JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

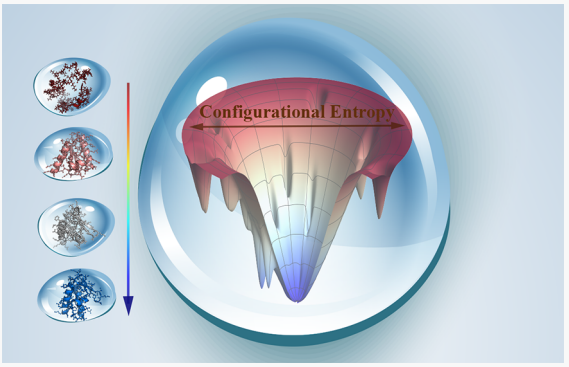
Protein Structure Prediction in CASP13 Using AWSEM-Suite

Shikai Jin,[#] Mingchen Chen,[#] Xun Chen,[#] Carlos Bueno, Wei Lu, Nicholas P. Schafer, Xingcheng Lin, José N. Onuchic, and Peter G. Wolynes*

Cite This: *J. Chem. Theory Comput.* 2020, 16, 3977−3988 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Recently several techniques have emerged that significantly enhance the quality of predictions of protein tertiary structures. In this study, we describe the performance of AWSEM-Suite, an algorithm that incorporates template-based modeling and coevolutionary restraints with a realistic coarse-grained force field, AWSEM. With its roots in neural networks, AWSEM contains both physical and bioinformatical energies that have been optimized using energy landscape theory. AWSEM-Suite participated in CASP13 as a server predictor and generated reliable predictions for most targets. AWSEM-Suite ranked eighth in both the free-modeling category and the hard-to-model category and in one case provided the best submitted prediction. Here we critically discuss the prediction performance of AWSEM-Suite using several examples from different categories in CASP13. Structure prediction tests on these selected targets, two of them being hard-to-model targets, show that AWSEM-Suite can achieve high-resolution structure prediction after incorporating both template guidances and coevolutionary restraints even when homology is weak. For targets with reliable templates (template-easy category), introducing coevolutionary restraints sometimes damages the overall quality of the predictions. Free energy profile analyses demonstrate, however, that the incorporations of both of these evolutionarily informed terms effectively increase the funneling of the landscape toward native-like structures while still allowing sufficient flexibility to correct for discrepancies between the correct target structure and the provided guidance. In contrast to other predictors that are exclusively oriented toward structure prediction, the connection of AWSEM-Suite to a statistical mechanical basis and affiliated molecular dynamics and importance sampling simulations makes it suitable for functional explorations.



1. INTRODUCTION

The structures of proteins largely determine protein function *in vivo*. Understanding the structures of proteins enables many applications in protein engineering and in the pharmaceutical field.[1] Due to the limitation and costs of experimental structure determination, computational structure prediction can play a significant role in practical work. Based on Anfinsen's studies it has long been held that, given the sequence of amino acids for a specific protein, we should be able to predict the native folded structure without any additional experimental input.[2] Our modern understanding of Anfinsen's experiment, i.e., how proteins fold to their stable states, has been clarified through the notion that folding energy landscapes are funneled and that this funneling is the result of evolution.[3−5] *Ab initio* protein structure prediction using only sequence information starting from extended peptide chains to simulate the folding process has been verified to be feasible for the smallest proteins, but such fully atomistic approaches still entail significant computational expense.[6]

CASP (critical assessment of structure prediction) is an experiment that aims to assess the state of the art in predicting protein structure every two years. Although the quality of predicted structures without evolutionary guidance has been increasing with each generation of CASP, the accuracy of the *ab initio* protein structure prediction is not satisfactory for many practical purposes.[7] Template-based modeling is the most reliable and time-efficient way to predict protein structure if a high sequence identity homologue can be found.[8] If two sequences are sufficiently similar, it can be inferred that they must have descended from a common ancestor and must therefore share similar tertiary structures. The template structures that are used for modeling are usually found by sequence−sequence comparison methods such as PSI-BLAST or by sequence−structure threading methods that, even in the absence of an obvious template, sometimes reveal more distant evolutionary relationships among related sequences.[9−12] Using sequence coevolutionary information is also helpful.[13] Strong covariance between a pair of residues suggests these residues are in contact using this idea. Coevolutionary-based

Received: February 24, 2020
Published: May 12, 2020

https://dx.doi.org/10.1021/acs.jctc.0c00188
J. Chem. Theory Comput. 2020, 16, 3977−3988

predictions have been used with success to predict not only the three-dimensional structures of proteins but also multiprotein complexes.[14] Due to the conservative nature of protein evolution, many diverse homologous proteins share the same global fold. Early methods of coevolutionary analysis used mean-field direct-coupling analysis (mfDCA), and Gremlin uses a pseudolikelihood approach (plmDCA) to predict the contact pairs.[15,16] Global statistical methods have more recently attempted to build models using whole protein sequence alignment rather than simply analyzing pairs of sites separately, which can improve the accuracy of contact prediction.[17]

Neural network ideas have been used to predict protein tertiary structure for quite a while.[18−20] Since the last CASP held in 2016, machine learning based algorithms have developed rapidly and performed very well in the free-modeling area. Nonetheless, the physics based methods also have displayed their power in some cases. Here we describe a prediction method that we have employed that combines elements of all these successful ideas. This method, the Associative memory, Water-mediated, Structure and Energy Model (AWSEM), is optimized based on the principles of the energy landscape theory of protein folding which provide a quantitative machine learning strategy.[5,18] AWSEM has witnessed success in a comprehensive set of applications including monomer *ab initio* predictions[21] and prediction of protein−protein associations,[22] and by joining it with coarse-grained DNA models.[23] By coupling to coarse-grained DNA models, AWSEM has been used to study nucleosomes.[24] AWSEM-Suite is presently being used in many functional investigations ranging from exploring the mechanisms of chromosome extrusion[25] to the formation of memory through actin network reorganization.[26] AWSEM-Suite is a realistic coarse-grained force field that employs transferable tertiary interactions along with homology modeling and knowledge-based local-in-sequence interaction terms with carefully optimized parameters based on experimentally determined protein structures.[21] AWSEM-Suite is a hybrid of earlier versions of the AWSEM force field including AWSEM-Template and AWSEM-ER, which have been used in structure prediction.[2,27] Here the abbreviation ER reflects the use of "evolutionary restraints". Across the more than 100 blind tests in CASP13, AWSEM-Suite outperforms the simpler versions, AWSEM-Template and AWSEM-ER, in most domains. AWSEM-Suite stands as eighth in the free-modeling and hard-to-model categories among all the server predictors in the last CASP experiment. In this paper, we describe the predictions of AWSEM-Suite and their critical evaluation in the recent CASP13 competition. This study shows that AWSEM-Suite predicts protein structures with relatively high accuracy especially for proteins where the only available templates are in the "twilight zone" of sequence identity. We also highlight how AWSEM-Suite provides better predictions than other physics based algorithms in those cases where water-mediated contacts in structures are structurally important.

2. METHOD

2.1. The AWSEM-Suite Force Field. AWSEM is a predictive, coarse-grained, protein folding force field that represents the conformation of each type of amino acid residue except glycine by using only three explicit atomic centers (C_α, C_β, and O) and by employing an ideal peptide geometry assumption to locate three other implicit atoms (N, C′, and H atom attached to N).[21] A review by Schafer et al. describing how AWSEM incorporates the principles of energy landscape theory can be consulted by readers who are interested in details of the model.[28] The AWSEM-Suite Hamiltonian also includes a template bias term and a coevolutionary term along with the transferable AWSEM force field.[2,27] These terms are combined as detailed in eq 1. The backbone term restricts the

$$V_{\text{total}} = V_{\text{backbone}} + V_{\text{contact}} + V_{\text{fragmem}} + V_{\text{hydrogen}} + V_{\text{template}} + V_{\text{coev}} \quad (1)$$

peptide backbone through a harmonic potential and also contains excluded volume terms for all atomic centers. The contact term describes direct contact interactions and water/protein-mediated interactions of somewhat longer range. The fragment memory term is a bioinformatically informed term that biases the formation of local structure based on the known structures of overlapping sequences of peptide fragments, while the hydrogen bond term allows helical structures and β-sheets to form. The newly added terms in AWSEM-Suite, the template term and the coevolutionary term, guide the formation of global tertiary interactions in the folding process along with the purely physical interactions which are short range in space.

2.2. Constructing Template Guidance from Protein Databases. In the present instantiation of the software, a template of presumed homologous structure is first sought by HHpred using a minimum threshold confidence score of 95%. Structure−sequence matching has previously been employed in the associative memory Hamiltonian framework.[11,29,30] Templates found by HHpred usually have sequence identity of more than 10% to the predicted sequence with an *e*-value lower than 1. The *e*-value estimates the statistical significance of a match. It indicates how many chance hits with a score better than this would be expected if the database were to contain only hits that are actually unrelated to the query.[31] The aligned regions from the templates were renumbered according to their corresponding regions in the query sequence. Then, a pairwise distance matrix is created and the entries in the matrix are used to define the interresidue distance r_{ij}^N used in the definition of Q_{template}.

A collective variable, Q_{template} (eq 2), which ranges between 0 and 1, is used, then, to measure the structural similarity of

$$Q_{\text{template}} = \frac{\sum_{i,j} \Theta(r_{ij}^N - 2\ \text{Å}) \exp[-(r_{ij} - r_{ij}^N)^2/2\sigma_{ij}^2]}{\sum_{i,j} \Theta(r_{ij}^N - 2\ \text{Å})},$$
$$\sigma_{ij} = |j - i|^{0.15}$$
$$V_{\text{template}} = k_{\text{template}}(Q_{\text{template}} - 1)^2 \quad (2)$$

structures to the template by comparing the pairwise distances between a structure and the template. This collective variable then is used for a soft bias in V_{template}.

The Q_{template} is computed only from the aligned region of the template. $\Theta(r_{ij}^N - 2\ \text{Å})$ is 1 for $r_{ij}^N \geq 2.0$ Å and 0 otherwise, r_{ij}^N is the residue−residue distance in the templates, and r_{ij} is the corresponding pairwise distance in the simulation snapshot to which the template refers. The strength of the template term can be scaled to allow a desired balance with other energy

https://dx.doi.org/10.1021/acs.jctc.0c00188
J. Chem. Theory Comput. 2020, 16, 3977−3988

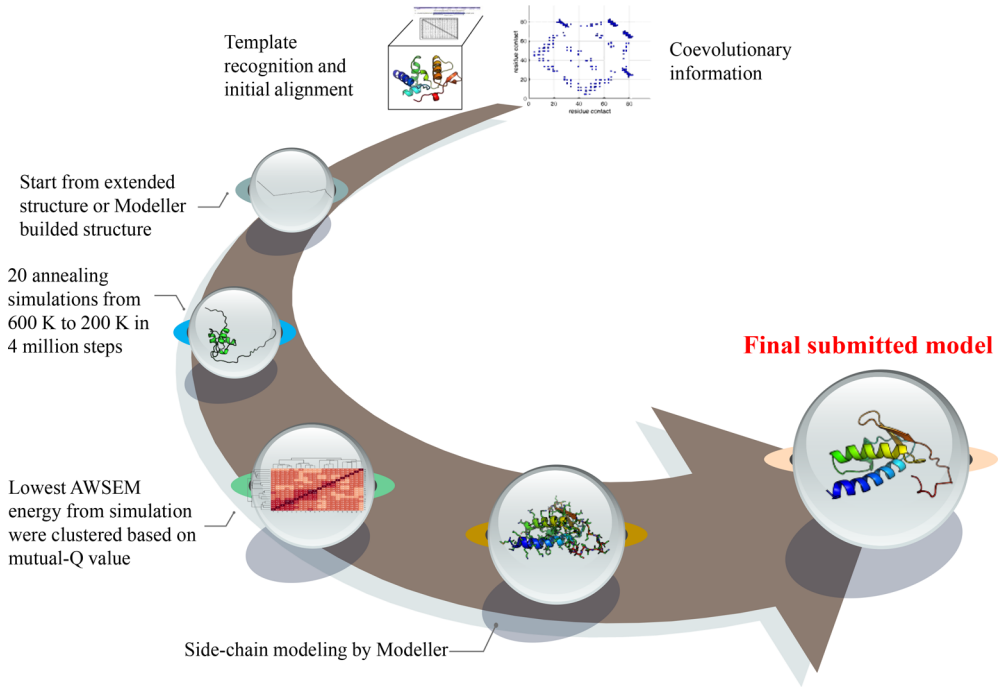


Figure 1. Protocol of structure prediction using AWSEM-Suite. The template term and the coevolutionary term are added to the transferable AWSEM force field. The template term uses tertiary information from a selected template based on HHpred multiple sequence alignment. The coevolutionary term uses coevolutionary information from an online server such as RaptorX-contact or DeepContact. The initial structures are chosen to be either an annealed extended structure generated by Pymol or a homology structure that is generated by Modeller if a template with an *e*-value less than 1 was found. Simulated annealing then starts from 600 K if the initial structure is an extended chain or from 400 K if the initial structure is based on an homologous model to cooling finally to 200 K. Twenty parallel jobs that have different initial velocity seeds were run. The lowest energy frames from each trajectory were chosen and clustered based on mutual-Qw values so as to pick five structures that are then rebuilt with side chains by Modeller. The final all-atom models were submitted to CASP13.

terms. The strength is scaled based on the strength and average length of protein sequence based on test cases using the following equation. The exponent of 1.5 was found to fit the increasing of σ_{ij} along with the sequence.

$$k_{\text{template}} = 200\left(\frac{\text{protein length}}{120}\right)^{1.5} \tag{3}$$

2.3. Coevolutionary Contact Restraints Inferred by RaptorX-Contact or DeepContact. The V_{coev} term is a pairwise additive contact term, which stabilizes any choice of contacts that can be specified as input. The location of the well centers for each contact pair were identified by specifying the expected contact length between two residues. The reference distance for each possible pair of amino acid types is based on a survey of thousands of Protein Data Bank (PDB) structures.[2] While any choice of contacts can be used, residue−residue contact pairs are generally predicted by either RaptorX-contact[32] or DeepContact.[33] The predicted pairs are incorporated into the force field using a restraint potential:

$$V_{\text{coev}} = -\frac{1}{2}k_{\text{coev}}\sum_{i,j}\Theta(\text{coev} - \text{predictions})$$
$$\exp[-(r_{ij} - r_{ij}^{\text{estimate}})^2/2\sigma_{ij}^{\ 2}], \quad \sigma_{ij} = |j - i|^{0.15} \tag{4}$$

In eq 4, the coefficient k_{coev} is used to scale the strength of this term relative to the other terms; here it is 1. Θ(coev − predictions) is 1 for the contact pairs that have been selected for use from the coevolutionary algorithms and 0 otherwise. The term r_{ij}^{estimate} is the predicted location of the well centers. The algorithms for contact pair prediction that were used depended on its availability when we submitted the jobs in the CASP competition.[32,33] Both of the contact prediction algorithms employ deep convolutional neural networks to learn structural interaction motifs from experimentally solved structures. A confidence score for each predicted contact pair is generated, reflecting the probability that it is correct. All predicted contacts with an estimated probability over 0.5 were used in the AWSEM-Suite prediction. Coevolutionarily inferred contacts tend to be less reliable with the decreasing estimation probability.

2.4. Obtaining Associative Memories from Fragments Based on Sequence Homology. The fragment memory term in the present instantiation utilizes the local-in-sequence structure information from known experimental structures to aid the prediction by applying a local bias. The database we used was the April 2018 version of the PDB database with the default culling threshold in PISCES Protein Sequence Culling Server.[34] The short peptide fragments were filtered by BLAST matrix to find the local match in sequence. In setting up the associative memory term, the fragments are taken both from the homologous sequences with more than 95% local sequence identity and from fragments of the nonhomologous sequences in places where there was poor coverage by the template. The 20 best matching fragments in the databases with a threshold of *e*-value over 0.0005 for the nonhomologous sequence database are selected.[21] The two sources of fragment memories are balanced to be weighted equally in the fragment memory term.

2.5. Simulation Protocol of AWSEM-Suite. The overall pipeline for predicting protein tertiary structures using

https://dx.doi.org/10.1021/acs.jctc.0c00188
J. Chem. Theory Comput. 2020, 16, 3977−3988

Table 1. Summary of Structure Prediction Results for AWSEM-Suite in CASP13 Competition[a]

target information			evaluation matrix							
			GDT-TS		Qw		RMSD		CE-RMSD	
target name	category	length	AWSEM	Rosetta	AWSEM	Rosetta	AWSEM	Rosetta	AWSEM	Rosetta
T0950-D1	FM	342	19.66	33.26	0.198	0.235	35.428	30.000	6.85/37%	3.29/44%
T0951-D1	TBM-easy	266	82.33	93.80	0.838	0.916	2.595	1.660	1.699/99%	1.118/99%
T0953s1-D1	FM	67	27.24	48.88	0.234	0.329	14.710	13.710	8.041/60%	3.594/60%
T0953s2-D1	FM/TBM	44	41.48	56.25	0.257	0.407	13.350	7.750	5.118/73%	4.258/73%
T0953s2-D2	**FM**	**111**	**61.26**	**45.27**	**0.573**	**0.307**	**4.738**	**13.520**	**3.182/94%**	**3.535/72%**
T0953s2-D3	**FM**	**93**	**39.52**	**21.77**	**0.280**	**0.224**	**12.890**	**13.038**	**2.69/34%**	**10.822/60%**
T0954-D1	TBM-hard	336	48.36	69.87	0.514	0.711	9.741	3.910	3.980/93%	2.902/98%
T0955-D1	FM/TBM	41	64.02	NaN	0.531	NaN	4.800	NaN	4.32/78%	NaN
T0957s1-D1	FM	108	35.42	39.35	0.294	0.361	13.850	11.970	6.125/59%	5.826/67%
T0957s1-D2	TBM-hard	54	52.78	65.28	0.369	0.525	6.900	4.950	4.408/89%	3.933/89%
T0957s2-D1	FM	155	43.55	45.81	0.430	0.465	9.110	6.180	5.375/83%	5.693/88%
T0958-D1	**FM/TBM**	**77**	**74.03**	**66.56**	**0.653**	**0.588**	**3.480**	**3.750**	**2.396/94%**	**2.943/94%**
T0960-D2	FM	84	35.42	56.55	0.260	0.458	9.820	8.030	5.054/67%	4.303/76%
T0960-D3	TBM-hard	89	54.21	73.88	0.449	0.670	10.050	5.190	5.210/72%	4.426/99%
T0960-D5	TBM-easy	105	62.86	70.24	0.550	0.652	4.780	4.060	2.931/91%	3.381/91%
T0963-D2	FM	82	36.28	38.41	0.306	0.337	10.046	7.520	6.621/68%	4.416/68%
T0963-D3	**TBM-hard**	**93**	**59.68**	**55.65**	**0.520**	**0.463**	**5.747**	**9.330**	**3.966/95%**	**3.204/77%**
T0963-D5	TBM-easy	94	61.97	70.21	0.566	0.553	4.480	4.940	2.989/94%	2.525/94%
T0965-D1	TBM-hard	313	63.90	66.05	0.651	0.665	5.213	4.390	3.467/97%	3.292/97%
T0966-D1	TBM-hard	492	54.22	61.08	0.608	0.685	5.530	4.080	4.013/93%	3.248/96%
T0967-D1	TBM-easy	79	81.01	93.67	0.765	0.908	3.090	1.540	1.841/91%	1.126/91%
T0968s1-D1	FM	119	44.49	66.74	0.409	0.668	7.835	3.890	5.587/87%	2.949/94%
T0968s2-D1	FM	116	41.74	71.30	0.385	0.708	9.878	4.700	4.619/69%	2.739/97%
T0969-D1	FM	354	23.38	30.30	0.194	0.255	29.195	22.450	5.278/34%	5.743/43%
T0970-D1	**FM/TBM**	**97**	**64.41**	**57.65**	**0.564**	**0.439**	**5.432**	**8.490**	**4.993/99%**	**3.117/66%**
T0971-D1	TBM-easy	130	83.85	96.35	0.803	0.951	2.740	1.520	2.087/92%	0.709/98%
T0976-D1	TBM-easy	120	70.83	77.92	0.669	0.721	3.737	6.310	2.718/93%	1.365/87%
T0976-D2	TBM-easy	124	72.18	73.59	0.714	0.654	3.861	6.520	2.426/97%	3.405/90%
T0980s1-D1	FM	105	32.69	40.14	0.304	0.332	13.394	14.360	6.679/69%	3.178/69%
T0984-D1	TBM-easy	504	40.68	60.10	0.490	0.628	6.910	5.360	4.586/86%	4.418/90%
T0984-D2	TBM-easy	147	44.03	71.27	0.480	0.649	6.850	4.580	3.384/87%	4.314/87%
T0986s1-D1	FM/TBM	92	49.19	68.21	0.492	0.660	7.040	6.400	4.761/87%	2.102/87%
T0986s2-D1	**FM**	**155**	**36.77**	**25.32**	**0.369**	**0.225**	**7.480**	**15.540**	**5.364/83%**	**4.396/41%**
T0990-D1	**FM**	**76**	**61.18**	**39.80**	**0.546**	**0.381**	**5.126**	**11.780**	**4.283/95%**	**6.471/74%**
T0990-D2	**FM**	**231**	**27.38**	**18.61**	**0.232**	**0.200**	**27.261**	**26.820**	**5.822/59%**	**6.583/45%**
T0990-D3	**FM**	**213**	**18.07**	**16.90**	**0.218**	**0.187**	**16.449**	**20.840**	**6.975/41%**	**11.157/34%**
T1003-D1	TBM-easy	437	68.89	89.17	0.706	0.846	5.911	8.550	3.174/94%	1.195/94%
T1005-D1	FM/TBM	326	42.33	55.83	0.444	0.512	10.131	8.480	5.134/71%	4.120/74%
T1006-D1	TBM-easy	77	86.69	92.53	0.867	0.912	2.277	1.810	1.219/94%	0.941/94%
T1008-D1	FM/TBM	77	45.45	68.18	0.436	0.619	8.440	3.610	7.839/94%	2.556/94%
T1011-D1	TBM-hard	302	38.12	65.36	0.446	0.596	7.342	5.220	4.907/85%	4.299/72%
T1016-D1	TBM-easy	202	70.55	81.56	0.701	0.782	3.696	3.640	2.217/95%	1.885/95%
T1018-D1	TBM-easy	334	71.26	88.10	0.744	0.880	3.540	2.090	2.637/98%	1.346/98%
T1021s1-D1	TBM-hard	149	48.32	66.44	0.462	0.632	7.070	4.750	3.858/86%	2.584/91%
T1021s2-D1	TBM-hard	349	47.28	66.26	0.537	0.644	6.210	9.720	4.38/94%	3.317/92%
T1021s3-D1	FM	178	33.28	40.66	0.324	0.333	10.620	8.600	6.82/67%	4.467/76%
T1021s3-D2	**FM**	**101**	**29.90**	**25.00**	**0.218**	**0.228**	**14.750**	**12.900**	**8.189/71%**	**8.241/55%**
T1022s1-D1	FM	156	22.76	34.78	0.223	0.295	16.720	15.090	6.658/36%	5.065/41%
T1022s1-D2	**TBM-hard**	**67**	**71.27**	**69.03**	**0.568**	**0.634**	**8.400**	**5.800**	**6.929/96%**	**4.976/96%**
T1022s2-D1	FM	525	39.38	62.10	0.473	0.683	8.011	5.850	5.283/64%	5.241/58%

[a]Bold rows show AWSEM-Suite has a better performance than Baker-RosettaServer in GDT-TS.

AWSEM-Suite is sketched in Figure 1. The initial structures for annealing the proteins are generated through the Modeller program when a homologue has been found; otherwise the initial structures are generated through Pymol as an extended structure.[35] β-Sheet secondary structure is difficult to form during a simple simulated annealing process. Bioinformatics tools do a generally good job in predicting the correct secondary structures without any templates, but of course they are not perfect.[36] We use a bias from the predicted secondary structures. The secondary structure predictions using the three states secondary structure output from RaptorX-property provides additional restraints to the backbone ϕ and ψ angles

https://dx.doi.org/10.1021/acs.jctc.0c00188
J. Chem. Theory Comput. 2020, 16, 3977−3988

biasing peptide fragments toward either α-helical or β-strand conformations in the backbone.[36]

AWSEM-Suite relies on the framework of the Large-Scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) open source software package and serves as an open source add-on.[37] Annealing simulations using AWSEM were carried out from 600 to 200 K if the initial structure was extended, or from 400 to 200 K if the initial structure was generated by Modeller cooling over 4 million steps. The total energy of each frame was calculated, and the lowest energy frame was picked out for further analysis. The resulting structures were then clustered based on mutual-Qw value, leading up to five representative structures from the three largest clusters that were finally picked. The output structures from AWSEM simulation include only three atoms, so rebuilding of the side chains on these coarse-grained structures was performed using Modeller version 9.18 with the default parameters.[35] The rebuilt structures were finally submitted to the CASP13 competition.

2.6. Free Energy Computed through Umbrella Sampling and Weighted Histogram Analysis Method. To study the landscapes of the models, we used umbrella sampling along an order parameter, Qw, with respect to the crystal structure of this protein to project the free energy landscapes of the proteins onto a single dimension. Qw is a metric with the form given by the following equation.

$$\mathrm{Qw} = \frac{2}{(N-2)(N-3)} \sum_{j-i\geq 2} \exp[-(r_{ij} - r_{ij}^{N})^2/2\sigma_{ij}^{2}],$$
$$\sigma_{ij} = |j-i|^{0.15} \quad (5)$$

Qw measures structural similarity by comparing pairwise distances between two structures. While emphasizing contacts, Qw resembles the root-mean-square deviation (RMSD) in depending on global accuracy but is less sensitive to the misprediction of dangling disordered segments. N is the total number of residues. The harmonic biasing potential used for constant temperature umbrella sampling simulations for 4 million steps was scaled to 200 kcal/mol. The biasing center values were chosen to be equally spaced from 0 to 1 with an increment of 0.02. The weighted histogram analysis method (WHAM) method is used to reconstruct the unbiased free energy landscapes from the umbrella sampling data.[38] The free energy landscapes can also be extrapolated to temperatures other than the temperature at which the simulation was performed.

2.7. Structure Prediction Accuracy Metrics. Four metrics are used to evaluate the accuracy of the structure predictions provided by the AWSEM-Suite protocol and by the other online server's algorithms: RMSD, CE-RMSD, Qw, and GDT-TS. All of these quantities measure the structural similarity between two structures, but they behave differently in different situations. These metrics can be used to judge the accuracy of each predicted model based on the corresponding native structure. RMSD describes the root-mean-square deviation of all atoms between the predicted and native structures when ideally aligned. CE (combinatorial extension)-RMSD calculates the RMSD of the best aligned regions of proteins after using an algorithm called CE alignment that discards those regions that would give a low match.[39] CE-RMSD reports also the percentage of aligned residues. GDT-TS (global distance test total score) is another matrix that is often associated with CASP to assess the quality of structure prediction.[40] GDT-TS is defined as the percentage of C_α carbons falling within 1, 2, 4, and 8 Å distances. These four scores are then added up and divided by 4.

Another metric, called Qo, is used to evaluate the formation of direct contacts with various distance cutoffs and compared to the reference structure. The Qo value evaluates the contacts with sequence separation of more than 4 under a distance cutoff D.

$$\mathrm{Qo} = \frac{2}{(N-3)(N-4)} \sum_{j-i\geq 4} \exp[-(r_{ij} - r_{ij}^{N})^2/2\sigma_{ij}^{2}]$$
$$\text{when } r_{ij}^{N} < D,\ \sigma_{ij} = |j-i|^{0.15} \quad (6)$$

Separate Qo values can be calculated based on the type of contact being considered. Short-range, long-range, and water-mediated contacts are separate types of interactions that are defined in previous papers.[21]

3. RESULTS

3.1. Performance of Tertiary Structure Prediction by AWSEM-Suite. To better understand the performance of the AWSEM-Suite relative to other cutting-edge physics based models, the complete set of best submitted structures from AWSEM-Suite (Group ID:124) and Baker-RosettaServer (Group ID:368)[41] in the CASP13 competition, segmented as domains, are summarized in Table 1 using four structure quality metrics: GDT-TS, Qw, RMSD, and CE-RMSD. Domains for which AWSEM-Suite has better performance are shown in bold in Table 1. The domain sequences were divided into three categories based on the modeling difficulty. These categories are the template-based-modeling (TBM) category which contains those targets where one or more structure templates can be identified from the sequence, the free-modeling (FM) category for sequences that do not have such templates, and the category TBM/FM which bridges these two categories.[42] The summary of 50 domains presents an analysis with deposited coordinates of the solved structures. Twenty of these domains belong to FM, seven of them belong to FM/TBM, and the remaining 23 domains belong to the TBM category. In the TBM groups, 13 of these domains were categorized as TBM-easy and the other 10 domains were categorized as TBM-hard. For both the FM/TBM category and the FM category, AWSEM-Suite ranked eighth among all 39 server predictors in the competition and yielded in several cases the best or second best structures among all server group algorithms. The performance indicates that using AWSEM-Suite is most advantageous when no good template information is available. We computed the average percentage of the GDT-TS value ratio from the AWSEM-Suite model to the best GDT-TS of the server predicted model to determine for which category AWSEM-Suite is the most optimal. The TBM-easy category shows the highest average GDT-TS ratio value (0.802), followed by FM/TBM (0.765), then TBM-hard (0.761), and last FM (0.666). The GDT-TS value ratio shows that, although the ranking of AWSEM-Suite in the TBM-easy category is not very high, nevertheless the models predicted by it are still quite reliable among all predictors.

3.2. Evaluation of Contact Accuracies for Different Contact Types. The performance of AWSEM-Suite depends on the resolution which is sought. To quantify this, we calculated the GDT values using 2, 4, 6, 8 and 10 Å as RMS cutoffs. The kernel density estimation of the histogram of these

https://dx.doi.org/10.1021/acs.jctc.0c00188
J. Chem. Theory Comput. 2020, 16, 3977−3988

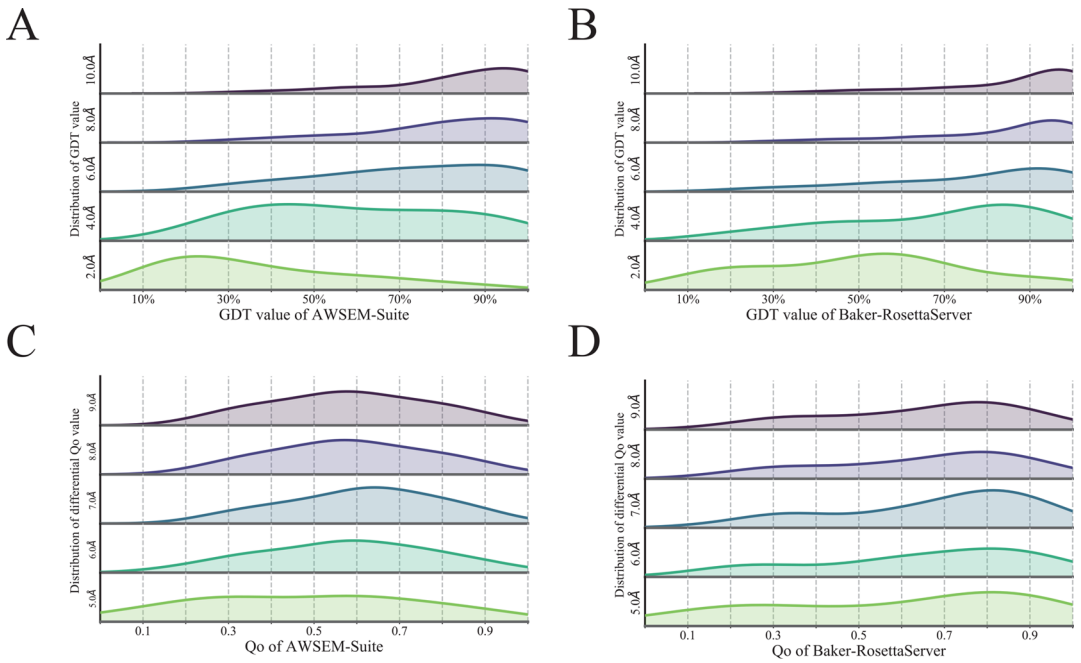


Figure 2. Distribution of GDT and Qo values from the AWSEM-Suite and from the Baker-RosettaServer predicted structures using different distance cutoffs. The histograms are calculated based on the 49 domains shown in Table 1 (except for one T0955-D1 for which Baker-RosettaServer group apparently did not submit a structure). A univariate or bivariate kernel density estimate is fitted based on the histogram that was separated into 10 bins. The GDT values are calculated using 2, 4, 6, 8, and 10 Å cutoffs. The Qo values are calculated using 5, 6, 7, 8, and 9 Å as defining a contact.

measurements from AWSEM-Suite and from Baker-RosettaServer are shown in Figure 2A,B. At short distances, the Rosetta-BakerServer pure fragment based approach works better. The GDT value distribution of AWSEM predicted structures of 2 Å resolution has a peak at around 0.2 compared to Rosetta's peak at 0.6. The peak of GDT distribution shifts to around 0.8 at the moderate resolution value of 6 Å, indicating AWSEM has good performance at medium ranges. AWSEM-Suite performance displays that one-third of the structures that AWSEM predicted have better quality than Baker-RosettaServer at 4 Å. This is also true at 8 Å resolution.

Next, we plotted the contact Qo values for various distance cutoffs for AWSEM-Suite and for Baker-RosettaServer. The Qo value provides a different perspective than comes from GDT-TS since it quantifies contact formation. The Qo values for both AWSEM-Suite and Rosetta show different distributions in Figure 2C,D. Again the distribution of Qo values from Baker-RosettaServer is bivariate, but the results for AWSEM-Suite display a univariate distribution that is nearly Gaussian. The different distributions indicate that in the range 5−7 Å there are some structures where AWSEM-Suite performs better than Baker-RosettaServer. From Figure S1, we see there are multiple domains, such as T0953s2-D3 and T0990-D2, for which AWSEM-Suite achieves a higher Qo value than Baker-RosettaServer. Most of these cases correspond to systems having a high Qo value coming from the water-mediated contacts. Water-mediated interactions are especially relevant in domain interactions. The Qo distributions for different distance cutoffs ranging from 4 to 12 Å are shown in Figure S1. While the performance of A7D (Google's Alphafold algorithm)[43] is best in most cases, the Qo values provided by AWSEM-Suite, Baker-RosettaServer, and RaptorX-DeepModeller[44] are comparable.

The water-mediated interactions described by Papoian et al. are most important for providing a funneled landscape for protein−protein binding.[22,45] The water degrees of freedom are usually averaged out in simulations because distant solvent structure relaxes rapidly compared with protein motions. Some water molecules however are directly involved with protein structure. These bound waters move very slowly compared with the bulk water molecules.[46] Waters near the surface of the protein are referred to as forming the hydration shell. Surface residues thus interact through the influence of water molecules in this shell giving rise to sometimes unexpected correlations. The water-mediated intermonomeric contacts are more distant than the direct protein residue contacts. Their threshold is within 6.5 Å based on previous calculation.[45] The Qo values were computed corresponding to the three different types of contacts—water-mediated, long-range, and short-range—for the results of AWSEM-Suite and Baker-RosettaServer. These are displayed in Table S1. There are several examples in the CASP13 where water-mediated protein association is clearly critical. For example, T0953s2-D2, which is a part of an adhesin protein in the fibers, involves the interplay with water molecules guiding the interactions with another trimeric β-helical tip.[47] T0970-D1, a homotetramer, has a clear water shell providing the interaction surface between two dimers. When we look at the water-mediated contacts, we see there are 27 domains for which the AWSEM-Suite has a higher Qo value than does Baker-RosettaServer (Table 2). AWSEM-Suite clearly has a better performance specifically for the water-mediated interaction, which is encoded by the multibody contact energy term in the AWSEM-Suite Hamiltonian.

3.3. Surveying the Performance of AWSEM-Suite by Using Selected Examples. The final structural results from AWSEM-Suite for eight selected domains are displayed in Figure 3. These models employed templates in the "twilight

https://dx.doi.org/10.1021/acs.jctc.0c00188
J. Chem. Theory Comput. 2020, 16, 3977−3988

Table 2. Statistics of Better Qo Value between AWSEM-Suite and Baker-RosettaServer

short contact		long contact		water-mediated contact	
AWSEM	Rosetta	AWSEM	Rosetta	AWSEM	Rosetta
18	32	13	37	27	23

zone" of low sequence identity. This class is one of the challenging parts of CASP13. The performance of AWSEM-Suite for these ordinarily difficult-to-predict domains shows the power of the AWSEM-Suite algorithm. Two of these were the best performance cases, T0958-D1 and T0970-D1, that belong to the TBM/FM modeling category. In order to visualize which regions of protein structure are predicted well, we show the contact maps of the AWSEM-Suite best predicted structure and the crystal structure for these domains. AWSEM-Suite most successfully predicts folds with almost all α-helix secondary structure pattern or where mostly one finds β-sheet secondary structure patterns. For those structures that are not well-predicted, an incorrect order or wrong alignment of the β-sheets hurts the quality of the structures. We note that in the laboratory such misaligned β-strands relax slowly, so it is

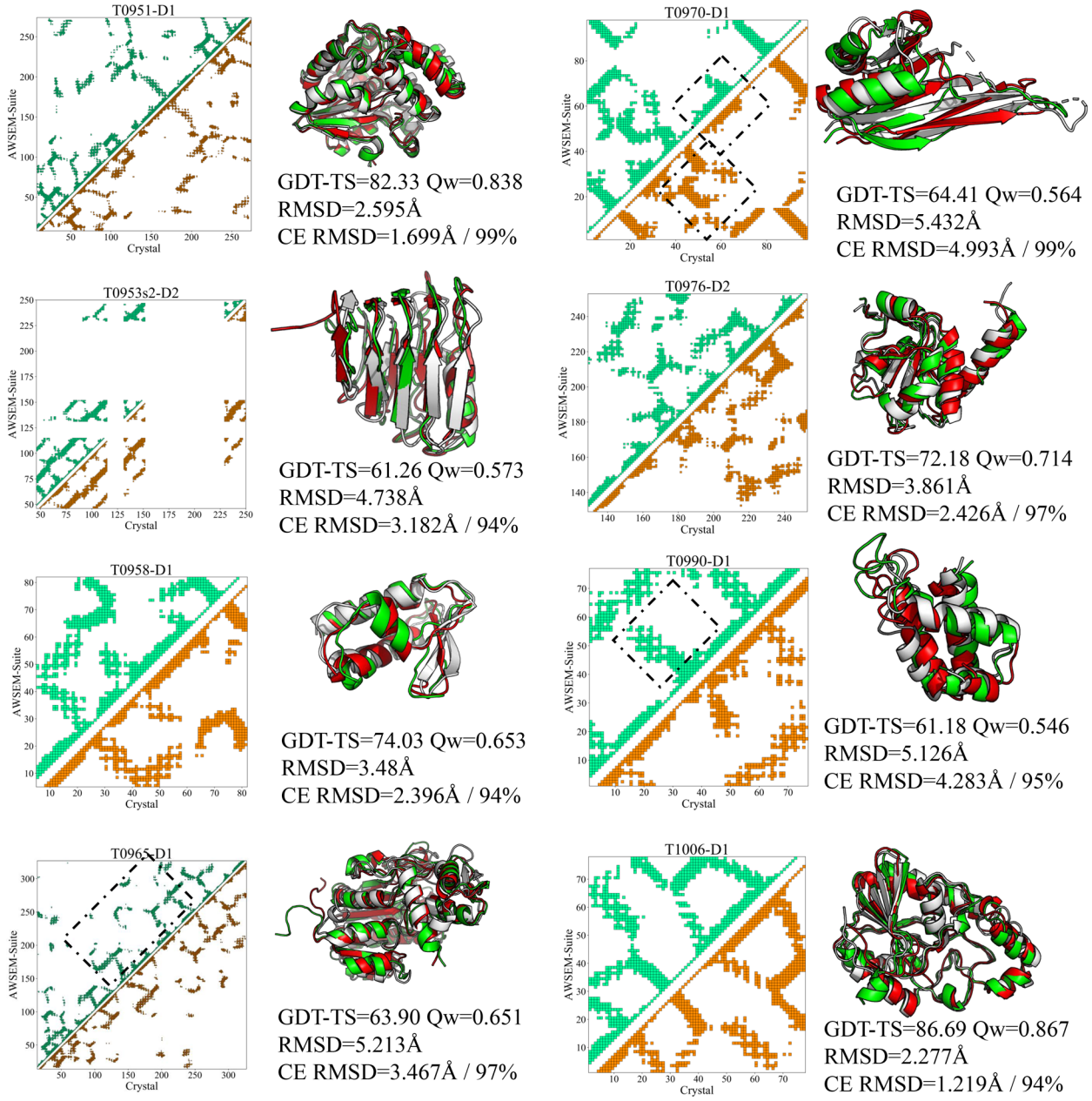


Figure 3. Prediction quality for eight targets using AWSEM-Suite in the CASP13 competition. The contact maps on the left side of each panel highlight any difference in structure with the crystal. Green squares correspond to contacts in the AWSEM predicted structure, while orange squares correspond to the crystal structure. The cutoff distance for contacts between C_α atoms has been set to 9.5 Å. The structural alignments of the submitted AWSEM structures with the corresponding X-ray crystal structures are shown on the right side of each panel. The best AWSEM-Suite predicted structures are shown in red, the best other server group structures are shown in green, and the corresponding X-ray crystal structures are shown in white. For these structures with RMSDs over 5, we marked the regions dissimilar to the crystal with black dashed−dotted squares.

https://dx.doi.org/10.1021/acs.jctc.0c00188
J. Chem. Theory Comput. 2020, 16, 3977−3988

possible that more extended simulation annealing would be helpful in these cases.

To evaluate whether the selection of the lowest energy frame structures in the simulation is effective, the Qw values for the best Qw frame, for the lowest energy frame, and for the ending frame for each of 20 simulations of several selected domains are shown in Figure S2. The difference in Qw value between that for the best Qw frame and that for the lowest energy frame is usually as low as 0.05 or less. The ending frame resembles very much the lowest energy frame since the energy generally decreases with the reduction of temperature. The lowest energy frame usually occurs in the last part of the simulation. In general, picking the lowest energy frame yields a reasonable result.

3.4. Analysis of the Contributions of Different Energy Terms to Structure Prediction Using AWSEM-Suite. Depending on the availability of an appropriate quality template and predicted coevolutionary contact information, AWSEM-Suite can emphasize the transferable interactions or add either or both of the template terms and the coevolutionary terms. To compare the performances of the different enhancements of previous AWSEM algorithms, we also carried out predictions using AWSEM-Template and using AWSEM-ER separately as controls in Figure 4. When compared with AWSEM-Suite, AWSEM-Template is missing the coevolutionary constraint term while AWSEM-ER does not contain the template term. The results indicate that different ones of these three AWSEM versions can yield the highest prediction quality. There are, however, two cases where AWSEM-Suite showed a worse performance, indicating that coevolutionary information sometimes conflicts with the template information. Both of these examples belong to the TBM-hard category, which means their predicted coevolutionary contact information contains some false-positive results or failed somehow to generate a sufficient number of accurate contacts and thus incorrectly guided the protein folding during the simulations. In T0976-D2, we also found the performance of AWSEM-ER is better than those of the other two algorithms. Table S2 shows the true positive rate of the predicted contacts and the sequence identity of the template used. Having a high true positive rate for predicted contacts greatly improves the performance of the AWSEM-Suite prediction. The same is true for the high sequence identity template. DeepContact performs somewhat worse than the RaptorX-contact prediction which displays many incorrect positive hits.

3.5. Comparing Different Combinations of Structural Guidance with AWSEM-Suite. As stated under Method, AWSEM-Suite is a Hamiltonian that combines many different energy contributions. In order to see how each part of the energy contributes to the process of simulated annealing, in Figure 5 we display a stacked bar plot for the different energy terms for 29 domains. The differences in energy contribution from the initial extended structure to the ending frame representing the final folded structure are calculated. For domains where there is a template term, the fragment library term contributes the largest part of the reduction of the energy, indicating the peptide structure at short sequence lengths plays the main role in the AWSEM-Suite force field. The contributions of the coevolutionary term and the template term are similar in magnitude, demonstrating the balance of the two parts is reasonable in the current pipeline. The coevolutionary term becomes the major contributor to the energy loss when there is no template information. This may

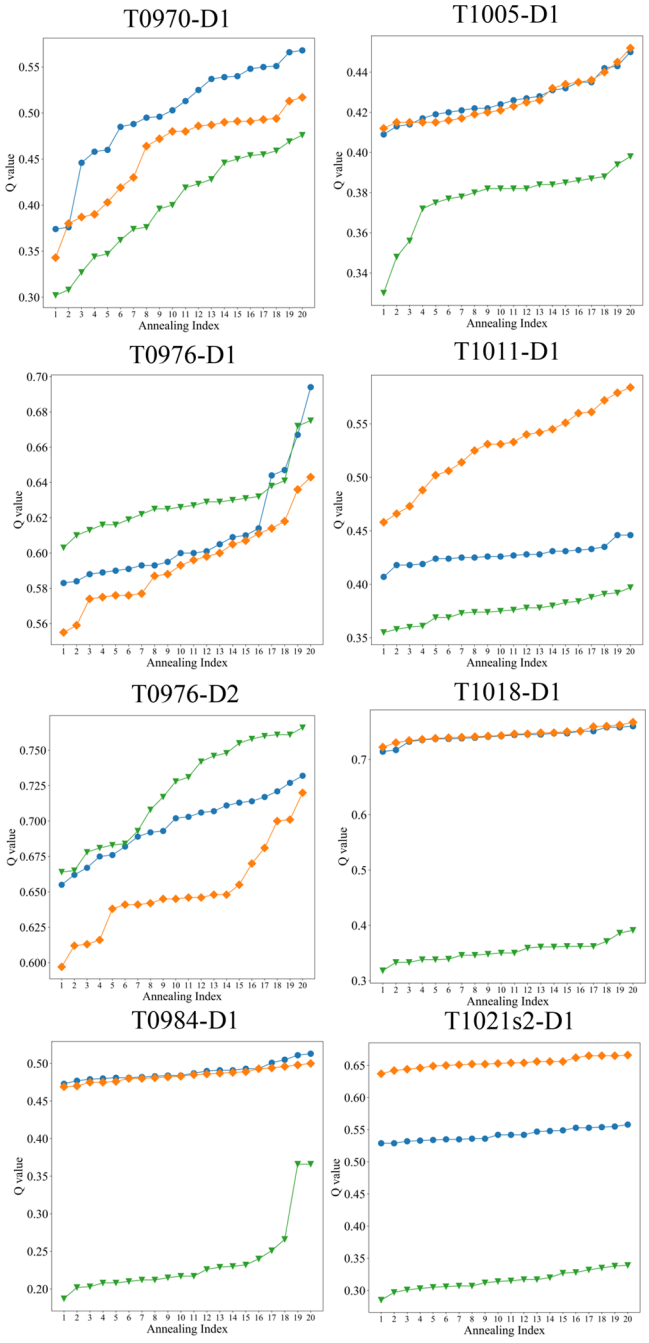


Figure 4. Comparison of the performances of different enhancements in AWSEM-Suite. In each case, the Qw value of the best energy frame from each of 20 simulated annealings with AWSEM-Suite (blue), AWSEM-Template (orange), and AWSEM-ER (green) force fields is plotted in descending order. In general, AWSEM-Template+ER compares favorably with AWSEM-Template.

suggest that sometimes the coevolutionary term is too strong and distorts the local structure in free-modeling predictions.

4. DISCUSSION

4.1. Funneling of the Folding Landscape of T0958 Enhanced by Differing Structural Guidance. Template-based modeling has provided the most successful structure prediction in several CASP competitions. Template-free targets or low sequence identity template targets are especially

https://dx.doi.org/10.1021/acs.jctc.0c00188
J. Chem. Theory Comput. 2020, 16, 3977−3988

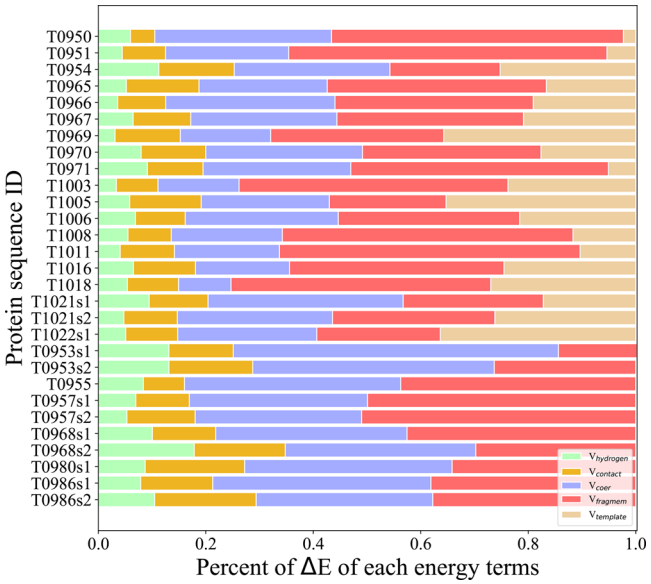


Figure 5. Contributions to the energy change from each energy term during the process of simulated annealing. The energy difference is calculated based on the difference of the initial extended structure and the ending frame of the submitted trajectory which is then divided by the total energy change. The hydrogen bonding term (green), contact term (orange), coevolutionary term (purple), fragment library term (red), and template term (pink) are shown in different colors in a stacked bar plot. Twenty-nine simulated proteins were evaluated. Ten of these did not have any available template information when the simulation was performed.

important because they may stand as pioneer members of new protein families.

T0958 has 96 residues, but only residues 5−81 are sufficiently resolved to be counted for domain 1. T0958-D1 includes two α-helices and a short antiparallel β-sheet in the C-terminal. These form a winged helix−turn−helix domain.[48] The template used for T0958 was the chain B of 3gva. This chain has a 20% sequence identity to T0958. The *e*-value of the template is 9.90×10^{-6}, which indicates however that this is a reliable template with a likely similar folding pattern. The GDT-TS for the template itself is 64.83, and the Qw value for its aligned region is 0.508. The upper left of Figure S3 shows the template only provides partial secondary structure, as its α-helices are not complete, and the whole β-sheet part is lost. The AWSEM-Suite force field folds the secondary structure which is not formed in the template in this case. The backbones are predicted with very high quality, but far-range side chains are not predicted very well. Comparison of the AWSEM-Suite prediction structure with the A7D and Baker-RosettaServer's best models is highlighted in Figure S3. We found that the AWSEM-Suite model has a better formed antiparallel β-sheet.

To highlight the role of the transferable AWSEM-Suite energy potential, we developed free energy profiles for T0958 as a function of different energy components using umbrella sampling simulations. The computed energy profiles were plotted as a function of Qw at 300 K. As shown in Figure 6, the combined AWSEM-Template+ER free energy shows the best equilibrium Qw value (around 0.53) at the energy minimum. AWSEM without template and coevolutionary bias has a lower Qw value at the minimum (around 0.35). The profiles for AWSEM with template bias and AWSEM with coevolutionary

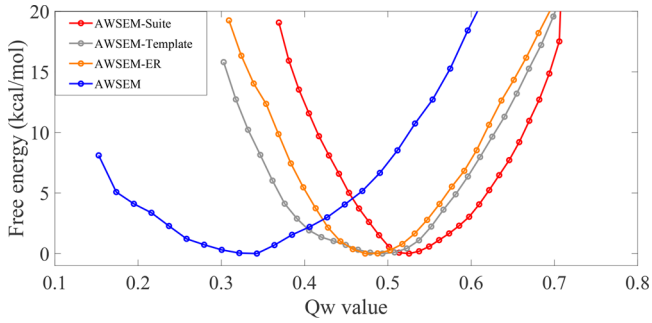


Figure 6. Restraints from both templates and coevolutionary information guiding the folding of T0958 toward a native-like basin. Free energy profiles for T0958 as a function of Qw are shown for four different prediction force fields at 300 K: AWSEM (blue circles), AWSEM-Template (gray circles), AWSEM-ER (orange circles), and AWSEM-Template+ER (red circles). AWSEM-Template+ER has the highest Qw value at the lowest free energy point, which indicates the force field prefers a higher accuracy model.

bias lie in the middle. These free energy profiles demonstrate that the incorporation of template and coevolutionary guided funneling of the landscape more strongly to a native-like basin. To evaluate the effect of varying the folding temperature, we also plotted the free energy profile of AWSEM force field at 200 K and at 400 K in Figure S4. For 200 K, the energy minimum has a Qw value similar to the one at 300 K. AWSEM force field in 400 K is predicted to be less favorable to the current native state at 300 K, suggesting that 400 K would be too high a temperature to use for simulation.

The variations of the different energy terms in the simulation of T0958 are plotted as a function of Qw in Figure S5. We found that the energy terms in the standard AWSEM force field are well funneled to a Qw value around 0.35 but level out at higher Qw values. The template bias term and coevolutionary term still fall at higher Qw values, so the total energy decreases monotonically to a higher Qw value. Thus, the adding of these two terms improves the performance of the structure prediction of standard AWSEM and helps the system sample structures closer to the native structure.

4.2. Future Prospects of AWSEM-Suite Tools for Structure Prediction. What factors influence the performance of AWSEM-Suite? The ratio of correctly predicted contacts to the domain length is plotted in Figure S6. The sequence identity of the template displays a stronger positive correlation to the structure quality as indicated in the same figure. AWSEM-Suite performs more favorably than Baker-RosettaServer does when the water-mediated contacts are specifically considered. It is clear however that contact distance prediction improves *ab initio* modeling. Unlike template information, coevolutionary constraints only provide partial contact information; this information is not as complete as the contact information inherent in template-based approaches. There is a strong correlation between the number of contacts correctly predicted versus the Qw or GDT-TS value in Figure S7. Clearly, more reliable inferences of the coevolutionary contacts will greatly improve the quality of predictions. Based on the analysis of these eight selected structures, we found the true positive rates from different algorithms of inference could differ. Currently, we suggest using RaptorX-contact as the best choice. We also noticed that using a higher probability threshold to incorporate the contact pairs into the coev term would decrease the false positive rate. We must point out that

https://dx.doi.org/10.1021/acs.jctc.0c00188
J. Chem. Theory Comput. 2020, 16, 3977−3988

sometimes the mutations of pairs of residues display correlation without the residues being in close proximity because they are involved in function. This phenomenon can hurt the performance of the coevolutionary information in pure structure prediction. Poor template information can also hurt the performance of prediction because proteins can have both open and closed forms, for example, for an ion channel. The energy contributions from different parts of the Hamiltonian are similar to each other, indicating perhaps a stronger weight should be applied to the template term once the sequence identity of the template is sufficiently high. In CASP13, the server provided relatively unrefined predictions. We have shown new types of refinement technologies such as PCA-guided refinement are able to improve significantly the quality of the final predicted structure for most CASP cases, even when contact information is not completely correct.[49] Another indicator of expected prediction quality is the preponderance of specific secondary structures.[50] The GDT-TS value from AWSEM-Suite correlates well with the fraction of specific secondary structures. The results showing a strong positive correlation to the percent of helix and β-strand in the crystal structure can be seen in Figure S8. Since the currently used secondary structure prediction prescribes only three categories (helix, sheet, and coil), converting from SS3 to SS8 may help AWSEM-Suite perform better especially for the barrel-like structures and when there are helix kinks in the native structure. The AWSEM-Suite algorithm has been implemented as an online server at https://awsem.rice.edu.[51]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.0c00188.

> Qo values under different types of contacts and over different distance cutoffs; template and coevolutionary contact pairs for selected structures; contact map comparison of T0958-D1; quality of energy-based blind selection from prediction trajectory; free energy profiles of T0958; energy terms in different AWSEM force fields for T0958; correct contact versus length and sequence identity; ratio of correct contacts in different parts divided by domain length vs Qw value of each domain; Qw values vs secondary structure percent (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Peter G. Wolynes − *Center for Theoretical Biological Physics, Department of Physics, Department of Chemistry, and Department of Biosciences, Rice University, Houston, Texas 77005, United States;* orcid.org/0000-0001-7975-9287; Phone: (713) 348-4101; Email: pwolynes@rice.edu

Authors

Shikai Jin − *Center for Theoretical Biological Physics and Department of Biosciences, Rice University, Houston, Texas 77005, United States;* orcid.org/0000-0001-9525-4166

Mingchen Chen − *Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005, United States*

Xun Chen − *Center for Theoretical Biological Physics and Department of Chemistry, Rice University, Houston, Texas 77005, United States*

Carlos Bueno − *Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005, United States*

Wei Lu − *Center for Theoretical Biological Physics and Department of Physics, Rice University, Houston, Texas 77005, United States*

Nicholas P. Schafer − *Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005, United States*

Xingcheng Lin − *Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0002-9378-6174

José N. Onuchic − *Center for Theoretical Biological Physics, Department of Physics, Department of Chemistry, and Department of Biosciences, Rice University, Houston, Texas 77005, United States;* orcid.org/0000-0002-9448-0388

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.0c00188

Author Contributions

#S.J., M.C., and X.C.: These three authors contributed equally to this work.

Notes

The authors declare no competing financial interest.

The source code for the AWSEM-Suite force field within the LAMMPS suite is available for download on GitHub (https://github.com/adavtyan/awsemmd). A server implementation of AWSEM-suite can be found at https://awsem.rice.edu. Academic users are encouraged to submit their sequences for structure predictions. Other documentations and references can be found on this Web site: http://awsem-md.org.

■ ACKNOWLEDGMENTS

This work was supported by the Center for Theoretical Biological Physics and sponsored by an NSF grant (PHY-1427654). Additional support was also provided by the D. R. Bullard-Welch Chair at Rice University, Grant C-0016. The authors would like to thank the Center for Research Computing at Rice University for the computational resources during CASP13, and Dr. Xiaoqin Huang and Mr. Mark Coccimiglio at Rice University for additional resources during server development.

■ REFERENCES

(1) Fetrow, J. S.; Babbitt, P. C. New computational approaches to understanding molecular protein function. *PLoS Comput. Biol.* **2018**, *14*, e1005756.

(2) Sirovetz, B. J.; Schafer, N. P.; Wolynes, P. G. Protein structure prediction: making AWSEM AWSEM-ER by adding evolutionary restraints. *Proteins: Struct., Funct., Bioinfo.* **2017**, *85*, 2127−2142.

(3) Service, R. F. Problem Solved* (*sort of). *Science (Washington, DC, U. S.)* **2008**, *321*, 784−786.

(4) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins: Struct., Funct., Genet.* **1995**, *21*, 167−195.

(5) Wolynes, P. G. Evolution, energy landscapes and the paradoxes of protein folding. *Biochimie* **2015**, *119*, 218−230.

(6) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How Fast-Folding Proteins Fold. *Science (Washington, DC, U. S.)* **2011**, *334*, 517−520.

(7) Jothi, A. Principles, Challenges and Advances in ab initio Protein Structure Prediction. *Protein Pept. Lett.* **2012**, *19*, 1194−1204.

(8) Baker, D. Protein Structure Prediction and Structural Genomics. *Science (Washington, DC, U. S.)* **2001**, *294*, 93−96.

(9) Schaffer, A. A. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* **2001**, *29*, 2994−3005.
(10) Zimmermann, L.; Stephens, A.; Nam, S.-Z.; Rau, D.; Kübler, J.; Lozajic, M.; Gabler, F.; Söding, J.; Lupas, A. N.; Alva, V. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J. Mol. Biol.* **2018**, *430*, 2237−2243.
(11) Goldstein, R. A.; Luthey-Schulten, Z. A.; Wolynes, P. G. Protein tertiary structure recognition using optimized Hamiltonians with local interactions. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 9029−9033.
(12) Goldstein, R. A.; Luthey-Schulten, Z. A.; Wolynes, P. G. In *Recent Developments in Theoretical Studies of Proteins*, 1st ed.; Elber, R., Ed.; World Scientific Publishing Co.: 1996; Chapter 6, pp 359−388.
(13) Morcos, F.; Pagnani, A.; Lunt, B.; Bertolino, A.; Marks, D. S.; Sander, C.; Zecchina, R.; Onuchic, J. N.; Hwa, T.; Weigt, M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, E1293−E1301.
(14) Morcos, F.; Jana, B.; Hwa, T.; Onuchic, J. N. Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 20533−20538.
(15) Weigt, M.; White, R. A.; Szurmant, H.; Hoch, J. A.; Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 67−72.
(16) Ovchinnikov, S.; Kinch, L.; Park, H.; Liao, Y.; Pei, J.; Kim, D. E.; Kamisetty, H.; Grishin, N. V.; Baker, D. Large-scale determination of previously unsolved protein structures using evolutionary information. *eLife* **2015**, *4*, e09248.
(17) Tetchner, S.; Kosciolek, T.; Jones, D. T. Opportunities and limitations in applying coevolution-derived contacts to protein structure prediction. *Bio-Algorithms and Med-Systems* **2014**, *10*, 243−254.
(18) Goldstein, R. A.; Luthey-Schulten, Z. A.; Wolynes, P. G. Optimal protein-folding codes from spin-glass theory. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 4918−4922.
(19) Friedrichs, M. S.; Wolynes, P. G. Toward Protein Tertiary Structure Recognition by Means of Associative Memory Hamiltonians. *Science (Washington, DC, U. S.)* **1989**, *246*, 371−373.
(20) Bohr, H. G.; Wolynes, P. G. In *Neural Networks: From Biology to High-Energy Physics*, 1st ed.; Benhar, O., Bosio, C., Giudice, P. D., Grandolfo, M., Eds.; World Scientific Publishing Co.: 1991; Chapter 3, pp 261−275.
(21) Davtyan, A.; Schafer, N. P.; Zheng, W.; Clementi, C.; Wolynes, P. G.; Papoian, G. A. AWSEM-MD: Protein Structure Prediction Using Coarse-Grained Physical Potentials and Bioinformatically Based Local Structure Biasing. *J. Phys. Chem. B* **2012**, *116*, 8494−8503.
(22) Zheng, W.; Schafer, N. P.; Davtyan, A.; Papoian, G. A.; Wolynes, P. G. Predictive energy landscapes for protein-protein association. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 19244−19249.
(23) Hinckley, D. M.; Freeman, G. S.; Whitmer, J. K.; de Pablo, J. J. An experimentally-informed coarse-grained 3-site-per-nucleotide model of DNA: Structure, thermodynamics, and dynamics of hybridization. *J. Chem. Phys.* **2013**, *139*, 144903.
(24) Zhang, B.; Zheng, W.; Papoian, G. A.; Wolynes, P. G. Exploring the Free Energy Landscape of Nucleosomes. *J. Am. Chem. Soc.* **2016**, *138*, 8126−8133.
(25) Krepel, D.; Davtyan, A.; Schafer, N. P.; Wolynes, P. G.; Onuchic, J. N. Braiding topology and the energy landscape of chromosome organization proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 1468−1477.
(26) Wang, Q.; Chen, M.; Schafer, N. P.; Bueno, C.; Song, S. S.; Hudmon, A.; Wolynes, P. G.; Waxham, M. N.; Cheung, M. S. Assemblies of calcium/calmodulin-dependent kinase II with actin and their dynamic regulation by calmodulin in dendritic spines. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 18937−18942.
(27) Chen, M.; Lin, X.; Lu, W.; Schafer, N. P.; Onuchic, J. N.; Wolynes, P. G. Template-Guided Protein Structure Prediction and Refinement Using Optimized Folding Landscape Force Fields. *J. Chem. Theory Comput.* **2018**, *14*, 6102−6116.
(28) Schafer, N. P.; Kim, B. L.; Zheng, W.; Wolynes, P. G. Learning To Fold Proteins Using Energy Landscape Theory. *Isr. J. Chem.* **2014**, *54*, 1311−1337.
(29) Hardin, C.; Eastwood, M. P.; Prentiss, M. C.; Luthey-Schulten, Z.; Wolynes, P. G. Associative memory Hamiltonians for structure prediction without homology: alpha/beta proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 1679−1684.
(30) Hegler, J. A.; Latzer, J.; Shehu, A.; Clementi, C.; Wolynes, P. G. Restriction versus guidance in protein structure prediction. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 15302−15307.
(31) Altschul, S. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.
(32) Xu, J. Distance-based protein folding powered by deep learning. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 16856−16865.
(33) Liu, Y.; Palmedo, P.; Ye, Q.; Berger, B.; Peng, J. Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks. *Cell Syst.* **2018**, *6*, 65−74.
(34) Wang, G.; Dunbrack, R. L. PISCES: a protein sequence culling server. *Bioinformatics* **2003**, *19*, 1589−1591.
(35) Webb, B.; Sali, A. *Current Protocols in Bioinformatics*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2016; pp 5.6.1−5.6.37.
(36) Wang, S.; Li, W.; Liu, S.; Xu, J. RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Res.* **2016**, *44*, W430−W435.
(37) Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comput. Phys.* **1995**, *117*, 1−19.
(38) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **1992**, *13*, 1011−1021.
(39) Shindyalov, I. N.; Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng., Des. Sel.* **1998**, *11*, 739−747.
(40) Zemla, A.; Venclovas, C.; Moult, J.; Fidelis, K. Processing and analysis of CASP3 protein structure predictions. *Proteins: Struct., Funct., Genet.* **1999**, *37*, 22−29.
(41) Ovchinnikov, S.; Park, H.; Kim, D. E.; DiMaio, F.; Baker, D. Protein structure prediction using Rosetta in CASP12. *Proteins: Struct., Funct., Bioinfo.* **2018**, *86*, 113−121.
(42) Kinch, L. N.; Kryshtafovych, A.; Monastyrskyy, B.; Grishin, N. V. CASP13 target classification into tertiary structure prediction categories. *Proteins: Struct., Funct., Bioinfo.* **2019**, *87*, 1021−1036.
(43) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, *577*, 706−710.
(44) Källberg, M.; Wang, H.; Wang, S.; Peng, J.; Wang, Z.; Lu, H.; Xu, J. Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.* **2012**, *7*, 1511−1522.
(45) Papoian, G. A.; Ulander, J.; Wolynes, P. G. Role of water mediated interactions in protein-protein recognition landscapes. *J. Am. Chem. Soc.* **2003**, *125*, 9170−9178.
(46) Gottschalk, M.; Dencher, N. A.; Halle, B. Microsecond exchange of internal water molecules in bacteriorhodopsin. *J. Mol. Biol.* **2001**, *311*, 605−621.
(47) Dunne, M.; Denyes, J. M.; Arndt, H.; Loessner, M. J.; Leiman, P. G.; Klumpp, J. Salmonella Phage S16 Tail Fiber Adhesin Features a Rare Polyglycine Rich Domain for Host Recognition. *Structure* **2018**, *26*, 1573−1582.e4.
(48) Lepore, R.; Kryshtafovych, A.; Alahuhta, M.; Veraszto, H. A.; Bomble, Y. J.; Bufton, J. C.; Bullock, A. N.; Caba, C.; Cao, H.; Davies, O. R.; Desfosses, A.; Dunne, M.; Fidelis, K.; Goulding, C. W.;

https://dx.doi.org/10.1021/acs.jctc.0c00188
J. Chem. Theory Comput. 2020, 16, 3977−3988

Gurusaran, M.; Gutsche, I.; Harding, C. J.; Hartmann, M. D.; Hayes, C. S.; Joachimiak, A.; Leiman, P. G.; Loppnau, P.; Lovering, A. L.; Lunin, V. V.; Michalska, K.; Mir-Sanchis, I.; Mitra, A.; Moult, J.; Phillips, G. N., Jr; Pinkas, D. M.; Rice, P. A.; Tong, Y.; Topf, M.; Walton, J. D.; Schwede, T. Target highlights in CASP13: Experimental target structures through the eyes of their authors. *Proteins: Struct., Funct., Bioinfo.* **2019**, *87*, 1037−1057.

(49) Lin, X.; Schafer, N. P.; Lu, W.; Jin, S.; Chen, X.; Chen, M.; Onuchic, J. N.; Wolynes, P. G. Forging tools for refining predicted protein structures. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 9400−9409.

(50) Zhang, C.; Mortuza, S. M.; He, B.; Wang, Y.; Zhang, Y. Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12. *Proteins: Struct., Funct., Bioinfo.* **2018**, *86*, 136−151.

(51) Jin, S.; Contessoto, V. G.; Chen, M.; Schafer, N. P.; Lu, W.; Chen, X.; Bueno, C.; Hajitaheri, A.; Sirovetz, B. J.; Davtyan, A.; Papoian, G. A.; Tsai, M.-Y.; Wolynes, P. G. AWSEM-Suite: a protein structure prediction server based on template-guided, coevolutionary-enhanced optimized folding landscapes. *Nucleic Acids Res.* **2020**.

https://dx.doi.org/10.1021/acs.jctc.0c00188
J. Chem. Theory Comput. 2020, 16, 3977−3988