JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

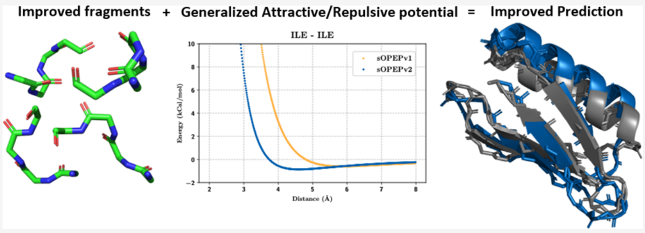
A Generalized Attraction−Repulsion Potential and Revisited Fragment Library Improves PEP-FOLD Peptide Structure Prediction

Vincent Binette, Normand Mousseau,* and Pierre Tuffery*

Cite This: https://doi.org/10.1021/acs.jctc.1c01293 Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: Fast and accurate structure prediction is essential to the study of peptide function, molecular targets, and interactions and has been the subject of considerable efforts in the past decade. In this work, we present improvements to the popular simplified PEP-FOLD technique for small peptide structure prediction. PEP-FOLD originality is threefold: (i) it uses a predetermined structural alphabet, (ii) it uses a sequential algorithm to reconstruct the tridimensional structures of these peptides in a discrete space using a fragment library, and (iii) it assesses the energy of these structures using a coarse-grained representation in which all of the backbone atoms but the α-hydrogen are present, and the side chain corresponds to a unique bead. In former versions of PEP-FOLD, a van der Waals formulation was used for non-bonded interactions, with each side chain being associated with a fixed radius. Here, we explore the relevance of using instead a generalized formulation in which not only the optimal distance of interaction and the energy at this distance are parameters but also the distance at which the potential is zero. This allows each side chain to be associated with a different radius and potential energy shape, depending on its interaction partner, and in principle to make more effective the coarse-grained representation. In addition, the new PEP-FOLD version is associated with an updated library of fragments. We show that these modifications lead to important improvements for many of the problematic targets identified with the former PEP-FOLD version while maintaining already correct predictions. The improvement is in terms of both model ranking and model accuracy. We also compare the PEP-FOLD enhanced version to state-of-the-art techniques for both peptide and structure predictions: APPTest, RaptorX, and AlphaFold2. We find that the new predictions are superior, in particular with respect to the prediction of small β-targets, to those of APPTest and RaptorX and bring, with its original approach, additional understanding on folded structures, even when less precise than AlphaFold2. With their strong physical influence, the revised structural library and coarse-grained potential offer, however, the means for a deeper understanding of the nature of folding and open a solid basis for studying flexibility and other dynamical properties not accessible to IA structure prediction approaches.



INTRODUCTION

Proteins are macromolecules involved in a wide variety of crucial biological processes. Their functions are determined by their tridimensional structure as well as their dynamics and thermodynamics properties. Thus, the characterization of protein folding, particularly how the tridimensional structure of protein is encoded in its amino acid sequence, is of great interest in molecular biology.[1] Since the end of the Human Genome Project, the development of next-generation sequencing has led to a drastic decrease in cost and a drastic increase in the number and diversity of determined genomes,[2] creating a growing gap between the number of known sequences and known structures. The development of fast and accurate protein structure prediction techniques is required to study not only the characteristics of the proteins themselves but also their interactions with partners such as peptides and small molecules, as protein structure predictions play a key role in the design of new therapeutic molecules. The discovery and development of 210 new molecules approved by the US Food and Drug Administration between 2010 and 2016, for example, were facilitated by structural information available in the Protein Data Bank.[3]

Progress in the numerical predictions of protein tridimensional structure is monitored by the Critical Assessment of Techniques for Protein Structure Prediction (CASP) meetings.[4] In recent years, the utilization of multiple sequence alignments (MSAs) with protein sequences derived from genomic sequencing taken from huge data sets combined with very successful machine learning techniques, such as RaptorX,[5−7] RosettaFold,[8] or the now state-of-the-art AlphaFold2,[9] has led to tremendous improvement of the predicted results nearing experimental accuracy for some targets.

Received: December 22, 2021

However, the CASP meetings are mainly focused on fairly large proteins of a couple hundred (to a few thousand) amino acids. For example, only three targets tested in CASP14 have less than 70 amino acids. However, many small peptides, of less than a few dozen amino acids, have interesting properties. For instance, antimicrobial peptides of such size[10,11] could be crucial in the mitigation of antibiotic resistance, which, according to some experts, could lead to 10 million yearly deaths by 2050.[12] Newly emerging interfering peptides also belong to these sizes.[13]

Small peptides present a unique challenge compared to large proteins,[14] and multiple computational approaches utilizing a wide variety of techniques have been developed to target specifically the peptide secondary and tertiary structure prediction. For example, PSSP-MVIRT is a successful deep-learning method for the prediction of peptide secondary structure.[15] PEPstr[16] (and its extension to nonstandard amino acid, PEPstrMod[17]) utilizes the observation on the prevalence of β-turn secondary structure to add constraints on molecular dynamics simulation to predict peptide tertiary structure. In the parallel microgenetic algorithm (PMGA)[18] techniques, peptides' structure predictions are done by utilizing a genetic algorithm with backbone dihedral angle correlations for sampling a density functional theory derived fitness function. Finally, the recently developed APPTest[19] was developed by combining distance/angle constraints derived by a neural network with simulated annealing, resulting in great structural predictions for small peptides.

In this study, we present improvements to PEP-FOLD,[20−22] a quick and highly simplified approach for small peptide structure prediction. The PEP-FOLD software is freely available as a Web server[23] and has been used in a variety of applications, such as the very recent research of a SARS-CoV-2 treatment.[24−26] The PEP-FOLD approach is based on three main features: (1) the concept of structural alphabet (SA), (2) discrete fragment assembly, and (3) a coarse-grained energy function. We present here improvements to two of these key features.

First, the fragment library is reworked to better sample the conformational variability associated with the letters of the structural alphabet. Second, we revisit the coarse-grained energy function. The coarse-grained energy function used in PEP-FOLD is based on the **O**ptimized **P**otential for **E**fficient Structure **P**rediction (OPEP). Compared to other force fields such as that of CABS-fold,[27] OPEP originality comes from the coarse-grained representation and the treatment of hydrogen bonds. The OPEP representation includes all atoms from the backbone except the α-hydrogen and represents side chains using only one bead. This detailed backbone representation makes possible an explicit account for hydrogen bonds, necessary to support the OPEP-specific treatment for cooperativity in hydrogen bonds, that favors secondary structure formation during folding. Over the years, the OPEP force field has been successfully applied to a wide variety of biophysical applications,[28] including the self-assembly of amyloid protein,[29] associated with many neurodegenerative diseases like Alzheimer's, the study of DNA/RNA systems,[30] the peptide/protein docking,[31] and many more. More specifically, PEP-FOLD predictions are guided by a **s**implified version of the OPEP force field named sOPEP, that ignores most of the bonded energy terms due to the PEP-FOLD-specific assembly procedure that does not occur in a continuous space but in a discretized space using a limited number of fragments representative of the structural alphabet.[22] This rigid assembly process challenges the relevance of the non-bonded energy terms that are based on a van der Waals formulation. In OPEP, each particle is associated with one radius. This fixed radius can be problematic to optimally parametrize interactions that can occur under contradictory circumstances. For instance, a large radius could be relevant for a large side chain interacting with another large side chain but irrelevant for interactions with small side chains or the beads describing the backbone, leading to high energy values for interbead distances observable in structures. Several ways to overcome this limitation have been proposed in the literature such as the use of soft-core potentials,[32] or variations in the exponent values of the van der Waals terms, as proposed by Mie a long time ago[33] or more recently in the context of long-range corrections for dispersion interactions in inhomogeneous simulations.[34] However, none of these solutions addresses satisfactorily the requirement to have simultaneous control over the optimal distance $r0$, the energy at this distance, and the distance at which the energy is 0. In a previous study, we had proposed a formulation making this requirement possible for disulfide bonds.[35] Here we generalize this formulation to any exponent combination.

In former studies, the optimization of sOPEP was done on large ensembles of decoys generated with a wide variety of sampling techniques: molecular dynamics, threading, greedy assembly, etc. These sampling algorithms have different search spaces compared to PEP-FOLD, which could have an impact on the effectiveness of sOPEP. Second, the classification score used for the optimization was the TMscore,[36] a score based mainly on geometric factors (mean distance between corresponding Cα-atoms), while sOPEP energy terms mainly involve interatomic interactions (contacts, explicit hydrogen bonds, etc.). Finally, only a small portion of the parameters were optimized, while most of them were derived from experimental structures, with little consideration for interactions interdependence.

In this study, we present a reworking of the non-bonded interactions of sOPEP as well as the reoptimization of all of its energy components. We analyze how the combination of the newly improved fragments library and the newly optimized sOPEP potential impacts the quality of PEP-FOLD predictions, and we compare these to state-of-the-art approaches for both peptide and structure predictions.

■ MATERIALS AND METHODS

PEP-FOLD. PEP-FOLD relies on a hidden Markov model (HMM) derived structural alphabet (SA).[37] It consists of 27 letters that correspond to fragments of four residues overlapping by three residues. Thus, the 3D conformation of a peptide of length L can be described by $L - 3$ SA fragments.

More specifically, PEP-FOLD prediction of the 3D structure from the amino acid sequence is performed according to a three-step protocol:

1. **SA profile prediction**: PEP-FOLD first converts the amino acid sequence into a sequence of letters taken from the structural alphabet (SA). This is achieved by using a support vector machine (SVM) that takes as input a matrix of eight series of 20 values. The 20 values correspond to the position-specific scoring matrix as determined by PSI-BLAST.[38] These eight series correspond to those of the four amino acids of the fragment, extended by two on each side. The SVM has been trained to predict from the 160 values of the input the probability of being associated with each of the 27 letters. For a peptide of length L, the SVM is used iteratively for all $L -$

3 fragments of four amino acids. The result is a SA profile of size 27 × $(L - 3)$ that gives the probability of each fragment of the protein to be described by each of the 27 letters of the SA. Given the SA profile, the forward-backtrack algorithm (FBT) or a taboo sampling algorithm is then used to generate a specified number of suboptimal trajectories in the SA letter space from the SA profile.[20]

2. **Tridimensional reconstruction**: Each of these SA trajectories is used to generate a 3D model. First, an initial model is built from the rigid assembly of the fragments associated with each SA letter sequentially. The polypeptide chain is built by adding amino acid by amino acid, starting from an initial fragment of four amino acids. At each step, all possible conformations are generated by superimposing the fragments associated with the SA letter at the current position to the last three amino acids of the conformations generated at the previous step. A modified greedy algorithm[39] is used to filter the generated conformations at each step of reconstruction. A portion of the structures are kept based on their predicted energy according to the sOPEP force field[22] with the rest selected at random among the remaining structures.
3. **Monte Carlo**: As a final step, the resulting best conformations associated with each SA trajectory are then refined using a Monte Carlo procedure; at each Monte Carlo step, a fragment is randomly replaced by another and the modification is accepted based on a Metropolis criterion. Note that the fragments themselves are not allowed to modify their structure. It is the replacement of one fragment by another which conditions the change in the conformation.

For a more thorough description of the PEP-FOLD protocol, we refer the reader to the Lamiable et al. article.[20]

Library of Fragments. The first part of PEP-FOLD, which is revisited here, is the fragment library. The structure of a nonredundant collection of proteins was decoded as a series of strings of SA letters using the Viterbi or the forward−backward algorithms (see Camproux et al.[37]). Fragments of four amino acids associated with each of the 27 letters were collected, and for each letter, the clustering of the fragments allowed representative fragments of the letter to be identified, with their number depending on the conformational variability of the letter.

Two main changes are made with respect to the initial design of the library of fragments. The first one concerns the approach used to superimpose the fragments in order to generate a distance matrix between each of them. While superimposition was performed originally using the backbone of the four amino acids of the fragments, we opt here to superimpose only the three first amino acids to measure the RMSD between the fragments. This modification delivers a scheme that is expected to be more consistent with the HMM concept, as it allows a better measurement of the diversity of the position of the fourth amino acid.

The second is a change in the clustering itself. Instead of using dynamics clustering, we now use the Ward algorithm, as implemented in the *hclust* module of R, using the squared dissimilarity values (ward.D2). The resulting tree is then used to identify clusters separated by some arbitrary cutoff value. A similar value was used for all SA letters. In order to keep the calculation tractable, only a limited number of fragments was randomly drawn from the complete sets. A set of 5000 was found to be sufficient to ensure a satisfactory reproducibility. Cluster centroids used for the fragment assembly are taken as the fragment closest to all other members of the cluster. Finally, outlier clusters, i.e., those including less than 2.5% of the total number of fragments, are discarded. This threshold, which would allow up to 40 equally distributed clusters, is much lower than the expected frequency of well populated clusters whose number is, in practice, of the order of 15−20. From now on, the original and updated libraries will be referred to as Lib1 and Lib2, respectively.

sOPEP. One of the PEP-FOLD particularities is the use of a physics-based/knowledge-based coarse-grained potential, sOPEP, to discriminate between structures. This description plays a crucial role in guiding the 3D reconstruction as well as in the Monte Carlo refinement step.

The coarse-grained representation used in sOPEP is based on the OPEP[22,40] force field representation. The OPEP force field is a coarse-grained model where each amino acid is represented by a total of six pseudoatoms, as shown in Figure 1: the backbone is represented by five pseudoatoms for atoms N, H, C_α, C, and O, and a single pseudoatom is used to represent the side chain (SC). The position for the side chain (i) is fixed based on the $C(i - 1)$, $N(i)$, and $C_\alpha(i)$ positions and using predetermined centroid values.[41]

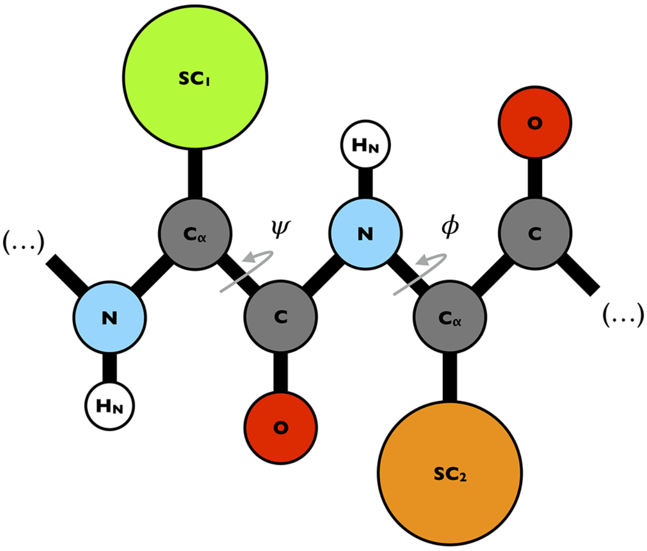


Figure 1. Coarse-grained representation in sOPEP. The backbone is represented in all-atoms format with all backbone atoms but HA—N, H_N, C_α, C, and O—present, while side chains are represented by a single interaction center. Also indicated are the ϕ and ψ dihedral angles.

The sOPEP potential is a variation on OPEP targeted at the specific structural alphabet approach of PEP-FOLD; it is composed of three main energy terms: bonded interactions (dihedral angles), non-bonded interactions (repulsion/dispersion effects), and explicit hydrogen bonds (with secondary structure cooperativity). In the following, the original formulation/parametrization is referred to as sOPEPv1, while the formulation/parametrization introduced here is called sOPEPv2. The complete formulation of sOPEPv2 as well as its key differences with sOPEPv1 are presented below.

Bonded Potential. The only bonded interaction considered in sOPEPv1 is the dihedral angle ϕ (Figure 1). In addition to the dihedral angle ϕ, sOPEPv2 also accounts for the dihedral angle ψ (Figure 1). These two dihedral angles are of crucial importance in the description of protein conformations, as demonstrated by the well-known Ramachandran plot. Since, in PEP-FOLD, the geometry is mainly imposed by the superimposition of the discrete SA letters, the impact of this addition is minimal, but it is added here for completeness.

As PEP-FOLD constrains the backbone by a rigid association of the fragments of the SA letters, sOPEPv2 uses a simple flat-bottomed quadratic potential to described the energy associated with dihedral angles ϕ described by

$$E_{\mathrm{rama}}(\phi_i) = \epsilon_{\phi}(\phi_i - \phi_{0\_\mathrm{sc}\_i})^2$$

where $\phi_{0\_\mathrm{sc}\_i} = \phi$ within the interval $[\phi_{\mathrm{low}\_\mathrm{sc}\_i}, \phi_{\mathrm{high}\_\mathrm{sc}\_i}]$ and $\phi_{0\_\mathrm{sc}\_i} = \min(\phi - \phi_{\mathrm{low}\_\mathrm{sc}\_i}, \phi - \phi_{\mathrm{high}\_\mathrm{sc}\_i})$ outside of the interval $\phi_{\mathrm{low}\_\mathrm{sc}\_i}$ and $\phi_{\mathrm{high}\_\mathrm{sc}\_i}$ are specific to each amino acid type.

sOPEPv2 uses the same equations for describing the dihedral ψ angle with adapted parameters.

Non-Bonded Potential. The potential associated with repulsion/dispersion effects was slightly reworked in sOPEPv2. For side chain−side chain interactions, sOPEPv1 adopted a dual formulation using either a repulsive term or a repulsive/attractive formulation based on the type of amino acids of the pairs.[22] In sOPEPv2, a repulsive/attractive formulation is adopted for all side chain pairs, and the same formalism is used for all non-bonded interactions between any pair of pseudoatoms, whereas, in sOPEPv1, the van der Waals formulation was in use for backbone−backbone and backbone−side chain interactions. Finally, non-bonded interactions including HN are not considered, and HN interactions are only considered for hydrogen bonds.

The repulsion/dispersion effects are described using the following Mie potential[33] given by

$$E_{\mathrm{vdw}\_ij} = \epsilon_{ij} \times \left[\frac{m}{n-m}\left(\frac{r_{ij}^0}{r_{ij}}\right)^n - \frac{n}{n-m}\left(\frac{r_{ij}^0}{r_{ij}}\right)^m \right] \qquad (1)$$

where ϵ_{ij} is the potential depth and r_{ij}^0 is the position of the potential minimum function of atomic types for i and j. The combination of exponents, n and m, gives the relationship between the position of the potential minimum (r^0) and the position where it is zero ($gR0$):

$$gR0 = \left(\frac{m}{n}\right)^{1/(n-m)} r_0 \qquad (2)$$

It is thus possible to have control over the well depth, its position, and the position where the potential is zero, but the slope at $gR0$ cannot be adjusted independently. This formulation makes it possible, to some extent, to limit the impact of the representation of the side chains using only one bead. sOPEPv1 parameters include ϵ_{ij} and $gR0_{ij}$ specific to each pseudoatom type pair and potential minimum defined by the sum of individual pseudoatom type radii: $r_{ij}^0 = r_i^0 + r_j^0$. sOPEPv2 retains the sOPEPv1 description for ϵ_{ij} and $gR0_{ij}$ (using a n_{ij}/m_{ij} combination) and optimizes r_{ij}^0 for each heavy atom type pair specifically. Moreover, as described above, all pseudoatom pair interactions include the attractive and repulsive terms. To make it compatible with sOPEPv1, the initial value for ϵ is set at 0.05 kcal/mol, similarly to side chain/backbone and backbone/backbone interactions.

Explicit Hydrogen Bond. Hydrogen bonds are considered explicitly in the OPEP family of potentials. sOPEPv2 keeps the same formulation as sOPEPv1: a hydrogen bond between residue i and residue j is characterized by the hydrogen/acceptor distance r_{ij} and the donor/hydrogen/acceptor angle α_{ij}. The hydrogen bond potential is defined as follows

$$E_{\mathrm{HB}}(r_{ij}, \alpha_{ij}) = \epsilon_{\alpha}^{\mathrm{HB}} \sum_{ij,j=i+4} \mu(r_{ij})\cdot\nu(\alpha_{ij}) + \epsilon_{\beta}^{\mathrm{HB}} \sum_{ij,j>4} \mu(r_{ij})\cdot\nu(\alpha_{ij}) \qquad (3)$$

$$\mu(r_{ij}) = \epsilon_{ij}\cdot\left[5\left(\frac{\sigma}{r_{ij}}\right)^{12} - 6\left(\frac{\sigma}{r_{ij}}\right)^{10}\right] \qquad (4)$$

$$\nu(\alpha_{ij}) = \begin{cases} \cos(\alpha_{ij}) & \text{if } \alpha_{ij} > 90^\circ \\ 0 & \text{otherwise} \end{cases} \qquad (5)$$

where σ is the position of the potential minimum and ϵ is the potential depth. We distinguish between α-helix-like hydrogen bonds defined by O(i)−H(i + 4) and other hydrogen bonds. Hydrogen bonds between a pair of residues separated by less than four amino acids are not considered.

sOPEP also includes a cooperativity term between hydrogen bond motifs present in secondary structure. In sOPEPv1, the cooperativity formulation involves a per-residue cooperativity propensity associated with α-helix and β-sheet.[22,41] sOPEPv2 integrates the cooperativity formulation of sOPEPv1 but does not include a residue-specific cooperativity propensity.

The cooperativity, which involves pairs of hydrogen bonds (between residues i and j and residues k and l), is used to stabilize secondary structure motifs and distinguishes between α-helix cooperativity and β-sheet cooperativity. The cooperativity energy is given by the following:

$$E_{\mathrm{coop}}(r_{ij}, r_{kl}) = \epsilon_{\alpha}^{\mathrm{coop}} \sum C(r_{ij}, r_{kl}) \times \Delta(ijkl) + \epsilon_{\beta}^{\mathrm{coop}} \sum C(r_{ij}, r_{kl}) \times \Delta'(ijkl)$$

$$C(r_{ij}, r_{kl}) = \exp(-0.5(r_{ij} - \sigma)^2)\exp(-0.5(r_{kl} - \sigma)^2)$$

$$\Delta(ijkl) = \begin{cases} 1 & \text{if } (k, l) = (i+1, j+1) \\ & \text{and } (j, l) = (i+4, k+4) \\ 0 & \text{otherwise} \end{cases}$$

$$\Delta'(ijkl) = \begin{cases} 1 & \text{if } (k, l) = (i+2, j-2) \\ & \text{or } (k, l) = (i+2, j+2) \\ 0 & \text{otherwise} \end{cases}$$

Optimization Protocol. The optimization of the sOPEPv2 parameters follows the basic optimization scheme developed for earlier versions of OPEP[22] and sOPEP.[22] The optimization process is designed to allow sOPEPv2 to discriminate between conformations almost identical to the experimental structure (native conformations), conformations resembling the experimental structure (near-native conformations), and the rest of the possible conformations (non-native conformations) without imposing additional biases associated with intermediary approximations such as all-atom force fields. This protocol mimics, in a general way, the funnel description of protein folding used to fit simplified force fields. However, while some optimization approaches focus on the native state,[42−44] we follow[41,45] by simply requiring that native structures have the lowest energy, followed native-like and non-native, defined using global scoring. We select to use this three-part classification, instead a more complex categorization built on more specific structural elements, such as in ref 46, due to the nature of the force field and its focus on shorter sequences.

More specifically, the decoys' classification defines a remarkably simple set of inequalities

$$\begin{aligned} &E(N_i) < E(L_j), \text{ for all } i, j \\ &E(N_i) < E(M_k), \text{ for all } i, k \\ &E(L_j) < E(M_k), \text{ for all } j, k \end{aligned} \tag{6}$$

where $E(X)$ is the sOPEP energy of a decoy X_h, being the $h = i, j, k$ element of the $X = N, L, M$, where N, L, and M correspond to the native, near-native, and non-native classes of decoys, respectively (see below). The optimization scheme uses these inequalities to classify an ensemble of decoys, on which the parameters are optimized.

The following sections will describe how (1) decoys are classified, (2) protein targets are selected for the parametrization/validation ensemble, (3) decoys are generated for each protein target, and finally (4) the optimization score and protocol are defined.

Decoys Classification. In order to define the set of inequalities given in eq 6, it is necessary to adopt a criterion for decoys classification, as there is no unique way to set up the classes. The optimization of sOPEPv1[22,41] used a decoy classification based on the TMscore.[36] Here, we select, rather, the CAD score, a score based on the similarity of interatomic contacts,[47,48] to classify decoys into the non-native, near-native, and native class. This score presents features that make it particularly suitable for optimizing sOPEPv2: (i) it is based on interatomic contacts, and it was shown (ii) to be more accurate in terms of the HB network similarity and (iii) to give more realistic stereochemical features according to MOLPROBITY[49] as compared to other highly used scores such as the TMscore (sOPEPv1), the GDT-TS, and the RMSD.[48] These features are well aligned with the sOPEP force field, which is based on interatomic interactions, explicit hydrogen bonds, and hydrogen bond cooperativity in secondary structure.

The CAD score[47,48] has been developed for the comparative analysis of protein structures at atomic resolution. It is defined as follows

$$\text{CAD} = 1 - \left(\frac{\sum_{(i,j)} \text{CAD}_{(i,j)}^{\text{bounded}}}{\sum_{(i,j)} T_{(i,j)}} \right)$$

$$\text{CAD}_{(i,j)}^{\text{bounded}} = \min(\text{CAD}_{(i,j)}, T_{(i,j)})$$

$$\text{CAD}_{(i,j)} = |T_{(i,j)} - M_{(i,j)}|$$

where $T_{(i,j)}$ and $M_{(i,j)}$ are the contact area between residues i and j for the target structure and model structure, respectively. Using an all-atom representation, the contact area is estimated using a Voronoi diagram of the heavy atoms described by hard spheres with a radius corresponding to their van der Waals radius. In order to compute the CAD score in the coarse-grained representation of OPEP, we define new atom types corresponding to the OPEP side chains with the radius taken from sOPEPv1. In the following, we refer to this score as CAD-CG.

Our classification is based on two main elements. We first consider the empirical distribution of the all-atom CAD score (CAD-AA) (and the cumulative distribution) presented by Olechnovic et al.[48] The overwhelming majority of the score is distributed between values of 0.3 and 0.7. More than 80% of the structures have a CAD-AA below 0.60, while more than 90% of the structures have a CAD-AA below 0.65. The second factor is based on the highest CAD-CG predictions generated by sOPEP1/Lib1. After visual inspection, we determine, in agreement with Olechnovic observations, that a CAD-CG above 0.60 is associated with largely correct secondary structure predictions while a CAD-CG of above 0.65 is associated with accurate secondary and tertiary structure. Thus, the native, near-native, and non-native classes are characterized by CAD score ranges of [0.65, 1.00], [0.60, 0.65], and [0.00, 0.60], respectively.

Selection of Protein Targets. The parametrization ensemble for optimizing sOPEPv1 contained 12 proteins or protein fragments: 1abz (α, 40 aa), 1dv0 (α, 47 aa), 1e0m (β, 37 aa), 1orc (α/β, 71 aa), 1pgb (α/β, 56 aa), 2gb1f (a β-fragment of 2gb1 spanning residues 41–56), 1qhk (α/β, 47 aa), 1shg (β, 62 aa), 1ss1 (α, 62 aa), 1vii (α, 36 aa), 2ci2 (α/β, 83 aa), and 2crofisa (α, 71aa).[22]

In order to improve the original sOPEP parametrization ensemble, we probe the Protein Data Bank for protein targets with the following characteristics: sequences that (1) have 70 amino acids or less, (2) are monomers, (3) contain only standard amino acids, (4) have a structure determined in a pH between 5.5 and 8.5, (5) are not membrane proteins, (6) are not making interactions with ions or ligands, and (7) show no more than 30% sequence similarity with others in the set. An additional six targets with more than 30% sequence similarity were added to the validation ensembles when they were considered in previous PEP-FOLD publications[20−22] (see the Supporting Information for the listing).

This leaves 135 protein targets, that we further divide into a parametrization and a validation ensemble. For each protein target, we generate the reference structure for the CAD-CG score computation in the following manner: we extract the first model from the Protein Data Bank, we minimize it using the all-atom force field AMBER99sb*-ILDN[50] using the GROMACS software,[51] and then we convert it to the sOPEP coarse-grained representation.

To try and minimize potential problems in the SVM part of PEP-FOLD and really focus on the potential optimization, we further classify the targets based on whether or not sOPEP1/Lib1 is able (i) to generate native predictions for the target, irrespective of their energy, and (2) to correctly assign a low energy to the native prediction with respect to near and non-native structures. The classification protocol is presented in Figure 2. For each target, we first generate 500 PEP-FOLD predictions with Lib1/sOPEPv1 and assign the resulting structure to one of the three classes using the CAD-CG (see the Decoys Classification section).

If the lowest energy prediction is in the same class (native/near-native) as the best generated prediction, the sequence target is placed in the *Generated/Correctly Classified* (G/CC) category because PEP-FOLD predictions (Lib1/sOPEPv1) are already able to provide reliable folding for this target. If, on the contrary, none of the generated structures are classified as native or near-native, the target sequence is placed in the *Not Generated* (NG) category: sOPEP1/Lib1 fails to produce a satisfactory folding. Finally, if predicted native or near-native structures are generated but do not correspond to the lowest energy prediction, the target sequence is placed in the *Generated/Incorrectly Classified* (G/IC) category, meaning that, for this target, sOPEP1/Lib1 is able to generate a good structural

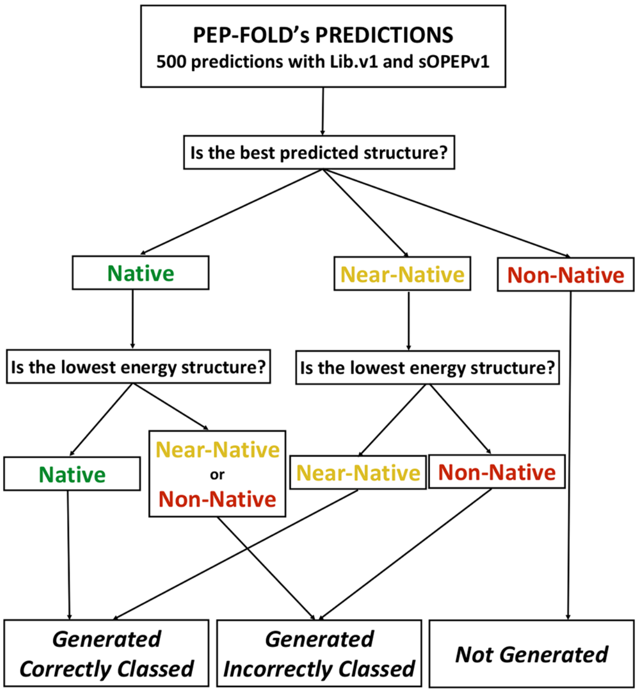

PEP-FOLD's PREDICTIONS
500 predictions with Lib.v1 and sOPEPv1
Is the best predicted structure?
Native
Near-Native
Non-Native
Is the lowest energy structure?
Is the lowest energy structure?
Native
Near-Native
or
Non-Native
Near-Native
Non-Native
Generated Correctly Classed
Generated Incorrectly Classed
Not Generated

Figure 2. Target classification based on PEP-FOLD predictions. Each protein target is classified in one of three ensembles: *Generated/Correctly Classified* (G/CC), *Generated/Incorrectly Classified* (G/IC), and *Not Generated* (NG). Targets for which the best predicted structure is in the non-native class are placed in ensemble NG (no predicted structure in the native or near-native class). Targets for which the lowest energy structure is in a worst class than the best predicted structure are placed in ensemble G/IC. Finally, targets for which the lowest energy structure is in the same class as the best predicted structure are placed in ensemble G/CC.

prediction, but its energy is high with respect to non-native and near-native structures.

Out of the 135 protein targets, 48, 64, and 23 sequences are placed in the G/CC, G/IC, and NG categories, respectively. The full list of targets is presented in the Supporting Information.

For the optimization, we focus our attention on the targets from the G/IC ensemble, since, for these targets, the potential is the primary hurdle to improvement of the predictions. From the 64 targets of the G/IC ensemble, we select 25 targets with special care given to contact and structural diversity in order to build the parametrization set. These selected targets are 1b03 (β, 18 aa), 1bhi (α/β, 38 aa), 1cpz (α/β, 68 aa), 1e0n (α/β, 27 aa), 1fex (α, 59 aa), 1g2h (α, 61 aa), 1go5 (α, 69 aa), 1i6c (β, 39 aa), 1jjs (α, 50 aa), 1spw (β, 39 aa), 1uxd (α, 65 aa), 1wcn (α, 70 aa), 1yiu (α/β, 37 aa), 1z4h (α/β, 66 aa), 1zv6 (α, 68 aa), 1zxg (α, 59 aa), 2b7e (α, 59 aa), 2bby (α/β, 69 aa), 2dt6 (α, 64 aa), 2fmr (α/β, 65 aa), 2l92 (β, 50 aa), 2l93 (α/β, 55 aa), 2lma (α, 22 aa), 2mwf (β, 32 aa), and 2ysb (β, 49 aa). More specifically, the optimization ensemble is composed of 11 α-proteins, 6 β-proteins, and 7 α/β-proteins. To show the diversity of included contacts, the side chain contact frequency of the targets in the parametrization ensemble is presented in Figure S1. Only the MET−MET contact is absent from the selected experimental structures. We also note few contacts with CYS, because targets forming disulfide bonds are excluded.

Decoys Generation. We then generate decoys on which to optimize the parameters for each protein target identified previously.

In sOPEPv1 optimization, decoys were generated using multiple techniques: (1) molecular dynamics, (2) threading, (3) greedy assembly, and (4) simulated annealing.[41] Between 430 and 928 decoys were generated for each target for an average of 550 decoys per target.

In the present work, all decoys are generated directly with the PEP-FOLD protocol. We use 500 suboptimal sequences in the SA letters space generated using the FBT algorithm. For the greedy algorithm, we use a heap size of 300, of which 100 are selected based on the sOPEP energy while 200 are randomly selected. Finally, the structures are refined using 30,000 Monte Carlo steps (see the PEP-FOLD section).

Parameters Optimization. Using the inequalities of eq 6 based on our decoys classification, we define the optimization score as follows

$$\text{Score} = -1.0 \cdot \frac{N_{\text{tot}}}{T} \sum_{t}^{T} \frac{1}{N_{\text{t}}} \sum_{i}^{C} \sum_{ii}^{D_t^i} \sum_{j>i}^{C} \sum_{jj}^{D_t^j} H(E_{jj} - E_{ii})$$

where N_{tot} is the total number of inequalities, T is the total number of targets, N_t is the number of inequalities associated with target t, C is the total number of decoy classes (native, near-native, and non-native) included in the evaluation, and D_t^i is the number of decoys for target t in class i. The sum over i and j is done over all decoy classes from native to non-native. $H(E_{jj} - E_{ii})$ is the Heaviside function which equals 1 if the energy of decoy ii is smaller than the energy of decoy jj: $E_{jj} - E_{ii} > 0$. To prevent that improvement in the resolution of inequalities from being dominated by a single target with more decoys, the score is normalized by the total number of inequalities for each target N_t.

The optimization of sOPEPv2 parameters is done using particle SWARM optimization (PSO),[52] as implemented in the *pyswarm* python package. This optimization technique works by moving a set of particles, each representing a candidate solution, iteratively in the search space according to the following velocity and position equations

$$\vec{V}_i(t+1) = \omega \vec{V}_i(t) + \phi_p r_p(\vec{p}_i - \vec{X}_i) + \phi_g r_g(\vec{g}_i - \vec{X}_i)$$

$$\vec{X}_i(t+1) = \vec{X}_i(t) + \vec{V}_i(t+1)$$

where ω, ϕ_p, and ϕ_g represent the inertia, the "cognitive" coefficient, and the "social" coefficient, respectively. $\vec{p}$ is the best position visited by each particle individually, and $\vec{g}$ is the global best position visited by the swarm. r_p and r_g are the real random numbers between 0 and 1. In this work, the particles correspond to the sOPEP2 parameters, and we use the default values for the inertia and the cognitive and social coefficients: $\omega = 0.5$, $\phi_p = 0.5$, and $\phi_g = 0.5$. Initial velocities are set randomly according to a uniform distribution; 500 particles are used for the optimization process, as described below.

Only a small fraction of the parameters were directly optimized for sOPEPv1: the parameters *r*0 and *gR*0 for the repulsion/dispersion interactions were determined directly from the distance distributions computed on the Protein Data Bank and were not optimized further; only the ϵ parameters of the repulsion/dispersion interactions were optimized.[22]

For sOPEPv2, all parameters are reoptimized. The pair potential involves 300 pairs of heavy atom type (210 side chain/side chain, 80 side chain/main chain, and 10 main chain/main chain), each associated with four parameters—ϵ, *r*0, *n*, and *m* (giving the value of *gR*0)—for a total of 1200 parameters. For the HB and cooperativity interactions, we have a total of five parameters for $\epsilon_{\alpha/\beta}^{\mathrm{HB}}$, $\epsilon_{\alpha/\beta}^{\mathrm{coop}}$, and σ. Finally, for the ϕ/ψ potential, we have 2 ϵ values and 40 lower and higher limits $\phi_{\mathrm{low/high\_sc\_}i}/\psi_{\mathrm{low/high\_sc\_}i}$ (2 values per amino acid). In preliminary tests to maximize the speed and efficacy of the optimization, increasing the number of particles from 100 to 250 improves the best score by ~25%. Further increasing the number of particles from 250 to 500 and to 1000 leads to more modest improvements of, respectively, ~5 and 8%. Therefore, we select to use 500 particles for a maximum of 75 iterations or until the score is stable for 10 iterations, whichever comes first. A final optimization step is also tested with 750 particles, but no improvements to the final score are noted.

Iterated Optimization Procedure. To take into account the close relationship between the conformational search and the force field, we use an iterative optimization procedure that is based specifically on PEP-FOLD generated decoys, as described in Figure 3.

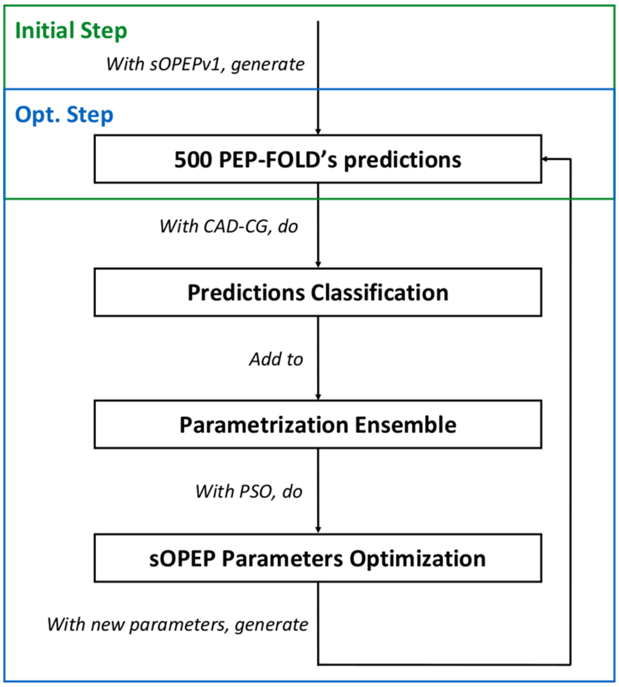

Initial Step
With sOPEPv1, generate
Opt. Step
500 PEP-FOLD's predictions
With CAD-CG, do
Predictions Classification
Add to
Parametrization Ensemble
With PSO, do
sOPEP Parameters Optimization
With new parameters, generate

Figure 3. Overview of the optimization protocol. In the initial step of the optimization (top green frame), the parametrization ensemble is composed of 500 PEP-FOLD predictions for each target. The sOPEP parameters are then optimized utilizing an iterative procedure (bottom blue frame), in which the parametrization ensemble is improved by adding newly generated PEP-FOLD predictions.

The optimization process for sOPEPv2 is presented in Figure 3. Each iteration involves the following three steps:

1. All 1287 (1200 non-bonded, five for hydrogen bonds and cooperativity and 82 for dihedral angles) sOPEP parameters are optimized. Ten independent optimizations (randomly generated SWARM positions and velocities) are launched. Only the optimized parameters leading to the best score are used for the next steps.
2. With the optimized parameters, new PEP-FOLD predictions are generated, as described in the Decoys Generation section on the protein sequences of the parametrization ensemble.
3. These newly generated decoys, that reflect the biases of the optimized potential, are added to the optimization ensemble. This approach allows the fitting procedure to include regions of the search space that could be available using the new parameters, mainly new wrong predictions with good energies.

In the full optimization cycle for sOPEP2, this whole procedure is repeated five times, leading to stable results. After the update of the library of fragments, from v1 to v2, we further optimize the bonded parameters, taking into consideration the difference in the local superposition of the new fragment. To reinforce this improvement, we use a more stringent score that requires a threshold of 0.6 for BC-WDC[53] in addition to the CAD-CG for the definition of native and near-native decoys. The BC-WDC is a nonlocal score based on the volume defined by the tetrahedrons formed by quadruplets of Cα and the geometric center of the protein. This added score helps with the identification of correct domain orientation, for which a local score such as the CAD score is less sensitive.[48]

COMPARISON WITH STATE-OF-THE-ART TECHNIQUES

In order to probe the quality of PEP-FOLD predictions on small peptides, we compare our results with three state-of-the-art machine learning techniques: the APPTest server,[19] the RaptorX server,[5−7] and AlphaFold2.[9]

APPTest uses constraints on distances and dihedral angles determined with a neural network in simulated annealing simulation for the prediction of small peptide structures.[19] It was recently tested against other software for the structural predictions of small peptides and showed a high rate of success.

RaptorX uses an ultradeep residual neural network (ResNet) on multiple sequence alignment to predict the probability distribution of interatomic distances and orientations. Then, a gradient-based minimization is used to build a 3-D model from the potential derived by ResNet. The RaptorX server had excellent results in CASP12 and CASP13.[5−7]

AlphaFold2[9] works by feeding a deep neural network with multiple sequence alignment features obtained from the UniRef90,[54] BFD,[55] and MGnify[56] databases. One particularity of AlphaFold2 is that is uses a novel attention-based deep learning architecture. This first step results in a sequence-specific probability distribution for interatomic distances and dihedral angles. The derived potential is then minimized via a gradient descent algorithm. AlphaFold2 showed tremendous results on the targets of CASP14.[9]

In order to compare the results with PEP-FOLD, the all-atom predictions of APPTest, the RaptorX server, and AlphaFold2 are converted into the sOPEP coarse-grained representation (the main chain stays unchanged) before comparison.

RESULTS

Updated Library of Fragments. In the original library (Lib1), a total of 182 four-residue fragments had been identified for the 27 structural alphabet (SA) letters. Seven of these fragments were associated with SA letters corresponding to α-helix (A, a, V, W), while 17 were associated with SA letters corresponding to β-sheet (L, N, M, T, X).

Using the new strategy and gradually decreasing the clustering cutoff from 2.0 Å^2 down to 1.5 Å^2, the number of separate clusters increases from 161 to 210. We select 1.9 Å^2 as a reasonable compromise between the number of clusters and the effectiveness of structure reconstruction.

This updated library, dubbed Lib2, contains 166 fragments, 7 of which are associated with SA letters corresponding to a α-helix conformation and 28 to β-sheets. This increase in the number of fragments associated with β-strands is a direct consequence of the change of strategy in fragment superimposition prior to clustering.

Optimization of the sOPEPv2 Parameters. To separate the impact of upgrading the fragment library from that of revising the force field, we first perform a full optimization cycle with five optimization steps on decoys generated using the original fragment library. Step to step improvements of the fraction of unsolved inequalities are presented in Figure 4.

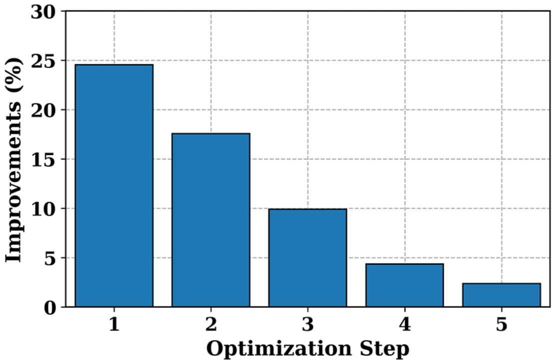


Figure 4. Improvements in the number of unsolved inequalities during the optimization protocol. For each optimization step (x-axis), the improvements are presented as the additional fraction in the number of unsolved inequalities over the previous optimization step.

Before optimization, with 500 decoys per target, 64.5% of the total number of inequalities are solved with the unoptimized second version of the potential, compared to 67.5% for the original potential. The optimization leads to an improvement of 24.6% in the number of unsolved inequalities. After five optimization steps, with 2500 decoys per target, associated with a 2.4% improvement in the number of unsolved inequalities, the reoptimized potential is able to solve 75.5% of the total number of inequalities.

To take into consideration the difference in the local superposition of the new fragments associated with Lib2, we only reoptimize the bonded potential (phi/psi parameters) while keeping the non-bonded parameters fixed to their previously optimized values. Additionally, we use a slightly more stringent classification score as described in the Materials and Methods section. The optimization is performed over 500 newly generated PEP-FOLD predictions using Lib2 and sOPEPv2 for each target in the parametrization set. This new optimization step leads to a 2.8% improvement in the number of unsolved inequalities. Since only a small improvement in the number of unsolved inequalities is observed, in addition to the fact that only the torsion angle parameters are optimized (82 parameters out of 1200 for the non-bonded parameters), we consider the optimization converged after this single step.

The optimization has a noticeable impact on the non-bonded energy terms. The exact values of the parameters are provided in Supporting Information Table S1. The optimization affects the $r0$ values only slightly, and its average variation during the optimization is of only −0.01 ± 0.79. Few large deviations occur for side chain−side chain interactions. The largest decrease occurs for the ASP−TRP pair (difference 1.91 Å) and the largest increase for HIS−GLN (1.82 Å) and MET−GLN (1.80 Å). For the ϵ parameters, we observe an average deviation of −0.09 ± 0.19 Å, with the largest decrease for CYS−CYS (−0.9 kcal/mol) and the largest increase for PHE−MET (0.66 kcal/mol). After optimization, the ϵ values for ASN−LEU and THR−LEU are at the maximum allowed value, indicating that these interactions stay mainly repulsive even with the attractive/repulsive formulation. Larger deviations are observed for $gR0$, which tends to decrease. On average, the difference is −0.20 ± 0.67 Å, with minimal and maximal deviations of −2.07 and 1.61 Å for ASN-THR and GLN-SER, respectively. In sOPEPv1, the repulsive strength at short distances is controlled by moving the asymptotic divergence around zero instead of directly changing the exponents which are still 12-6.[22] With this in mind, the most striking difference is observed for the exponents n and m. Their variation during the optimization is, on average, −4.96 ± 3.95 and −3.13 ± 2.36, respectively. For close to 160 interactions involving the side chain pseudoatom, n tends to be close to 4 (compared to the initial value of 12), while m ranges from only 0.6 to close to only 4. Although such exponents are less repulsive than the original 12-6, the fact that the asymptote stays at zero for a Mie potential[33] can still lead to sharper repulsion at shorter distances, as discussed in the following paragraph. Strikingly, it is mostly side chain−side chain interactions that are modified, whereas side chain−backbone interactions or backbone−backbone interactions tend to be less impacted.

Overall, we observe that, for the interactions that were already attractive/repulsive in sOPEPv1, two-thirds (62/93) are more permissive at short distances with a smaller value of $gR0$ in sOPEPv2. These interactions are mainly between apolar/apolar residues (such as ILE−ILE, ILE−LEU, or LEU−VAL), between pairs of aromatic residues (like PHE−PHE, PHE−TRP, or PHE−TYR) and between some pairs of oppositely charged residues (ASP−ARG, GLU−LYS), as shown in Figure S4. For their part, attractive/repulsive interactions that are less permissive in sOPEPv2 occur mainly between polar/polar residues (like ASN−GLN or GLN−TYR) or between polar/apolar residues (like MET−GLN or HIS−PRO), as shown in Figure S5. For the interactions using the repulsive formulation in sOPEPv1, shown in Figure S6, we observe that sOPEPv2 is less permissive at shorter distances; energies obtained for r values of 2 and 2.5 Å are higher using sOPEPv1 for only 30(/117) and 31(/117) pairs, respectively. These interactions involve mainly the small polar residue SER, with polar and charged residues (like ASN−SER or GLN−SER) and the small polar residue THR (like THR−TYR), as well as the positively charged residues LYS, ARG, and, depending on the pH, HIS (like ARG−ARG, LYS−ARG, or HIS−LYS).

Impact on Structure Prediction. In order to simplify the following discussion, we focus on the results associated with sOPEP1/Lib1 (Lib1/sOPEPv1) and sOPEP2/Lib2 (Lib2/sOPEPv2). Results for the other studied combinations (Lib1/sOPEPv2 and Lib2/sOPEPv1) are all presented in Tables S2, S3, and S4.

The Parametrization Ensemble. We first consider the impact of sOPEP reoptimization on the 25 protein target sequences retained in the parametrization ensemble (see Table S2 for details), using either Lib1 or Lib2. Results for this ensemble are presented as the left bar of each panel in Figure 5. Overall, one notes an improvement in model quality moving from sOPEP1 to sOPEP2 and an improvement in model quality moving from Lib1 to Lib2. This was expected due to the optimization protocol.

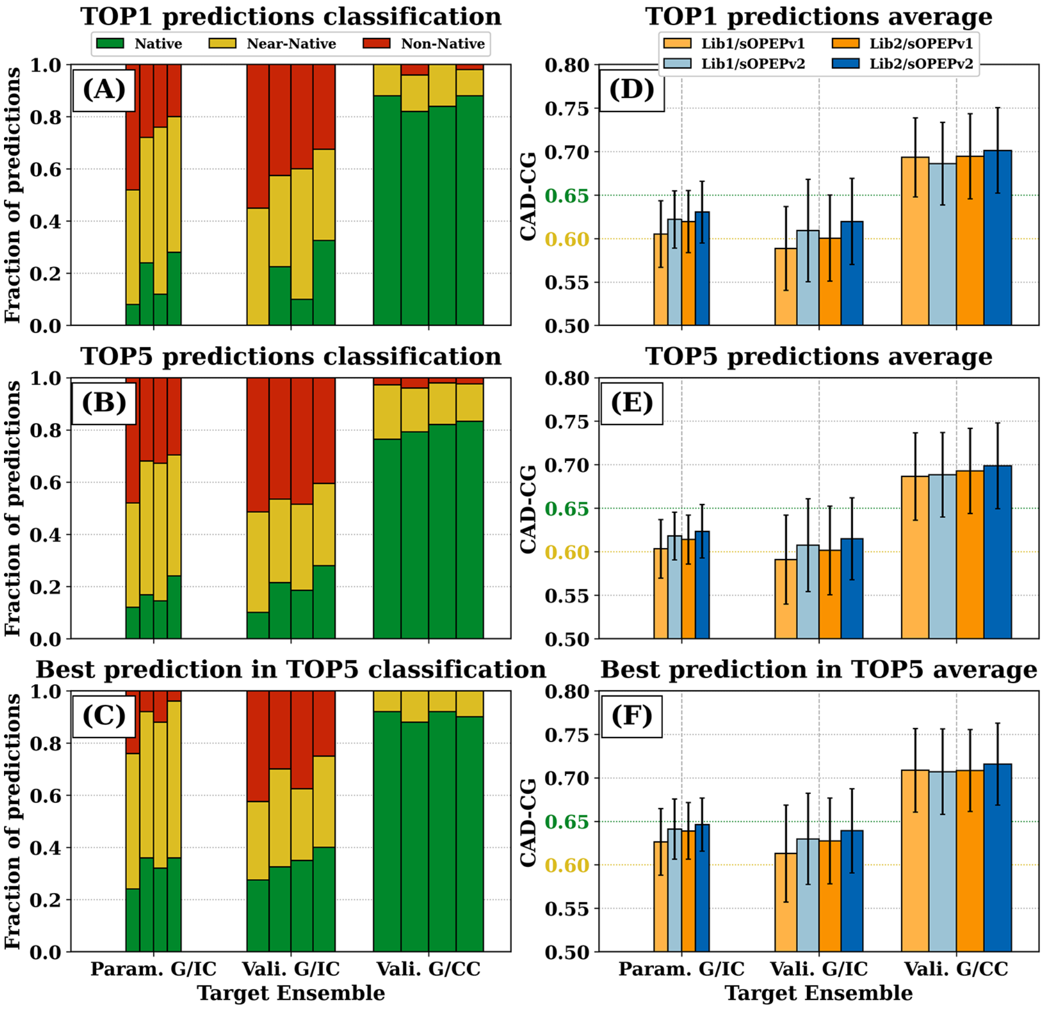


Figure 5. Classification of PEP-FOLD predictions and average CAD-CG of PEP-FOLD predictions. *x*-axis: name of the proteins' sets. *Param. G/IC*, *Vali. G/IC*, and *Vali. G/CC* refer to the parametrization, the validation G/IC, and the validation G/CC set containing 25, 39, and 48 proteins, respectively. Bar widths are proportional to the number of proteins of the set. Left side: classification of PEP-FOLD predictions. The native, near-native, and non-native classifications are shown in green, yellow, and red, respectively. For each set, the four columns represent from left to right the original library/original potential, the original library/reoptimized potential, the new library/original potential, and the new library/reoptimized potential, respectively. Panels A, B, and C: fraction of proteins per class considering the lowest energy model only (TOP1), the five lowest energy models (TOP5), and the best CAD-CG in the TOP5, respectively. Right side: average CAD-CG of 3D predictions. For each set, the four columns represent from left to right the original library/original potential (light orange), the original library/reoptimized potential (light blue), the new library/original potential (orange), and the new library/reoptimized potential (blue), respectively. The CAD-CG associated with the near-native and native classification are shown, respectively, in yellow and green (*y*-axis).

Considering the *best rank only* (TOP1) (Figure 5, panel A), out of 25 targets, the optimization increases the number of targets having native and near-native conformations from 2 (8%) and 11 (44%) up to 6 (24%) and 12 (48%), respectively, using Lib1. Using Lib2, this increases up to 7 (28%) and 13 (52%). The combined impact of sOPEPv2/Lib2 leads to a decrease by more than a factor of 2 in the number of targets with non-native predictions, from 12 (48%) to 5 (20%) targets. The average CAD-CG (Figure 5, panel D) increases from 0.605 ± 0.038, slightly above the near-native threshold (0.6), up to 0.630 ± 0.035, well into the near-native interval ([0.6, 0.65]).

Considering the *best in TOP5* prediction, presented in panel C of Figure 5, the same trends are observed. For sOPEP1/Lib1, six (24%) targets have a native prediction in the TOP5 and six (24%) have only non-native predictions in the TOP5. For sOPEP2/Lib2, nine (36%) targets have a native prediction in the TOP5 and only one (4%) target has only non-native predictions (1jjs; see Table S11 and the discussion). The average CAD-CG of the best in five predictions—panel F—slightly increases from 0.626 ± 0.038 up to 0.646 ± 0.030.

Finally, panels B and E present an analysis over the *five lowest energy* (TOP5) predictions. Using sOPEP1/Lib1, only 12% of the predictions in the TOP5 are native, while 48% are non-native. With sOPEP2/Lib2, the fraction of native predictions in the five lowest-energy structures for the parametrization targets increases to 24%, while the number of non-native predictions decreases to 30%. The average CAD-CG values increase from 0.603 ± 0.034, slightly above the near-native threshold (0.6), up to 0.623 ± 0.031, corresponding to the near-native definition.

In summary, for the parametrization ensemble, not only does the optimization make it possible to generate better models among the TOP5, but these are also of better quality on average. sOPEP2 outperforms sOPEP1, Lib2 outperforms Lib1, and there is an added value in combining sOPEP2 and Lib2.

The Validation (G/IC) Ensemble. As shown in Figure 5, the transferability of the improvements observed for the parametrization ensemble to the 40 protein targets of the validation (G/IC) ensemble is obvious (details regarding this ensemble are presented in Tables S3, S8, and S12).

Considering only the lowest energy model, panels A and D, using sOPEP1/Lib1, zero (0%) of the lowest-energy structures correspond to the native state of a target, with 18 (45%) classified as near-native and 22 (55%) as non-native. With sOPEP2/Lib2, 13 (33%) of the generated lowest energy structures correspond to a native state, 14 (35%) are near-native, and only 13 (33%) are non-native. This is a clear improvement. The average CAD-CG is 0.589 ± 0.048 for sOPEP1/Lib1, a score corresponding to the non-native classification. Using sOPEP2/Lib2, it increases up to 0.620 ± 0.049, well above the near-native threshold of 0.6. As before, sOPEP2/Lib2 gives the best results among the various combinations of force-field/library of fragments.

Considering the *best in TOP5* prediction, panels C and F, and using sOPEP1/Lib1, 11 (28%) sequences have at least one native structure among the prediction with the lowest five energies and 17 (43%) have only non-native among those. With sOPEP2/Lib2, 16 (40%) sequences have a predicted native structure among the TOP5 and only 10 (25%) have only non-native structures among the TOP5. The associated CAD-CG averages are of 0.613 ± 0.056, corresponding to the near-native class, and of 0.639 ± 0.048, closer to our native threshold of 0.65. The only target with a native prediction in the TOP5 with sOPEP1/Lib1 but only non-native prediction with sOPEP2/Lib2 is 5y22 (see the discussion).

Finally, considering the *five lowest energy* (TOP5) predictions, panels B and E, 10% of the predictions in the TOP5 are in the native class while 52% of the predictions in the TOP5 are non-native using sOPEP1/Lib1, whereas using sOPEP2/Lib2 almost triples the fraction of native predictions in the TOP5, to 28%, while the number of non-native predictions in the TOP5 decreases to 41%.

In summary, a clear improvement is observed for the targets that were incorrectly ranked using sOPEP1/Lib1.

The Validation (G/CC) Ensemble. We now look at the 50 protein targets of the validation (G/CC) ensemble that were correctly generated and ranked using sOPEP1/Lib1. Results for this target ensemble are presented in Tables S4, S9, and S13.

Overall, the results correspond to the expectation of a preserved performance. This is observed in terms of the *lowest energy* prediction, panels A and D of Figure 5, for which the lowest energy prediction is native for 44 (88%) protein targets and non-native for zero (0%) protein targets using sOPEP1/Lib1, while it is native for 44 (92%) targets and non-native for 1 structure (1rzs) but with a CAD-CG of 0.597, i.e., very close to near-native with a CAD score, using sOPEP2/Lib2. The average CAG-CG is 0.686 ± 0.050 with sOPEP1/Lib1 with a very slight improvement, at 0.699 ± 0.049, for sOPEP2/Lib2.

Considering the *best in TOP5* prediction, the results are very similar for all potential/library pairs with, for example, 0.709 ± 0.048 for sOPEP1/Lib1 and 0.716 ± 0.047 for sOPEP2/Lib2. The same is observed considering the *five lowest energy* (TOP5) predictions with 76% of the predictions in the TOP5 native and only 3% non-native for sOPEP1/Lib1, and 83 and 2% of the predictions are, respectively, native and non-native for sOPEP2/Lib2.

Overall, improving predictions of targets correctly predicted by sOPEP1/Lib1 does not lead to a deterioration for those correctly predicted: sOPEP2/Lib2 leads to similar or slightly better predictions than sOPEP1/Lib1 for almost all protein targets tested.

The NG Ensemble. Using our classification procedure of the targets, described in the Selection of Protein Targets section, we identified a series of 23 proteins for which sOPEP1/Lib1 is *unable* to generate a native (or near-native) prediction (ensemble NG), irrespective of its energetic classification. These proteins are mainly longer sequences (19 out of 23 are between 50 and 70 amino acids, i.e., longer than the original PEP-FOLD maximal size of 50) dominated by β-sheet secondary structures (10, 9, and 4 out of 23 are, respectively, β-protein, α/β-protein, and α-protein).

Modifications at the level of the library of fragments and the potential have no impact on the PEP-FOLD ability to generate native predictions for these targets; as for sOPEP1/Lib1, sOPEP2/Lib2 only generates non-native predictions. In order to better understand where the limitations lie and whether they are related to the discrimination by sOPEPv2, we compute the energy of the experimental structure, following a minimization. The energy of the experimental structure is then compared to that of the 3D predictions, as shown in Table 1. For 14 sequences out of 23, the energy of the native structure is lower than that of the best prediction and, for one additional sequence, the energy of the native structure is in the five lowest energy predictions. A more thorough analysis of the significance of these results is provided in the PEP-FOLD Limitations section.

Table 1. Energy Ranking for Targets in the NG Ensemble[a]

NG target	LowE energy (kcal/mol)	native energy (kcal/mol)	native rank
1gyf (α/β, 62)	−138	−184	0
1nd9 (α/β, 49)	−132	−98	501
1ne3 (β, 68)	−137	−149	0
1qxf (β, 66)	−170	−183	0
1vpu (α, 45)	−111	−40	501
1y2y (β, 68)	−149	−38	501
2cw1 (α/β, 65)	−180	−196	0
2do3 (β, 69)	−163	−213	0
2dy8 (α/β, 69)	−158	−173	0
2eqi (β, 69)	−127	−184	0
2gdl (α, 31)	−58	−21	501
2jrr (β, 67)	−147	−169	0
2jtv (α/β, 65)	−150	−182	0
2kaf (α/β, 67)	−176	−218	0
2l8d (β, 66)	−140	−208	0
2lhc (α, 56)	−161	−111	231.5
2lss (α/β, 70)	−162	−251	0
2m2l (α/β, 67)	−147	−138	4.5
2m4y (β, 56)	−128	−70	493.5
2m7o (α/β, 70)	−191	−205	0
2mck (α, 69)	−149	−127	202.5
2mdu (β, 29)	−76	−70	18.5
2xk0 (β, 69)	−158	−174	0

[a] LowE and Native Energy: energy of the lowest energy model generated using sOPEP2/Lib2 and experimental structure using sOPEPv2, respectively. Native Rank: ranking of the experimental structure compared to models generated using sOPEP2/Lib2. PEP-FOLD predictions are ordered from 1 to 500 in order of increasing energy; rank 0 means that the experimental structure has a lower energy than all predictions, while a rank of 501 means the experimental structure has a higher energy than all predictions.

Table 2. Proteins for Which the Classification of the Lowest Energy Prediction (TOP1) Is Improved[a]

Improved Proteins				
	sOPEPv1 - Lib1		sOPEPv2 - Lib2	
Target	LowE	Best in TOP5	LowE	Best in TOP5
1b03 (β, 18)	0.567 (0.280)	0.592 (0.496)	0.621 (0.553)	0.621 (0.553)
1i6c (β, 39)	0.589 (0.683)	0.645 (0.776)	0.608 (0.741)	0.608 (0.741)
1spw (β, 39)	0.590 (0.232)	0.652 (0.902)	0.683 (0.841)	0.693 (0.927)
1zv6 (α, 68)	0.545 (-0.159)	0.563 (-0.289)	0.625 (0.516)	0.629 (0.267)
2dt6 (α, 64)	0.630 (0.231)	0.636 (-0.146)	0.650 (0.604)	0.665 (0.718)
2fmr (α/β, 65)	0.562 (0.392)	0.571 (0.043)	0.680 (0.939)	0.680 (0.939)
2l92 (β, 50)	0.579 (0.016)	0.610 (0.845)	0.631 (0.314)	0.647 (0.502)
2lma (α, 22)	0.555 (0.101)	0.560 (0.031)	0.642 (0.051)	0.659 (0.047)
2mwf (β, 32)	0.575 (0.803)	0.625 (0.893)	0.671 (0.881)	0.675 (0.860)
2ysb ((β, 49)	0.642 (0.799)	0.659 (0.861)	0.658 (0.823)	0.697 (0.882)
1ify (α, 49)	0.638 (-0.363)	0.689 (0.912)	0.688 (0.850)	0.706 (0.764)
1k8b (α/β, 52)	0.499 (0.485)	0.526 (-0.023)	0.607 (0.660)	0.607 (0.660)
1pgb (α/β, 56)	0.551 (0.452)	0.580 (0.762)	0.664 (0.898)	0.664 (0.898)
1q1v (α, 70)	0.639 (0.613)	0.678 (0.873)	0.697 (0.904)	0.697 (0.904)
1zrj (α, 50)	0.639 (0.782)	0.682 (0.812)	0.687 (0.843)	0.695 (0.851)
1zwv (α, 58)	0.594 (0.335)	0.627 (0.286)	0.668 (0.877)	0.668 (0.877)
2a63 (α/β, 66)	0.643 (-0.540)	0.648 (0.651)	0.692 (0.945)	0.739 (0.955)
2bn6 (α, 33)	0.585 (0.027)	0.599 (0.007)	0.685 (0.865)	0.701 (0.878)
2coo (α, 70)	0.646 (0.539)	0.654 (0.191)	0.670 (0.761)	0.706 (0.935)
2jof (α, 20)	0.639 (0.294)	0.715 (0.492)	0.680 (0.156)	0.680 (0.156)
2jtm (β, 60)	0.540 (-0.008)	0.589 (-0.188)	0.602 (0.715)	0.669 (0.849)
2k2a (α, 70)	0.632 (-0.043)	0.634 (0.151)	0.685 (0.924)	0.704 (0.924)
2k57 (β, 61)	0.579 (0.304)	0.634 (0.899)	0.631 (0.204)	0.631 (0.204)
2kt2 (α/β, 69)	0.585 (0.764)	0.611 (0.488)	0.607 (0.789)	0.633 (0.624)
2l4m (α/β, 69)	0.546 (0.314)	0.553 (-0.432)	0.604 (0.195)	0.604 (0.195)
2msu (α, 20)	0.572 (0.128)	0.584 (0.170)	0.607 (0.036)	0.624 (-0.004)
2v0e (α/β, 55)	0.570 (0.070)	0.572 (0.244)	0.604 (0.849)	0.612 (0.849)
2ysh (β, 40)	0.646 (0.839)	0.646 (0.839)	0.654 (0.819)	0.665 (0.882)
2zaj (β, 49)	0.635 (0.784)	0.656 (0.824)	0.676 (0.826)	0.695 (0.807)
1wr3 (β, 36)	0.645 (0.837)	0.645 (0.837)	0.664 (0.808)	0.671 (0.882)
1wr4 (β, 36)	0.631 (0.859)	0.657 (0.884)	0.671 (0.879)	0.676 (0.837)
2cp9 (α, 64)	0.636 (-0.361)	0.685 (0.787)	0.689 (0.719)	0.693 (0.761)

[a] The notations are identical to those of Table 1. Columns 2 and 3 present the results for the lowest energy prediction and the best prediction in the TOP5 for sOPEPv1/Lib1, while columns 4 and 5 present the same results for sOPEPv2/Lib2. Each column presents the quality assessment in terms of CAD-CG and, in parentheses, BC-WDC. Color coding: CAD-CG scores corresponding to the native, near-native, and non-native classification are shown, respectively, in green, yellow, and red.

Table 3. Targets for Which the Classification of the Lowest Energy Prediction (TOP1) Is Deteriorated[a]

Deteriorated Proteins				
	sOPEPv1 - Lib1		sOPEPv2 - Lib2	
Target	LowE	Best in TOP5	LowE	Best in TOP5
1g2h (α, 61)	0.645 (0.447)	0.645 (0.447)	0.579 (0.140)	0.628 (0.428)
2l4j (β, 46)	0.607 (0.787)	0.618 (0.913)	0.593 (0.747)	0.635 (0.863)
1qpm (α/β, 69)	0.685 (0.925)	0.685 (0.925)	0.630 (0.300)	0.646 (0.301)
1rzs (α, 61)	0.631 (0.476)	0.638 (0.665)	0.597 (0.227)	0.625 (0.355)
2m8j (β, 48)	0.652 (0.776)	0.652 (0.776)	0.634 (0.794)	0.666
2wqg (α, 51)	0.703 (0.940)	0.705 (0.925)	0.624 (0.839)	0.646 (0.856)

[a] The notations are identical to those of Table 2.

Improvements. To better understand the underlying effects of revised PEP-FOLD, we focus on the protein targets, among all ensembles, that see their lowest energy prediction change classification when going from sOPEP1/Lib1 (Lib1/sOPEPv1) to sOPEP2/Lib2 (Lib2/sOPEPv2). All predictions using sOPEPv2/Lib2 are available in the Supporting Information. A total of 32 targets are shifted to a better class, as shown in Table 2: 14 go from non-native to near-native, 12 go from near-native to native, and 6 move directly from non-native to native. Only six target sequences move down in classification with sOPEP2/Lib2, with respect to the sOPEP1/Lib1 as shown in Table 3: three move from native to near-native and three from near-native to non-native. When analyzing the best structure in the five lowest energy (TOP5) predictions, we do however note the

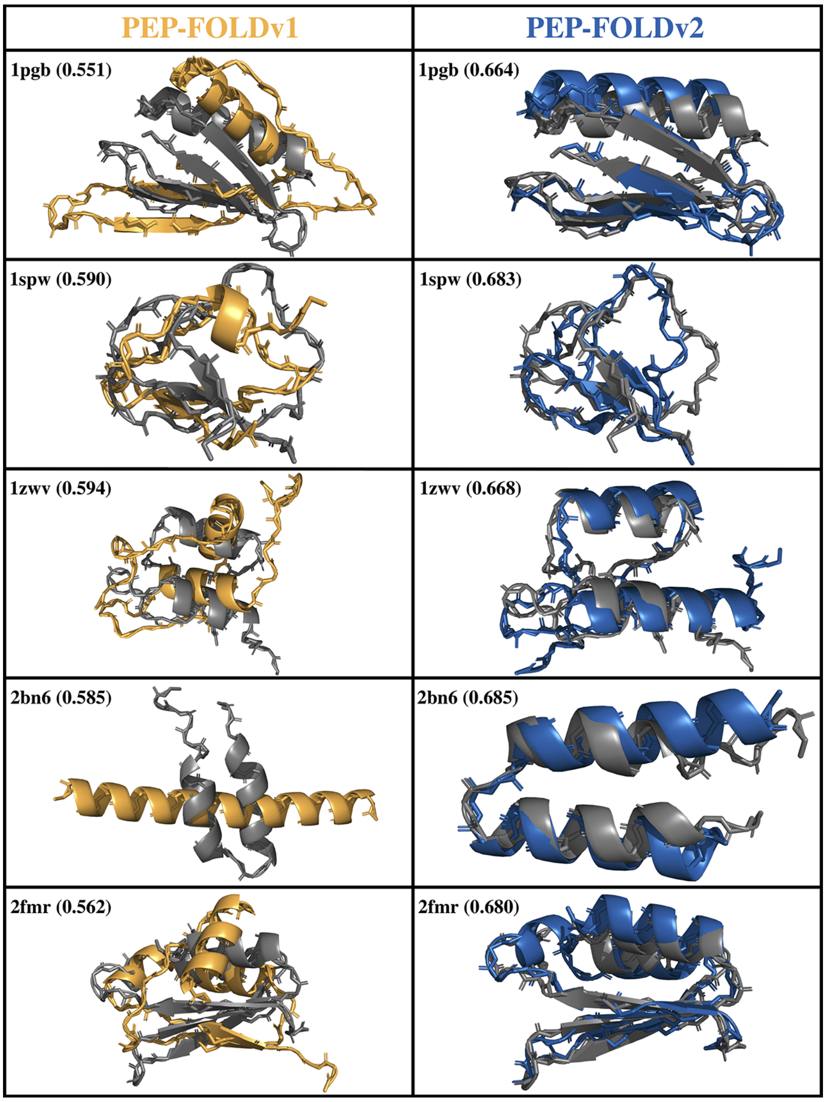


Figure 6. Lowest energy predictions for proteins going from non-native to native predictions. The left and right columns show the results for the original library with the original potential, in orange, and the new library with the new potential, in blue, respectively. The experimental structure is shown in gray. The structures are aligned on C_α of residues of the well-defined core, as presented in the Protein Data Bank. Pictures were generated using Pymol,[57] and secondary structure elements were determined using STRIDE.[58]

presence of at least a native prediction for 2m8j and at least a near-native prediction for the five others (1g2h, 2l4j, 1qpm, 1rzs, and 2wqg). The classification of the best prediction in the TOP5 deteriorates for only two out these six targets (1qpm and 2wqg).

In terms of secondary structure, out of a total of 60 α-targets, 13 are improved (1zv6, 2dt6, 2lma, 1ify, 1q1v, 1zrj, 1zwv, 2bn6, 2coo, 2jof, 2k2a, 2msu, and 2cp9) and 3 are deteriorated (1g2h, 1rzs, and 2wqg). Out of the 31 β-targets, 12 of them move up in classification (1b03, 1i6c, 1spw, 2l92, 2mwf, 2ysb, 2jtm, 2k57, 2ysh, 2zaj, 1wr3, and 1wr4) and 2 of them (2l4j and 2m8j) move down. Finally, out of the 21 α/β-targets, 7 of them see improved predictions (2fmr, 1k8b, 1pgb, 2a63, 2kt2, 2l4m, and 2v0e) with one of them (1qpm) deteriorating.

Overall, sOPEP2/Lib2 generates improved predictions across targets, irrespective of length: out of the 62 targets below 50 amino acids, 16 targets are improved (1b03, 1i6c, 1spw, 2l92, 2lma, 2mwf, 2ysb, 1ify, 1zrj, 2bn6, 2jof, 2msu, 2ysh, 2zaj, 1wr3, and 1wr4) and only 2 targets move down in classification (2l4j and 2m8j). Similar results are obtained for longer sequences with 15 targets (1zv6, 2dt6, 2fmr, 1k8b, 1pgb, 1q1v, 1zwv, 2a63, 2coo, 2jtm, 2k2a, 2k57, 2kt2, 2l4m, 2v0e, and 2cp9) out of the 50 between 50 and 70 amino acids moving to a higher classification and predictions for 4 targets (1g2h, 1qpm, 1rzs, and 2wqg) deteriorating.

The lowest energy prediction for both sOPEP1/Lib1 and sOPEP2/Lib2 for the six sequences that move from non-native to native class are presented in Figure 6. Improved predictions can be subtle, introducing a turn or perfecting the alignment, but they can also be fundamental, correcting badly predicted secondary structure, as shown by these examples.

For 1pgb, the sOPEP1/Lib1 prediction only identified two out of the four β-strands and the alignment of the α-helix and β-sheet is off. This prediction has a CAD-CG of 0.551 and a BC-WDC of 0.452. The prediction with sOPEP2/Lib2 correctly identifies the four β-strands and their alignment is fairly well reproduced, although a small deviation in the alignment of the α-helix remains. The new prediction has a CAD-CG of 0.664 and BC-WDC of 0.898. For 1spw, the sOPEP1/Lib1 prediction incorrectly predicts a small helix around residues 34−36, and while both small β-strands are present, their alignment is incorrect. For this prediction, the CAD-CG is 0.590 and the BC-WDC is 0.232. sOPEP2/Lib2 correctly reproduces the secondary structure elements and their alignment, leading to a CAD-CG of 0.683 and a BC-WDC of 0.841.

While almost all secondary structure motifs for 1zwv are correctly predicted with sOPEP1/Lib1, their alignment is completely wrong, leading to a CAD-CG of 0.594 and a BC-WDC of 0.335. It is correctly predicted with sOPEP2/Lib2, leading to a CAD-CG of 0.668 and a BC-WDC of 0.877.

sOPEP1/Lib1 overstabilized α-helical structures for 2bn6, predicting a single straight helix as its best structure leading to low CAD-CG (0.585) and BC-WDC (0.027) scores. sOPEP2/Lib2 correctly generates a turn, breaking the two α-helices, with the best predicted structure showing a CAD-CG of 0.685 and a BC-WDC of 0.865.

Finally, for 2fmr, sOPEP1/Lib1 has mostly the correct α-helical content, but the β-sheet content is seriously underestimated, leading to an overall alignment that is off, as shown by the low CAD-CG (0.562) and BC-WDC (0.392). sOPEP2/Lib2 correctly identifies 100% of the α-helical residues and 75% of the β-sheet residues, as well as the overall alignment with a CAD-CG of 0.680 and a BC-WD of 0.939.

The predictions' quality assessment is, in addition to the CAD-CG score, computed with the BC score[53] defined by the residues of the well-defined core (BC-WDC). The results are presented in Figure S2 for the lowest energy prediction (panel A), the five lowest energy predictions (panel B), and the best prediction in the five lowest energy predictions (panel C). The improvements observed in terms of CAD-CG are also compatible with the observed trend in terms of BC-WDC. For the lowest energy prediction, the BC-WDC goes from 0.347 ± 0.448 to 0.502 ± 0.380, from 0.330 ± 0.346 to 0.502 ± 0.380, and from 0.624 ± 0.400 to 0.707 ± 0.350 for the parametrization G/IC, the validation G/IC, and the validation G/CC ensemble, respectively. For the five lowest energy predictions, the BC-WDC goes from 0.327 ± 0.290 to 0.445 ± 0.345, from 0.327 ± 0.290 to 0.445 ± 0.345, and from 0.626 ± 0.361 to 0.721 ± 0.275 for the parametrization G/IC, the validation G/IC, and the validation G/CC ensemble, respectively. Finally, the BC-WDC of the best prediction in the five lowest energy predictions goes from 0.601 ± 0.274 to 0.652 ± 0.317, from 0.601 ± 0.274 to 0.652 ± 0.317, and from 0.810 ± 0.283 to 0.827 ± 0.193. These observations are compatible with those obtained for the CAD-CG.

Comparison. We now compare the predictions from sOPEP2/Lib2 to three state-of-the-art machine learning techniques: APPTest server,[19] the RaptorX server,[5−7] and AlphaFold2.[9]

APPTest is limited to sequences of 40 amino acids or less. For the very small peptides, 25 amino acids or less, the results with this tool are very good with an average CAD-CG of 0.698 ± 0.084. In the tested targets, the score is mainly dragged down by 1b03, for which the β-sheet secondary structure is not correctly reproduced (CAD-CG of 0.507), and 1s4j, which is predicted as almost fully extended (CAD-CG of 0.532). For larger targets between 26 and 40 amino acids, the prediction quality is decreased compared to the smaller targets with an average CAD-CG of 0.634 ± 0.075. As shown in panel A of Figure 7, this is mainly due to the β-sheet targets (CAD-GG score: 0.612 ± 0.077), as α-helical targets are very well predicted (CAD-GG score: 0.722 ± 0.046). This is caused by a slight shift in the hydrogen bond network between the β-strands, leading to incorrect prediction of side chain/side chain interactions, captured by the CAD score. Overall, panels A and B of Figure 7 show that APPTest does better on smaller, α-helical targets, while PEP-FOLD does better for larger, β-sheet targets, as it predicts the correct hydrogen bond network between the strands. Only two sequences from the NG ensemble are below the APPTest threshold of 40 amino acids: 2gdl and 2mdu. Similarly to PEP-FOLD using sOPEP2/Lib2, APPTest has trouble with these two targets with a CAD-CG of 0.534 and 0.592, respectively. When considering the best prediction in the TOP5, we note that the classification for six targets is improved, most notably for 2luf and 2mwf that go from the near-native to the native prediction (Table S13).

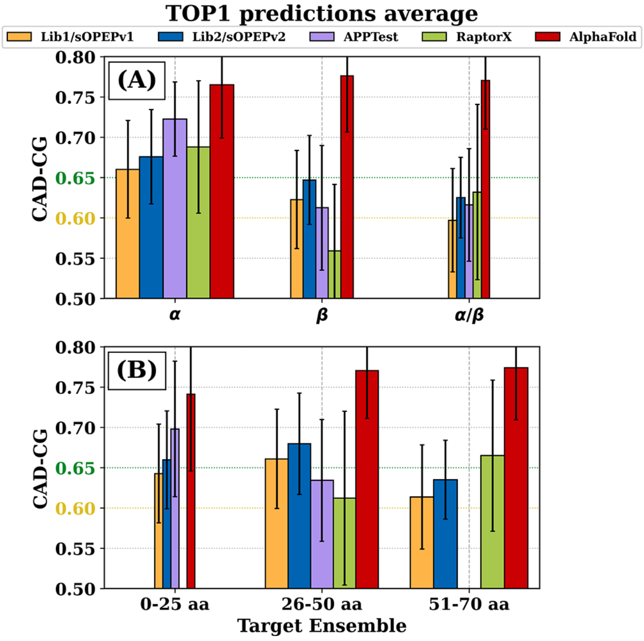


Figure 7. Average CAD-CG score for the TOP1 prediction using five prediction approaches. Panel A: results when targets are classified by structural class, with respectively 60, 32, and 21 proteins in the α, β, and α/β categories. Panel B: results when targets are classified by length, with respectively 17, 48, and 50 targets with less than 26 amino acids, between 26 and 50 amino acids, and between 51 and 70 amino acids. Of note, PEP-FOLD is usually limited to up to only 50 amino acids. The RaptorX server minimum accepted length is 26 amino acids. The APPTest does not consider sequences with more than 40 amino acids. The CAD-CG associated with the near-native and native classification are shown, respectively, in yellow and green (y-axis). Protein targets from the NG ensemble are excluded for this figure.

For the RaptorX server, a clear distinction of strengths and weaknesses, both in terms of secondary structure and length, emerges with respect to sOPEP2/Lib2. Panel A of Figure 7 shows that the predictions for α-targets of the RaptorX server are slightly better than those obtained using sOPEP2/Lib2 (average CAD-CG of 0.688 ± 0.082 vs 0.676 ± 0.059), while sOPEP2/Lib2 predictions are more reliable for β-targets (average CAD-CG of 0.647 ± 0.055 for sOPEP2/Lib2 vs 0.559 ± 0.082 for the RaptorX server). In terms of target length, panel B of Figure 7 shows that, for smaller targets, between 26 and 50 amino acids, sOPEP2/Lib2 gives better predictions than RaptorX (average CAD-CG of 0.680 ± 0.063 vs 0.612 ± 0.108), while RaptorX predictions are better for longer targets, between 50 and 70 amino acids (average CAD-CG of 0.635 ± 0.049 vs 0.665 ± 0.094). For the NG ensemble, RaptorX server predictions are also below its average over the other sequences: 9 of the 23 predicted targets in the NG ensemble are classified as native, with 1 near-native and 13 non-native. Focusing on the best prediction in the TOP5, we note only two targets for which the classification is improved: 2ysi (from non-native to native) and 2jrr (from non-native to near-native) (Table S13).

For its part, AlphaFold2 predictions are excellent for all secondary structure types, with an average CAD-CG of above 0.75 for α-, β-, and α/β-targets (Figure 7A) and for all protein lengths, with an average CAD-CG of above 0.70 for targets below 26 amino acids and above 0.75 for targets between 25 and 50 amino acids and between 50 and 70 amino acids (Figure 7B). For the smallest targets tested, the average of AlphaFold2 is only dragged down by 1b03, for which the hydrogen bond network of the β-sheet is shifted, leading to incorrect side chain/side chain interactions and a CAD-CG in the non-native realm (0.583) and 1s4j, which is predicted as a small α-helix as opposed to the native structure with two turns and no secondary structure elements. Finally, AlphaFold2 is the only method tested here that is able to correctly predict the overwhelming majority of the targets in the NG ensembles. Only four are incorrectly predicted by AlphaFold2, 1nd9, 1vpu, 1y2y, and 2gdl with a CAD-CG of, respectively, 0.558, 0.534, 0.494, and 0.595. For AlphaFold2, considering the best prediction inside the TOP5 does change the results with a single target, 2kya, that goes from the non-native to the near-native class (Table S12).

■ DISCUSSION

PEP-FOLD[20−22] is a quick, simplified, and successful approach to peptide structure prediction that is freely available as a Web server.[59] It is specialized in the structure prediction of small peptides up to 50 amino acids. In this study, we question PEP-FOLD applicability from 50 amino acids to 70 but still keep the focus on relatively short sequences, as these present a unique challenge compared to the prediction of larger proteins.[14]

In this study, we present improvements to two of the core aspects of PEP-FOLD: an updated library of fragments (Lib2) and a reoptimized version of sOPEP (sOPEPv2). sOPEPv2 introduces a new formulation for non-bonded interactions, and it is parametrized by using a self-consistent iterative process, using a philosophy similar to that developed for optimizing OPEP[41] and sOPEPv1.[22] The parametrization is designed to maximize the discrimination on an ensemble of decoys classified using simple criteria on the CAD score, with no further information on the distributions of interatomic distances. The improvement associated with the update of the library of fragments alone appears mainly limited to the targets below 50 amino acids, while the generalized formulation of the non-bonded interactions leads to improvements across all peptide lengths tested, as can be seen in Table S5.

Furthermore, despite its simplicity (discrete assembly, coarse-grained potential, etc.), the updated version of PEP-FOLD presented here shows improvements compared to state-of-the-art machine learning approaches such as APPTest and RaptorX.

Dependence on Target Size and Secondary Structure. sOPEP2/Lib2, with an updated fragment library and reoptimized potential, improves the accuracy of predicted structures for targets of 50 amino acids or less compared to sOPEP1/Lib1. The average CAD-CG goes from 0.656 to 0.675, placing most of these proteins in the native class, as we define it, while the average BC-WDC goes up from 0.519 to 0.596 (shown in Figure S3 and Table S5). In spite of the introduction of longer targets in the learning set, results for smaller targets do not deteriorate but improve in terms of both CAD-CG and BC-WDC.

As expected with a focus on longer sequences, sOPEP2/Lib2 delivers a pronounced improvement for peptides between 50 and 70 amino acids of the results in terms of CAD-CG, from 0.614 to 0.635, placing most of these proteins in the near-native class, with the average BC-WDC moving from 0.373 to 0.608. This improvement, that does not, for such sizes, bring PEP-FOLD to the level of performance of an approach such as AlphaFold, strongly suggests, however, that the generalized formulation proposed here is an effective direction. For such large sizes, other aspects of PEP-FOLD can limit the effective generation of accurate models, namely, the sequential assembly process in a discrete space.

Considering the secondary structure class of the targets (see Figure S7 and Table S6), α-proteins tend to be more often correctly predicted compared to β-proteins and α/β-proteins. This trend is observed for both sOPEP1/Lib1 and sOPEP2/Lib2, although with more important improvements in terms of tertiary structure, indicated by the CAD-CG and BC-WDC S7, for the latter compared to the former. The fact that α-helices are correctly predicted compared to other secondary structure classes is most likely a consequence of the PEP-FOLD assembly process rather than the force field itself.

Indeed, because α-helices are local in structure, correctly predicted structural alphabet letters associated with α-helices can therefore be immediately identified as favorable, during the amino acid by amino acid model generation process, whereas this is not possible for β-strands. More specifically, the lowest-energy structures predicted by sOPEP1/Lib1 and sOPEP2/Lib2 reproduce 93 and 96% of the experimental α-structures, respectively, for α-targets (shown in Figure S8).

By definition, β-targets contain no experimental α-helix. This feature is perfectly reproduced with PEP-FOLD: the folds contain no α-helix; hence, there is a 100% success rate for predictions. For α/β-targets, the predicted amount of correct experimental α-structure is also very high, with 94% for both sOPEP1/Lib1 and sOPEP2/Lib2, respectively.

The prediction of β-sheets is trickier. Indeed, slight deviations in the hydrogen bond network between strands can lead to completely new interactions between the side chains. Compared to other similarity scores, the CAD-CG score captures a small mismatch between β-strands very well.[48] With this in mind, we can clearly see improvements in the prediction of β-sheets with sOPEP2/Lib2. For the β-proteins, the average CAD-CG goes from 0.623 to 0.647 (shown in Figure S7) and the average fraction of β-sheet of the experimental structure correctly modeled increases from 0.817 to 0.871% (shown in Figure S8). For the α/β-proteins, the change is even more noticeable with the CAD-CG going from 0.597 to 0.625 and the average of reproduced β-sheet content from the experimental structure going from 45 to 78% for sOPEP1/Lib1 and sOPEP2/Lib2, respectively. Beyond the reoptimized potential, the use of the new library of fragments also contributes to this improvement, as the number of fragments associated with β-sheet letters of the SA goes from 17 to 28. This leads to more residues adopting the correct β-sheet secondary structure, as shown in Figure S8, with both sOPEPv2 and, to a lesser extent, sOPEPv1 when using Lib2.

PEP-FOLD Limitations. In spite of the overall prediction improvements realized with the revision of the sOPEP potential, we identify a few proteins for which PEP-FOLD is unable to make a correct prediction within the five lowest energy predictions.

One of these proteins, 1jjs (α, 50 amino acids), is in the parametrization ensemble. For this target, the use of sOPEPv2/Lib2 results in extending two of the three α-helices (from residue 5−13 and 19−31 and compared to 2−14 and 25−31) and in predicting a fourth helix between residues 45−49. Additionally, the relative positioning of the first helix with respect to the two others is off. A near-native structure (CAD-CG of 0.600 and BC-WDC of 0.566) is however present just outside the TOP5, at rank 8, as shown in Table 4.

Table 4. Ranking of Incorrectly Predicted Targets[a]

	native	first non-native prediction	
target	rank	CAD-CG (BC-WDC)	rank
1s4j (coil, 13)	501	0.610 (−0.525)	122
5y22 (α, 22)	236.5	0.724 (0.977)	61
2kya (α, 34)	2.5	0.611 (0.017)	119
1k91 (β, 37)	0	0.601 (−0.059)	10
1ed7 (β, 45)	0	0.602 (0.769)	102
2m6o (β, 48)	4.5	0.621 (0.877)	17
1jjs (α, 50)	158.5	0.600 (0.566)	8
1n87 (α/β, 56)	3.5	0.602 (0.871)	10
2mdj (β, 56)	217.5	0.613 (−0.445)	122
2mi6 (β, 62)	0	0.612 (0.693)	13
1f0z (α/β, 66)	0	0.600 (0.867)	77

[a] For each target incorrectly predicted within the five lowest energies, the ranking of the experimental structure and the quality assessment in terms of CAD-CG(BC-WDC) and ranking of the first non-native prediction are presented, respectively, for columns 2−4. PEP-FOLD predictions are ordered from 1 to 500 in order of increasing energy; rank 0 means that the experimental structure has a lower energy than all predictions, while a rank of 501 means the experimental structure has a higher energy than all predictions.

In the validation G/IC ensemble, PEP-FOLD is unable to identify a near-native or native prediction among the five lowest energy structures for 10 out of the 40 targets, five of which are β-protein (1ed7, 1k91, 2m6o, 2mdj, and 2mi6), two are α/β-protein (1f0z, 1n87), two are α-protein (2kya, 5y22), and one has no secondary structure elements (1s4j). To identify the source of this difficulty, we compute the energy of the experimental structure with the reoptimized potential after relaxation and compare its ranking with sOPEP2/Lib2 predictions. For 4 of these 10 sequences (1ed7, 1f0z, 1k91, and 2mi6), the experimental structure ranks before the best prediction, and for 2 others (1n87 and 2m6o), the experimental structure ranks in the five lowest energy predictions, as presented in Table 4. For 1k91 and 1n87, a near-native prediction is present just outside the TOP5 at rank 10 for both (see Table 4).

We now have a look at the remaining sequences. For 1s4j, sOPEPv2/Lib2 predicts a small β-hairpin, similarly to sOPEPv1/Lib1 instead of the two turns and no secondary structure elements of the native structure. For 2kya, the SVM predicts the position of the α-helix around residues 24−33, while the experimental α-helix is around residues 12−28 and it also predicts a nonexisting β-strand around residues 11−20. Finally, for 5y22, sOPEPv2/Lib2 predicts that the second half of the α-helix, between residues 3 and 15 in the experimental structure, instead forms a small β-hairpin. With sOPEPv1/Lib1, the second half of the experimental α-helix is instead mainly disordered, except for the fourth prediction which correctly predicts the correct α-helix (see Table S12). For sOPEPv2/Lib2, the correctly predicted structure is not present in the five lowest energy predictions. 5y22 is the only case for which the results in terms of the best prediction in the TOP5 is deteriorated by using sOPEPv2/Lib2 compared to sOPEPv1/Lib1.

Similarly to what we observed for 2kya, we find some limitation for the SVM on the targets from the Not Generated (NG) ensemble, as we identify multiple incorrect secondary structure predictions made by the SVM. For example, in 2gdl (α, 31 aa), the SVM predicts the localization of the α-helix around residues 21−29, instead of around residues 5−18 in the experimental structure. For 1vpu (α, 45 aa), the experimental helix around residues 23−28 is shifted in the SVM predictions to around residues 26−35, in addition to the helix between residues 39 and 43 not being identified by the SVM. Finally, for 2lhc (α, 56 aa), two of the three experimental α-helices, between residues 9 and 14 and residues 39 and 51, are identified as β-sheet by the SVM.

Together, these results show that the updated library and potentials are able to identify correctly the native structure of most of the problematic sequences. This confirms that the simplified representation adopted here, both in terms of structure, including the coarse-graining of the side chain, and interactions, manages to capture the essential features responsible for folding.

The results also show that, for most sequences, the SVM approach to structure prediction excels at generating the relevant structures, both secondary and tertiary, that can then be classified using the energy model. With the current structural alphabet, however, this approach can fail for a relatively small subset of sequences, particularly for sequences where the tertiary structure is essential to enforce the secondary structure. While a more detailed analysis of these cases could allow us to better understand the delicate balance between these two levels of organization for some sequences, the SVM remains a powerful tool for exploring the structures of peptidic sequences.

■ CONCLUSION

Small peptides can play an important role in the development of novel therapeutic approaches[10,11] and represent a unique challenge compared to larger proteins. Indeed, the same amino acid sequence can adopt very different structures whether it is a peptide or a fragment of a larger protein.[4,14] In this work, we present improvements to the popular, freely available online,[59] PEP-FOLD method for small peptide structure predictions and extend its application from sequences of 50 amino acids to 70.

These improvements focus on two aspects of PEP-FOLD. First, using a new superimposition and clusterization scheme, we update the PEP-FOLD library of fragments associated with each letter of the structural alphabet (SA). This leads to an overall decrease in the total number of fragments, from 182 to 166 but with a larger number of fragments associated with β-sheet letters (from 17 to 28). Second, the parameters of the sOPEP force field, used in PEP-FOLD for prediction classification during (and after) greedy assembly of the fragments, are reoptimized using an iterative self-consistent process. sOPEP2/Lib2 leads to improved predicted structures for targets found problematic with sOPEP1/Lib1, both in terms of the lowest energy and five lowest energy predictions, while maintaining the quality for targets already correctly predicted by sOPEP1/Lib1. While PEP-FOLD is the only approach of this study not going to the all-atom level and using a discrete space search, sOPEP2/Lib2 predictions compare well with other state-of-the-art protein/peptide structure prediction techniques—the recently developed APPTest[19] and RaptorX[5−7]—but are behind the recently proposed AlphaFold2.[9]

Therefore, with its overall high reliability for shorter sequences (50 amino acids and less), the original approach retained by PEP-FOLD, including the use of a structural alphabet, of a sequential growth algorithm, and of a rich coarse-grained potential optimized using a very general classification scheme, this improved version of PEP-FOLD offers a solid prediction tool that can provide physical insights into the folding process. The analysis presented with this revised parametrization shows, in particular, the importance of better understanding the link between tertiary and secondary structure, particularly for these smaller fragments, but also the strength of the local approach retained here. As updated, sOPEP2/Lib2 remains, therefore, an important tool for structure prediction of short sequences. In addition, the quality of the structure prediction provides a strong support for the simplified sOPEP2 potential, developed here, that could serve as a solid basis for dynamical studies, unreachable by purely IA folding techniques.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.1c01293.

Detailed comparison for all tested targets for all tested prediction techniques (ZIP)

PDB predictions (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

Normand Mousseau − *Départment de Physique, Université de Montréal, Montréal, QC H3C 3J7, Canada*; Email: normand.mousseau@umontreal.ca

Pierre Tuffery − *Université de Paris, INSERM U1133, CNRS UMR 8251, F-75205 Paris, France*; orcid.org/0000-0003-1033-9895; Email: pierre.tuffery@u-paris.fr

Author

Vincent Binette − *Départment de Physique, Université de Montréal, Montréal, QC H3C 3J7, Canada*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.1c01293

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

V.B. is grateful to the Fonds de recherche du Québec − Nature et technologie through a postgraduate fellowship and to MITACS for a travel grant. This program is partially supported through a Discovery Grant of the Natural Sciences and Engineering Research Council of Canada to Normand Mousseau. This research was made possible through generous computer allocations from Calcul Québec (www.calculquebec.ca) and Compute Canada (www.computecanada.ca).

■ REFERENCES

(1) Dill, K. A.; MacCallum, J. L. The protein-folding problem, 50 years on. *science* **2012**, *338*, 1042−1046.

(2) Goodwin, S.; McPherson, J. D.; McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **2016**, *17*, 333.

(3) Westbrook, J. D.; Burley, S. K. How structural biologists and the Protein Data Bank contributed to recent FDA new drug approvals. *Structure* **2019**, *27*, 211−217.

(4) Kryshtafovych, A.; Schwede, T.; Topf, M.; Fidelis, K.; Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Struct., Funct., Bioinf.* **2019**, *87*, 1011−1020.

(5) Xu, J.; Mcpartlon, M.; Li, J. Improved protein structure prediction by deep learning irrespective of co-evolution information. *Nature Machine Intelligence* **2021**, *3*, 601−609.

(6) Xu, J. Distance-based protein folding powered by deep learning. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 16856−16865.

(7) Wang, S.; Sun, S.; Li, Z.; Zhang, R.; Xu, J. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS computational biology* **2017**, *13*, No. e1005324.

(8) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871.

(9) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W.; Bridgland, A.; et al. Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, *577*, 706−710.

(10) Chen, C. H.; Lu, T. K. Development and challenges of antimicrobial peptides for therapeutic applications. *Antibiotics* **2020**, *9*, 24.

(11) Seo, M.-D.; Won, H.-S.; Kim, J.-H.; Mishig-Ochir, T.; Lee, B.-J. Antimicrobial peptides for therapeutic applications: a review. *Molecules* **2012**, *17*, 12276−12286.

(12) Willyard, C. The drug-resistant bacteria that pose the greatest health threats. *Nature News* **2017**, *543*, 15.

(13) Bruzzoni-Giovanelli, H.; Alezra, V.; Wolff, N.; Dong, C.-Z.; Tuffery, P.; Rebollo, A. Interfering peptides targeting proteinprotein interactions: the next generation of drugs? *Drug Discovery Today* **2018**, *23*, 272−285.

(14) Singh, H.; Singh, S.; Raghava, G. P. S. Peptide secondary structure prediction using evolutionary information. 2019, 558791v1. biorxiv.org e-Print archive. https://www.biorxiv.org/content/10.1101/558791v1.

(15) Cao, X.; He, W.; Chen, Z.; Li, Y.; Wang, K.; Zhang, H.; Wei, L.; Cui, L.; Su, R.; Wei, L. PSSP-MVIRT: peptide secondary structure prediction based on a multi-view deep learning architecture. *Briefings in Bioinformatics* **2021**, *22*, bbab203.

(16) Kaur, H.; Garg, A.; Raghava, G. P. S. PEPstr: a de novo method for tertiary structure prediction of small bioactive peptides. *Protein and peptide letters* **2007**, *14*, 626−631.

(17) Singh, S.; Singh, H.; Tuknait, A.; Chaudhary, K.; Singh, B.; Kumaran, S.; Raghava, G. P. PEPstrMOD: structure prediction of peptides containing natural, non-natural and modified residues. *Biology direct* **2015**, *10*, 73.

(18) Ru, X.; Lin, Z. Genetic Algorithm Embedded with a Search Space Dimension Reduction Scheme for Efficient Peptide Structure Predictions. *J. Phys. Chem. B* **2021**, *125*, 3824−3829.

(19) Timmons, P. B.; Hewage, C. M. APPTEST is a novel protocol for the automatic prediction of peptide tertiary structures. *Briefings in Bioinformatics* **2021**, *22*, bbab308.

(20) Lamiable, A.; Thévenet, P.; Rey, J.; Vavrusa, M.; Derreumaux, P.; Tuffery, P. PEP-FOLD3: faster de novo structure prediction for linear peptides in solution and in complex. *Nucleic acids research* **2016**, *44*, W449−W454.

(21) Shen, Y.; Maupetit, J.; Derreumaux, P.; Tuffery, P. Improved PEP-FOLD approach for peptide and miniprotein structure prediction. *J. Chem. Theory Comput.* **2014**, *10*, 4745−4758.

(22) Maupetit, J.; Derreumaux, P.; Tuffery, P. A fast method for large-scale De Novo peptide and miniprotein structure prediction. *Journal of computational chemistry* **2010**, *31*, 726−738.

(23) Maupetit, J.; Derreumaux, P.; Tuffery, P. PEP-FOLD: an online resource for de novo peptide structure prediction. *Nucleic Acids Res.* **2009**, *37*, W498−W503.

(24) Maffucci, I.; Contini, A. In Silico Drug Repurposing for SARS-CoV-2 Main Proteinase and Spike Proteins. *J. Proteome Res.* **2020**, *19*, 4637−4648.

(25) Singh, A.; Thakur, M.; Sharma, L. K.; Chandra, K. Designing a multi-epitope peptide based vaccine against SARS-CoV-2. *Sci. Rep.* **2020**, *10*, 16219.

(26) Tahir ul Qamar, M.; Rehman, A.; Tusleem, K.; Ashfaq, U. A.; Qasim, M.; Zhu, X.; Fatima, I.; Shahid, F.; Chen, L.-L. Designing of a next generation multiepitope based vaccine (MEV) against SARS-COV-2: Immunoinformatics and in silico approaches. *PloS one* **2020**, *15*, No. e0244176.

(27) Blaszczyk, M.; Jamroz, M.; Kmiecik, S.; Kolinski, A. CABS-fold: server for the de novo and consensus-based prediction of protein structure. *Nucleic Acids Res.* **2013**, *41*, W406−W411.

(28) Sterpone, F.; Melchionna, S.; Tuffery, P.; Pasquali, S.; Mousseau, N.; Cragnolini, T.; Chebaro, Y.; St-Pierre, J.-F.; Kalimeri, M.; Barducci, A.; et al. The OPEP protein model: from single molecules, amyloid formation, crowding and hydrodynamics to DNA/RNA systems. *Chem. Soc. Rev.* **2014**, *43*, 4871−4893.

(29) Nasica-Labouze, J.; Nguyen, P. H.; Sterpone, F.; Berthoumieu, O.; Buchete, N.-V.; Cote, S.; De Simone, A.; Doig, A. J.; Faller, P.; Garcia, A.; et al. Amyloid *β* protein and Alzheimers disease: When computer simulations complement experimental studies. *Chem. Rev.* **2015**, *115*, 3518−3563.

(30) Pasquali, S.; Derreumaux, P. HiRE-RNA: a high resolution coarse-grained energy model for RNA. *J. Phys. Chem. B* **2010**, *114*, 11957−11966.

(31) Kynast, P.; Derreumaux, P.; Strodel, B. Evaluation of the coarse-grained OPEP force field for protein-protein docking. *BMC biophysics* **2016**, *9*, 4.

(32) Levitt, M. Protein folding by restrained energy minimization and molecular dynamics. *Journal of molecular biology* **1983**, *170*, 723−764.

(33) Mie, G. Zur kinetischen Theorie der einatomigen Körper. *Annalen der Physik* **1903**, *316*, 657−697.

(34) Janeček, J.; Said-Aizpuru, O.; Paricaud, P. Long Range Corrections for Inhomogeneous Simulations of Mie nm Potential. *J. Chem. Theory Comput.* **2017**, *13*, 4482−4491.

(35) Thévenet, P.; Shen, Y.; Maupetit, J.; Guyon, F.; Derreumaux, P.; Tufféry, P. PEP-FOLD: an updated de novo structure prediction server for both linear and disulfide bonded cyclic peptides. *Nucleic Acids Res.* **2012**, *40*, W288−W293.

(36) Zhang, Y.; Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins: Struct., Funct., Bioinf.* **2004**, *57*, 702−710.

(37) Camproux, A.-C.; Gautier, R.; Tuffery, P. A hidden markov model derived structural alphabet for proteins. *Journal of molecular biology* **2004**, *339*, 591−605.

(38) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **1997**, *25*, 3389−3402.

(39) Tuffery, P.; Guyon, F.; Derreumaux, P. Improved greedy algorithm for protein structure reconstruction. *Journal of computational chemistry* **2005**, *26*, 506−513.

(40) Derreumaux, P. From polypeptide sequences to structures using Monte Carlo simulations and an optimized potential. *J. Chem. Phys.* **1999**, *111*, 2301−2310.

(41) Maupetit, J.; Tuffery, P.; Derreumaux, P. A coarse-grained protein force field for folding and structure prediction. *Proteins: Struct., Funct., Bioinf.* **2007**, *69*, 394−408.

(42) Micheletti, C.; Seno, F.; Maritan, A.; Banavar, J. An optimal procedure to extract interaction potentials for protein folding. *Computational materials science* **2001**, *20*, 305−310.

(43) Fain, B.; Levitt, M. Funnel sculpting for in silico assembly of secondary structure elements of proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 10700−10705.

(44) Liwo, A.; Sieradzan, A. K.; Lipska, A. G.; Czaplewski, C.; Joung, I.; Żmudzińska, W.; Ołdziej, S. A general method for the derivation of the functional forms of the effective energy terms in coarse-grained energy functions of polymers. III. Determination of scale-consistent backbone-local and correlation potentials in the UNRES force field and force-field calibration and validation. *J. Chem. Phys.* **2019**, *150* (15), 155104.

(45) Qiu, J.; Elber, R. Atomically detailed potentials to recognize native and approximate protein structures. *Proteins* **2005**, *61*, 44−55.

(46) Liwo, A.; Khalili, M.; Czaplewski, C.; Kalinowski, S.; Ołdziej, S.; Wachucik, K.; Scheraga, H. A. Modification and optimization of the united-residue (UNRES) potential energy function for canonical simulations. I. Temperature dependence of the effective energy function and tests of the optimization method with single training proteins. *J. Phys. Chem. B* **2007**, *111*, 260−285.

(47) Olechnovič, K.; Kulberkytė, E.; Venclovas, Č. CAD-score: a new contact area difference-based function for evaluation of protein structural models. *Proteins: Struct., Funct., Bioinf.* **2013**, *81*, 149−162.

(48) Olechnovic, K.; Monastyrskyy, B.; Kryshtafovych, A.; Venclovas, C.; Valencia, A. Comparative analysis of methods for evaluation of protein models against native structures. *Bioinformatics* **2019**, *35*, 937−944.

(49) Chen, V. B.; Arendall, W. B.; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography* **2010**, *66*, 12−21.

(50) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. How robust are protein folding simulations with respect to force field parameterization? *Biophysical journal* **2011**, *100*, L47−L49.

(51) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1*, 19−25.

(52) Eberhart, R.; Kennedy, J. Particle swarm optimization. *Proceedings of the IEEE international conference on neural networks*; 1995; pp 1942−1948.

(53) Guyon, F.; Tuffery, P. Fast protein fragment similarity scoring using a binet−cauchy kernel. *Bioinformatics* **2014**, *30*, 784−791.

(54) The UniProt Consortium.. UniProt: a worldwide hub of protein knowledge. *Nucleic acids research* **2019**, *47*, D506−D515.

(55) Steinegger, M.; Mirdita, M.; Söding, J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods* **2019**, *16*, 603−606.

(56) Mitchell, A. L.; Almeida, A.; Beracochea, M.; Boland, M.; Burgin, J.; Cochrane, G.; Crusoe, M. R.; Kale, V.; Potter, S. C.; Richardson, L. J.; et al. MGnify: the microbiome analysis resource in 2020. *Nucleic acids research* **2020**, *48*, D570−D578.

(57) Schrödinger, L., DeLano, W. *PyMOL*; 2020. Retrieved from http://www.pymol.org/pymol.

(58) Frishman, D.; Argos, P. Knowledge-based protein secondary structure assignment. *Proteins: Struct., Funct., Bioinf.* **1995**, *23*, 566−579.

(59) Néron, B.; Ménager, H.; Maufrais, C.; Joly, N.; Maupetit, J.; Letort, S.; Carrere, S.; Tuffery, P.; Letondal, C. Mobyle: a new full web bioinformatics framework. *Bioinformatics* **2009**, *25*, 3005−3011.