminatory power was achieved by mere weight optimization, that is, without changing the functional form of any of the eight energy terms. The robustness of our optimization procedure to changes in the initial weights (Table S2, Figures 1 and S2) and in the set of training peptides (Figures 2B–C and S3, Tables 1 and S4) suggests that native discrimination, especially where it remained poor, such as for peptides 1ntv, 1rxz, 1ou8, and 2o02, can hardly be improved if these energy terms are left unchanged. Improvement may rather occur by substitution of the more knowledge-based terms with equivalent terms derived from structural databases that represent peptide–protein interactions either directly[22, 56] or indirectly[20].

The optimized weights of five energy terms (*env*, *cbeta*, *pair*, *hbond_sr_bb*, and *rg*) are negative (Table 1). This prompted a comparative analysis of energy-term distributions between native-like

and non-native decoys, which verified our initial hypothesis that negative weights should correspond to unfavorable energies for native-like decoys (Figure 4, Table S3). Hence, given that *env*, *cbeta*, and *pair* were derived from protein statistics[33, 37, 38], the analysis suggests that peptide native conformations recapitulate worse than non-native conformations the solvation patterns and non-local electrostatic interactions observed in protein native structures. Peptide native conformations are also characterized by uncompactedness, by favorable local electrostatic and van der Waals interactions, and by few backbone–backbone H-bonds (Figures 4 and 5). Having unfavorable H-bond energies could therefore be part of the price that native conformations pay in solution for being selected by receptors as optimal fits to their binding site. Binding may also impose energetic costs to solvation and non-local electrostatics, but it is quite unwarranted to link them to the above-discussed lower recapitulation of native protein structures, particularly at shorter peptide lengths, where recapitulation becomes harder.

A natural follow-up to the present study is to test whether the optimized *cen_std+score4L* function can bias a conformational search towards native-like conformations. We anticipate that successful convergence of the search will depend on the type of sampling mechanism that will be coupled to the energy function. For example, if sampling is not based on fragment assembly, the current optimized *rama* weight may be too low in magnitude for an efficient rejection of sterically forbidden (φ, ψ) pairs. Also, because the optimized *vdw* weight has a low magnitude too, it may be more efficient to enforce the self-avoidance of conformations within the sampling mechanism itself, rather than allowing and then probabilistically discarding major atomic collisions. A classical way to measure the quality of conformational sampling is to compute how often the ensemble of selected dockable candidates contains a native-like conformation. However, the optimal size of that ensemble, one that affords the best trade-off between native coverage and

computational cost of the subsequent docking, may vary with peptide length and secondary structure[57]. Hence, a fair evaluation of conformational search under our optimized weights requires consideration of a spectrum of sampling options and selection thresholds, and so it is better deferred to a dedicated study.

The past decade has seen a proliferation of methods for peptide–protein docking[9] in response to the emergence of peptides as a new class of inhibitors of protein–protein interactions[1]. While the usefulness of these methods is undisputable, all of them might greatly benefit, with little or no adjustment, from the possibility to start docking from a peptide conformation that resembles the native bound conformation. In this work we have reported on the ability of the Rosetta *cen_std+score4L* energy function to discriminate in favor of such native conformations. Rosetta is a popular macromolecular modeling suite[31] that has been successfully used in modeling tasks that range from protein structure prediction[58] to the design of high-affinity protein binders[59] to the docking of small molecules[60] and peptides[10]. While Rosetta has a default energy function for high-resolution modeling[32], each modeling task performed at low resolution must explicitly specify its own energy function. As described above, *cen_std+score4L* is a low-resolution energy function essentially identical to that adopted in Rosetta protocols of loop modeling and peptide–protein docking. Moreover, Rosetta *ab initio* protein structure prediction[61] is guided in its low-resolution stages by energy functions that incorporate the *env*, *pair*, *vdw*, *cbeta*, and *rg* terms. We therefore conclude that our work has shed light on receptor-free peptide modeling by energy terms that are central to Rosetta low-resolution modeling. Finally, because native discrimination under the optimized weights has improved to an encouraging 50% success rate, we recommend that Rosetta users attempting to pregenerate dockable peptide conformations should try these weights first.

Availability of data and materials

Additional data and material are provided as Supplementary Information. These include:

Table S1. Distributions of decoy conformations

Table S2. Convergence of the ten weight optimizations on similar weight sets

Table S3. Negative optimized weights often correspond to unfavorable energies of native-like conformations

Table S4. Optimized weights produced with Alam et al.'s training set

Figure S1. Starting decoy conformations

Figure S2. Optimized weights from the *cbeta_10x* and *uniform* initial weights are essentially equivalent in discriminating native-like decoys

Figure S3. Optimized weights improve discrimination of native-like peptide conformations even with Alam et al.'s training set

Figure S4. Optimized weights often define a better classifier of native-like decoys than default weights

Competing interests

The authors declare no competing interests

Funding

No specific funding was received for this work.

Author Contributions

AB performed coding, experiments, analyses and wrote the article. AC conceived the work, supervised experiments, critically analyzed results and edited the final version of the article.

Acknowledgements

We thank Dr. Maxim Shapovalov for help with the *rama* term.

Abbreviations

PPIs: protein–protein interactions

PDB: Protein Data Bank

RMSD: root mean square displacement

AUC: area under the curve

REFERENCES

1. Bruzzoni-Giovanelli H, Alezra V, Wolff N, et al (2018) Interfering Peptides Targeting Protein–Protein Interactions: the Next Generation of Drugs? Drug Discov Today 23:272–285. https://doi.org/10.1016/j.drudis.2017.10.016

2. Ryan DP, Matthews JM (2005) Protein-Protein Interactions in Human Disease. Curr Opin Struct Biol 15:441–446. https://doi.org/10.1016/j.sbi.2005.06.001

3. Arkin MR, Tang Y, Wells JA (2014) Small-Molecule Inhibitors of Protein-Protein Interactions: Progressing toward the Reality. Chem Biol 21:1102–1114. https://doi.org/10.1016/j.chembiol.2014.09.001

4. Nero TL, Morton CJ, Holien JK, et al (2014) Oncogenic Protein Interfaces: Small Molecules, Big Challenges. Nat Rev Cancer 14:248–262. https://doi.org/10.1038/nrc3690

5. Petsalaki E, Russell RB (2008) Peptide-mediated Interactions in Biological Systems: New Discoveries and Applications. Curr Opin Biotechnol 19:344–350. https://doi.org/10.1016/j.copbio.2008.06.004

6. Stumpf MPH, Thorne T, De Silva E, et al (2008) Estimating the Size of the Human Interactome. Proc Natl Acad Sci U S A 105:6959–6964. https://doi.org/10.1073/pnas.0708078105

7. Venkatesan K, Rual JF, Vazquez A, et al (2009) An Empirical Framework for Binary Interactome Mapping. Nat Methods 6:83–90. https://doi.org/10.1038/nmeth.1280

8. Berman HM, Westbrook J, Feng Z, et al (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242. https://doi.org/10.1093/nar/28.1.235

9. Ciemny M, Kurcinski M, Kamel K, et al (2018) Protein–Peptide Docking: Opportunities and Challenges. Drug Discov Today 23:1530–1537. https://doi.org/10.1016/j.drudis.2018.05.006

10. Raveh B, London N, Zimmerman L, Schueler-Furman O (2011) Rosetta FlexPepDock ab-Initio: Simultaneous Folding, Docking and Refinement of Peptides onto Their Receptors. PLoS One 6:e18934. https://doi.org/10.1371/journal.pone.0018934

11. Trellet M, Melquiond ASJ, Bonvin AMJJ (2013) A Unified Conformational Selection and Induced Fit Approach to Protein-Peptide Docking. PLoS One 8:e58769. https://doi.org/10.1371/journal.pone.0058769

12. Schindler CEM, De Vries SJ, Zacharias M (2015) Fully Blind Peptide-Protein Docking with pepATTRACT. Structure 23:1507–1515. https://doi.org/10.1016/j.str.2015.05.021

13. Ben-Shimon A, Niv MY (2015) AnchorDock: Blind and Flexible Anchor-driven Peptide Docking. Structure 23:929–940. https://doi.org/10.1016/j.str.2015.03.010

14. Kurcinski M, Jamroz M, Blaszczyk M, et al (2015) CABS-dock Web Server for the Flexible Docking of Peptides to Proteins without Prior Knowledge of the Binding Site. Nucleic Acids Res 43:W419–W424. https://doi.org/10.1093/nar/gkv456

15. Yan C, Xu X, Zou X (2016) Fully Blind Docking at the Atomic Level for Protein-Peptide Complex Structure Prediction. Structure 24:1842–1853. https://doi.org/10.1016/j.str.2016.07.021

16. Peterson LX, Roy A, Christoffer C, et al (2017) Modeling Disordered Protein Interactions from Biophysical Principles. PLoS Comput Biol 13:e1005485. https://doi.org/10.1371/journal.pcbi.1005485

17. Alam N, Goldstein O, Xia B, et al (2017) High-Resolution Global Peptide-Protein Docking Using Fragments-Based PIPER-FlexPepDock. PLoS Comput Biol 13:e1005905. https://doi.org/10.1371/journal.pcbi.1005905

18. Lamiable A, Thévenet P, Rey J, et al (2016) PEP-FOLD3: Faster Denovo Structure Prediction for Linear Peptides in Solution and in Complex. Nucleic Acids Res 44:W449–W454. https://doi.org/10.1093/nar/gkw329

19. Zhou P, Li B, Yan Y, et al (2018) Hierarchical Flexible Peptide Docking by Conformer Generation and Ensemble Docking of Peptides. J Chem Inf Model 58:1292–1302. https://doi.org/10.1021/acs.jcim.8b00142

20. Vanhee P, Stricher F, Baeten L, et al (2009) Protein-Peptide Interactions Adopt the Same

Structural Motifs as Monomeric Protein Folds. Structure 17:1128–1136. https://doi.org/10.1016/j.str.2009.06.013

21. Diella F, Haslam N, Chica C, et al (2008) Understanding Eukaryotic Linear Motifs and Their Role in Cell Signaling and Regulation. Front Biosci 13:6580–6603. https://doi.org/10.2741/3175

22. London N, Movshovitz-Attias D, Schueler-Furman O (2010) The Structural Basis of Peptide-Protein Binding Strategies. Structure 18:188–199. https://doi.org/10.1016/j.str.2009.11.012

23. Tsai CJ, Buyong M, Sham YY, et al (2001) Structured Disorder and Conformational Selection. Proteins Struct Funct Genet 44:418–427. https://doi.org/10.1002/prot.1107

24. Csermely P, Palotai R, Nussinov R (2010) Induced Fit, Conformational Selection and Independent Dynamic Segments: An Extended View of Binding Events. Trends Biochem Sci 35:539–546. https://doi.org/10.1016/j.tibs.2010.04.009

25. Camacho CJ, Gatchell DW, Kimura SR, Vajda S (2000) Scoring Docked Conformations Generated by Rigid-body Protein-Protein Docking. Proteins Struct Funct Genet 40:525–537. https://doi.org/10.1002/1097-0134(20000815)40:3<525::AID-PROT190>3.0.CO;2-F

26. Tsai J, Bonneau R, Morozov A V., et al (2003) An Improved Protein Decoy Set for Testing Energy Functions for Protein Structure Prediction. Proteins Struct Funct Genet 53:76–87. https://doi.org/10.1002/prot.10454

27. Niu Huang, Shoichet BK, John J. Irwin (2006) Benchmarking Sets for Molecular Docking. J Med Chem 49:6789–6801. https://doi.org/10.1021/nl061786n.Core-Shell

28. Maffucci I, Contini A (2016) An Updated Test of AMBER Force Fields and Implicit Solvent Models in Predicting the Secondary Structure of Helical, β-Hairpin, and Intrinsically Disordered Peptides. J Chem Theory Comput 12:714–727. https://doi.org/10.1021/acs.jctc.5b01211

29. Shao Q, Shi J, Zhu W (2017) Determining Protein Folding Pathway and Associated Energetics through Partitioned Integrated-Tempering-Sampling Simulation. J Chem Theory Comput 13:1229–1243. https://doi.org/10.1021/acs.jctc.6b00967

30. Wang W, Liang T, Sheong FK, et al (2018) An Efficient Bayesian Kinetic Lumping Algorithm to Identify Metastable Conformational States via Gibbs Sampling. J Chem Phys 149:072337. https://doi.org/10.1063/1.5027001

31. Leman JK, Weitzner BD, Lewis SM, et al (2020) Macromolecular Modeling and Design in Rosetta: Recent Methods and Frameworks. Nat Methods 17:665–680. https://doi.org/10.1038/s41592-020-0848-2

32. Alford RF, Leaver-Fay A, Jeliazkov JR, et al (2017) The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J Chem Theory Comput 13:3031–3048. https://doi.org/10.1021/acs.jctc.7b00125

33. Rohl CA, Strauss CEM, Misura KMS, Baker D (2004) Protein Structure Prediction Using Rosetta. Methods Enzymol 383:66–93. https://doi.org/10.1016/S0076-6879(04)83004-0

34. Das R, Baker D (2008) Macromolecular Modeling with Rosetta. Annu Rev Biochem 77:363–382. https://doi.org/10.1146/annurev.biochem.77.062906.171838

35. Mandell DJ, Coutsias EA, Kortemme T (2009) Sub-angstrom Accuracy in Protein Loop

Reconstruction by Robotics-Inspired Conformational Sampling. Nat Methods 551–552. https://doi.org/10.1038/nmeth0809-551

36. Stein A, Kortemme T (2013) Improvements to Robotics-Inspired Conformational Sampling in Rosetta. PLoS One. https://doi.org/10.1371/journal.pone.0063090

37. Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Functions. J Mol Biol 268:209–225. https://doi.org/10.1006/jmbi.1997.0959

38. Simons KT, Ruczinski I, Kooperberg C, et al (1999) Improved Recognition of Native-Like Protein Structures Using a Combination of Sequence-Dependent and Sequence-Independent Features of Proteins. Proteins Struct Funct Genet 34:82–95. https://doi.org/10.1002/(SICI)1097-0134(19990101)34:1<82::AID-PROT7>3.0.CO;2-A

39. O'Meara MJ, Leaver-Fay A, Tyka MD, et al (2015) Combined Covalent-Electrostatic Model of Hydrogen Bonding Improves Structure Prediction with Rosetta. J Chem Theory Comput 11:609–622. https://doi.org/10.1021/ct500864r

40. Shapovalov M V., Dunbrack RL (2011) A Smoothed Backbone-dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. Structure 19:844–858. https://doi.org/10.1016/j.str.2011.03.019

41. Shmygelska A, Levitt M (2009) Generalized Ensemble Methods for de Novo Structure Prediction. Proc Natl Acad Sci U S A 106:1415–1420. https://doi.org/10.1073/pnas.0812510106

42. Wang C, Bradley P, Baker D (2007) Protein-Protein Docking with Backbone Flexibility. J Mol Biol 373:503–519. https://doi.org/10.1016/j.jmb.2007.07.050

43. Gront D, Kulp DW, Vernon RM, et al (2011) Generalized Fragment Picking in Rosetta: Design, Protocols and Applications. PLoS One 6:e23294. https://doi.org/10.1371/journal.pone.0023294

44. Buchan DWA, Jones DT (2019) The PSIPRED Protein Analysis Workbench: 20 years on. Nucleic Acids Res 47:W402–W407. https://doi.org/10.1093/nar/gkz297

45. Metropolis N, Rosenbluth AW, Rosenbluth MN, et al (1953) Equation of State Calculations by Fast Computing Machines. J Chem Phys 21:1087–1092. https://doi.org/10.1063/1.1699114

46. Frishman D, Argos P (1995) Knowledge-Based Protein Secondary Structure Assignment. Proteins Struct Funct Bioinforma 23:566–579. https://doi.org/10.1002/prot.340230412

47. Bazzoli A, Kelow SP, Karanicolas J (2015) Enhancements to the Rosetta Energy Function Enable Improved Identification of Small Molecules That Inhibit Protein-Protein Interactions. PLoS One 10:e0140359. https://doi.org/10.1371/journal.pone.0140359

48. He Z, Zhou J (2008) Empirical Evaluation of a New Method for Calculating Signal-to-Noise Ratio for Microarray Data Analysis. Appl Environ Microbiol 74:2957–2966. https://doi.org/10.1128/AEM.02536-07

49. Nelder JA, Mead R (1965) A Simplex Method for Function Minimization. Comput J 7:308–313. https://doi.org/10.1093/comjnl/7.4.308

50. Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCR: Visualizing classifier

performance in R. Bioinformatics 21:3940–3941. https://doi.org/10.1093/bioinformatics/bti623

51. Yoon MK, Venkatachalam V, Huang A, et al (2009) Residual Structure within the Disordered C-terminal Segment of p21 Waf1/Cip1/Sdi1 and Its Implications for Molecular Recognition. Protein Sci 18:337–347. https://doi.org/10.1002/pro.34

52. Cino EA, Choy WY, Karttunen M (2013) Conformational Biases of Linear Motifs. J Phys Chem B 117:15943–15957. https://doi.org/10.1021/jp407536p

53. Krieger JM, Fusco G, Lewitzky M, et al (2014) Conformational Recognition of an Intrinsically Disordered Protein. Biophys J 106:1771–1779. https://doi.org/10.1016/j.bpj.2014.03.004

54. Olson MA (2018) Conformational Selection of a Polyproline Peptide by Ebola Virus VP30. Proteomics 18:1800081. https://doi.org/10.1002/pmic.201800081

55. Miyanabe K, Akiba H, Kuroda D, et al (2018) Intramolecular H-bonds Govern the Recognition of a Flexible Peptide by an Antibody. J Biochem 164:65–76. https://doi.org/10.1093/jb/mvy032

56. Vanhee P, Reumers J, Stricher F, et al (2009) PepX: A Structural Database of Non-redundant Protein Peptide Complexes. Nucleic Acids Res 38:D545–D551. https://doi.org/10.1093/nar/gkp893

57. Yan Y, Zhang D, Huang S-Y (2017) Efficient Conformational Ensemble Generation of Protein-Bound Peptides. J Cheminform 9:59. https://doi.org/10.1186/s13321-017-0246-7

58. Ovchinnikov S, Kinch L, Park H, et al (2015) Large-Scale Determination of Previously

Unsolved Protein Structures Using Evolutionary Information. Elife 4:e09248. https://doi.org/10.7554/eLife.09248

59. Fleishman SJ, Whitehead TA, Ekiert DC, et al (2011) Computational Design of Proteins Targeting the Conserved Stem Region of Influenza Hemagglutinin. Science (80-) 332:816–821. https://doi.org/10.1126/science.1202617

60. DeLuca S, Khar K, Meiler J (2015) Fully Flexible Docking of Medium Sized Ligand Libraries with RosettaLigand. PLoS One 10:e0132508. https://doi.org/10.1371/journal.pone.0132508

61. Bradley P, Misura KMS, Baker D (2005) Toward High-Resolution de Novo Structure Prediction for Small Proteins. Science (80-) 309:1868–1871. https://doi.org/10.1126/science.1113801

Figures

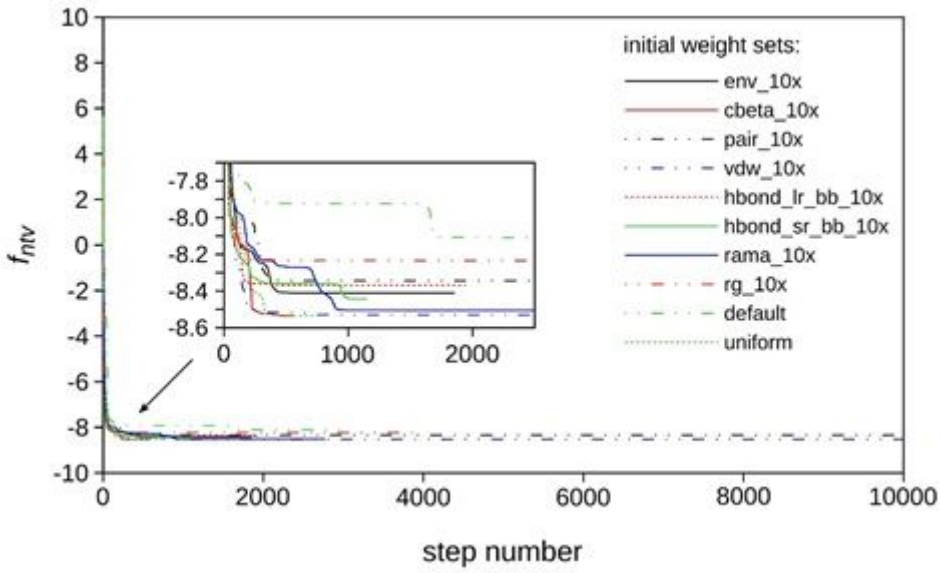

Figure 1

Convergence of the native discrimination function in the ten weight optimizations. The value of the native discrimination function (fntv) is plotted as a function of step number for the ten different runs of the Nelder–Mead optimization algorithm. Each run started from a different weight set. For each energy term eterm, the set of initial weights labeled eterm_10x assigned a weight of 1.0 to eterm and a weight of 0.1 to all other energy terms. The default initial set assigned energy terms their default weights in the cen_std and score4L energy functions. The uniform initial set assigned all energy terms a weight of 1.0. The optimal fntv value of −8.54 was achieved by two runs, started from weight sets cbeta_10x and uniform, respectively.

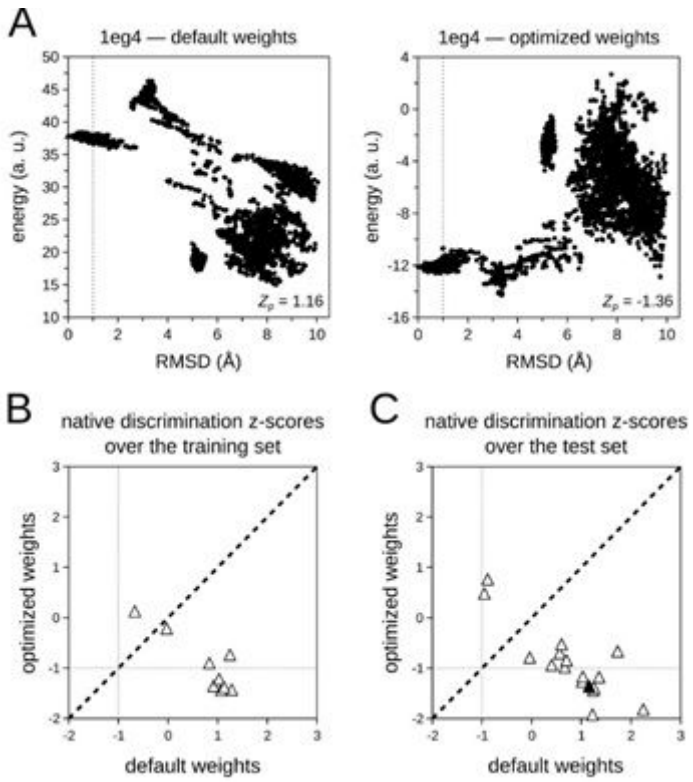

Figure 2

Weight optimization improved discrimination of native-like peptide conformations. (A) Example of the improvement in native discrimination z-score (Zp) caused by weight optimization. For all 7,054 decoy conformations of test peptide 1eg4, energy (arbitrary units) was computed with the cen_std+score4L function under default (left) or optimized (right) weights. Clearly, weight optimization turned native-like decoys (points to the left of the 1-Å RMSD line) from energetically unfavorable to favorable compared to non-native decoys (points to the right). (B–C) Native discrimination z-scores of the cen_std+score4L energy function under optimized versus default weights, for the 9 training peptides (B) and the 18 test peptides (C). Negative z-scores indicate discrimination in favor of native-like conformations, as defined in eq 2. If z-scores are < −1, the discrimination signal is larger than noise. The diagonal line denotes equal performance. The filled triangle in (C) represents peptide 1eg4. Here note, too, that only 17 triangles are visible because those of peptides 1eg4 and 1awr are perfectly overlapping.

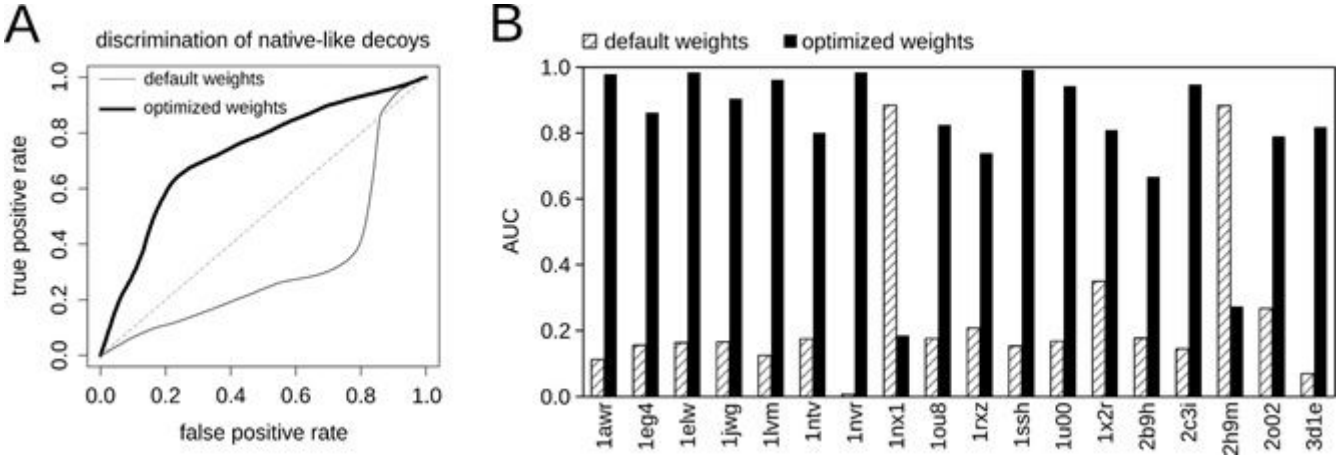

Figure 3

Optimized weights often define a better classifier of native-like decoys than default weights. (A) ROC curves for the discrimination of native-like decoys (RMSD from native ≤ 1 Å) by default and optimized weights, respectively, as plotted over all 126,972 decoys of the test peptides with discrimination thresholds given by the energy ranks ({1,…,7054}). The diagonal line denotes random classification. The plot was made with the ROCR package[50]. (B) For each test peptide, area under the ROC curve (AUC) for the default weights versus that for the optimized weights, where both curves were plotted as in (A) over the peptide's 7,054 decoys.

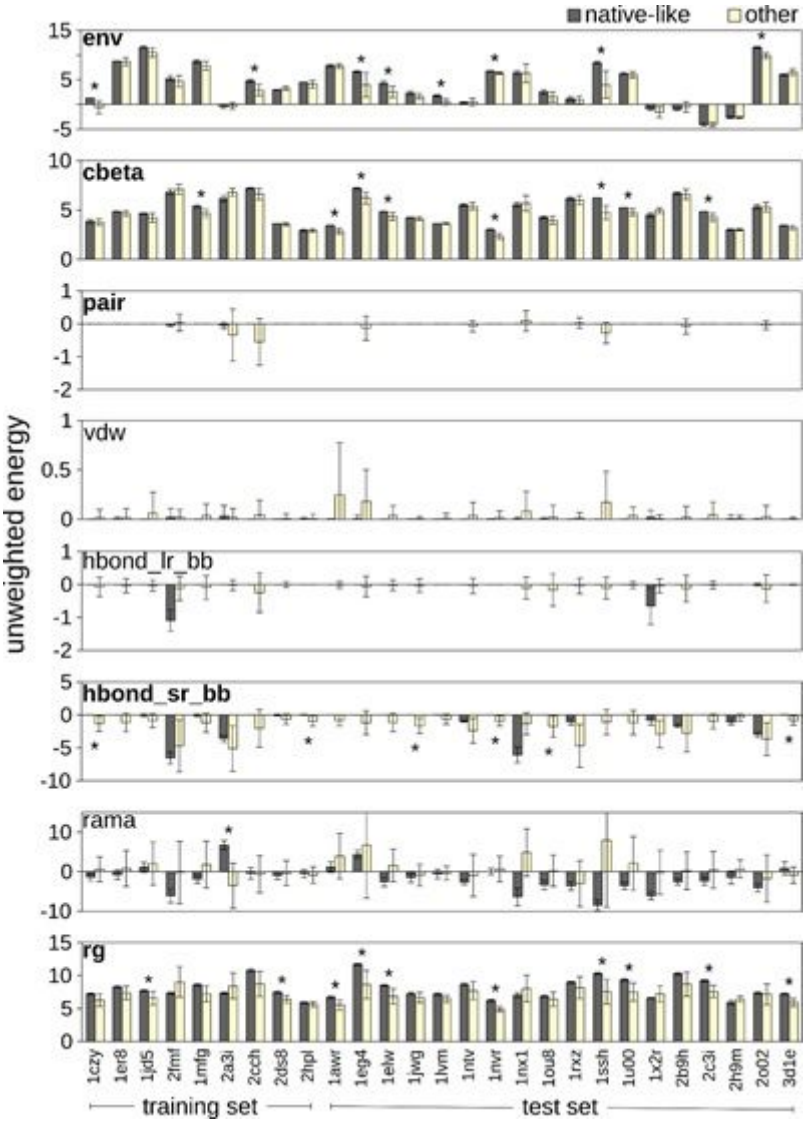

Figure 4

Negative optimized weights often correspond to unfavorable energies of native-like conformations. Average unweighted energies (arbitrary units) over native-like and non-native (other) decoy conformations of the benchmark peptides are reported for each of the eight energy terms of the cen_std+score4L function. Error bars represent standard deviations. Energy terms whose optimized weights are negative are printed in bold. Asterisks mark peptides for which the discrimination against native-like

conformations (ntv) by the energy term has a signal that is larger than noise, namely, μntv − μother > σntv + σother, where μD and σD represent, respectively, average and standard deviation of the energy term over decoy set D.

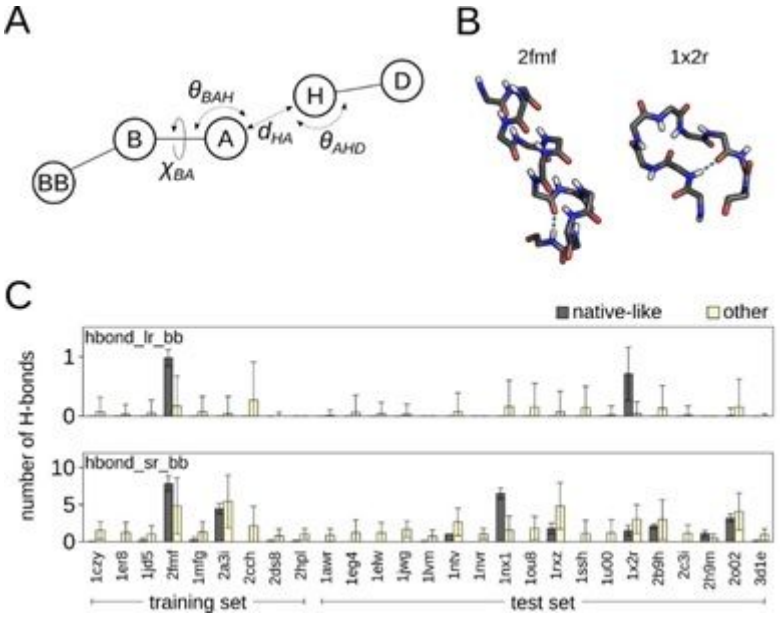

Figure 5

Backbone–backbone H-bonds are less frequent in native-like conformations than in non-native conformations. (A) Geometric parameters defining H-bond energy in Rosetta. dHA: hydrogen–acceptor distance; θAHD: acceptor–hydrogen–donor angle; θBAH: acceptor's base–acceptor–hydrogen angle; χBA: acceptor type-specific dihedral angle around the acceptor's base–acceptor bond. (B) Native backbone structures of peptides 2fmf and 1x2r. These are the only two native structures in the benchmark to have a backbone–backbone H-bond with sequence separation ≥ 5 (cyan broken line). (C) Average number of backbone–backbone H-bonds at sequence separations ≥ 5 (hbond_lr_bb) and < 5 (hbond_sr_bb) over native-like and non-native (other) decoys, respectively, for each peptide of the benchmark; error bars represent standard deviations.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation310820.docx