Fall 9-29-2021

DISCRETIZED GEOMETRIC APPROACHES TO THE ANALYSIS OF PROTEIN STRUCTURES

John Holland
John.E.Holland.GR@Dartmouth.edu

Fall 9-29-2021

DISCRETIZED GEOMETRIC APPROACHES TO THE ANALYSIS OF PROTEIN STRUCTURES

John Holland

DISCRETIZED GEOMETRIC APPROACHES TO THE ANALYSIS OF PROTEIN STRUCTURES

A Thesis
Submitted to the Faculty
in partial fulfillment of the requirements for the
degree of

Doctor of Philosophy

in

Computer Science

by John Holland

Guarini School of Graduate and Advanced Studies
Dartmouth College
Hanover, New Hampshire

September 2021

Examining Committee:

_____
(chair) ***Gevorg Grigoryan***

_____
Chris Bailey-Kellogg

_____
Michael Ragusa

_____
Roland Dunbrack

_____
F. Jon Kull, Ph.D.
Dean of the Guarini School of Graduate and Advanced Studies

Proteins play crucial roles in a variety of biological processes. While we know that their amino acid sequence determines their structure, which in turn determines their function, we do not know why particular sequences fold into particular structures. My work focuses on discretized geometric descriptions of protein structure—conceptualizing native structure space as composed of mostly discrete, geometrically defined fragments—to better understand the patterns underlying why particular sequence elements correspond to particular structure elements. This discretized geometric approach is applied to multiple levels of protein structure, from conceptualizing contacts between residues as interactions between discrete structural elements to treating protein structures as an assembly of discrete fragments. My earlier work focused on better understanding inter-residue contacts and estimating their energies statistically. By scoring structures with energies derived from a stricter notion of contact, I show that native protein structures can be identified out of a set of decoy structures more often than when using energies derived from traditional definitions of contact and how this has implications for the evaluation of predictions that rely on structurally defined contacts for validation. Demonstrating how useful simple geometric descriptors of structure can be, I then show that these energies identify native structures on par with well-validated, detailed, atomistic energy functions. Moving to a higher level of structure, in my later work I demonstrate that discretized, geometrically defined structural fragments make good objects for the interactive assembly of protein backbones and present a software application which lets users do so. Finally, I use these fragments to generate structure-conditioned statistical energies, generalizing the classic idea of contact energies by incorporating specific structural context, enabling these energies to reflect the interaction geometries they come from. These structure-conditioned energies contain more information about native sequence preferences, correlate more highly with experimentally determined energies, and show that pairwise sequence preferences are tightly coupled to their structural context. Considered jointly, these projects highlight the degree to which protein structures and the interactions they comprise can be understood as geometric elements coming together in finely tuned ways.

Acknowledgements

I would like to thank my advisor, family, colleagues, and friends for their endless support, without which I would not have been able to complete my research. My advisor, Dr. Gevorg Grigoryan, has been an incredible mentor whose expertise, support, and guidance have been essential. I am deeply appreciative and thankful for my committee members, Dr. Chris Bailey-Kellogg, Dr. Michael Ragusa, and Dr. Roland Dunbrack, for all of their helpful guidance and feedback. My wife, Lillie, has been nothing short of amazing in her dedication and love during my degree. I cannot express how much I appreciate the support my parents have provided me throughout my life and during my degree. My brother, James, has been a constant and excellent source of support and advice. Finally, I would like to thank my colleagues and friends, in particular Katherine McCoy and Yu Zhao, for all of their invaluable help and wisdom.

Table of Contents

List of Tables

List of Figures

1 Introduction

Proteins are an essential building block of biological systems and play diverse roles throughout them, acting as enzymes, regulating cellular activity, responding to internal and external stimuli, and forming functional and morphological structures. Despite their critical role in this array of crucial biological processes, many aspects of their behavior are poorly understood. In particular, while we know that their amino acid sequence determines their structure, which in turn determines their function, we do not have models that adequately explain why particular sequences result in particular structures (or structural ensembles). Learning the relationship between sequence and structure is key to understanding how proteins behave and can be made to behave and underlies a variety of long-standing problems—not only the central challenges of structure prediction and sequence design but many offshoots of these such as achieving high binding specificity, controlling allosteric networks, generating designable backbones, and predicting the effects of mutations.

Given the vast sizes of both sequence and structure space and the complex physics driving how proteins fold and interact, finding a definite relationship between sequence and structure may at first glance appear daunting or even intractable. However, a closer look at the literature reveals decades of work demonstrating that both folding sequence space and designable structure space are highly patterned. That is to say, most sequences do not fold into a stable structure but those that do collectively exhibit patterns, or degeneracy in the space, such that a sequence design method, tasked with finding a sequence that folds into the structure of interest, need not consider every possible sequence but only those with particular features. Analogously, most structures are not designable (i.e., there exist no sequences that fold into them) but those that do collectively exhibit patterns such that a structure prediction method, tasked with finding the structure that the sequence of interest folds into, need not consider every possible structure but only those with particular features.

Excitingly, the joint space of foldable sequences and designable structures is so highly patterned that fundamental statements can be made about how it works. Contact potentials have shown that when a pair of residues interacts, particular pairs of amino acids are more likely than others, with some pairs so unfavorable they almost never

appear[1]. The Ramachandran plot demonstrates that nearly all backbone dihedral angles adhere to a strict distribution, with various parts of the distribution corresponding to specific secondary structural elements, and that some amino acids occupy even more specific parts of that distribution[2,3]. Work on contact prediction has revealed that for families of sequences, many mutations do not occur independently but are instead coupled to mutations at other positions[4]; these patterns of coupled mutations not only suggest which parts of the sequence are folded together in the native structure as contacts but, more fundamentally, that these contacts form the way they do because of the intricate relationship between sequence and structure.

Recent work in the Grigoryan Lab[5–9] has focused on the patterns that can be found in designable structure space. In particular, it has been shown that only a small set of structural motifs—termed tertiary motifs because they are in general not contiguous in sequence but span contacts and other higher order interactions—is needed to cover, to a high degree of accuracy, most native structures (essentially, just several hundred are needed to cover half of all experimentally determined structures)[7]. In other words, despite the dazzling diversity of native and designed structures, most parts of most known structures can be reduced to a configuration of the same pseudo-discrete tertiary motifs. This insight allows us to imagine new strategies for learning the relationship between sequence and structure; if designable structures can be understood as comprising an arrangement of tertiary motifs, then it might be possible to discover patterns in the distribution of these motifs' sequences that explain why proteins behave the way they do. A model that can explain why particular sequence motifs result in particular structure motifs would be invaluable for tackling the aforementioned challenges in structural biology.

This work builds on these insights by building discretized geometric models of protein structure in a mostly pairwise manner. That is, this work focuses on building an understanding of protein structure and sequence-structure relationships by defining (pairwise) contacts geometrically and utilizing discretization, both in the definition via rotamers and in order to construct discrete fragments of contact-centered structure. By using this discretized geometric notion of contacts to estimate pairwise statistical energies, it is shown that this definition better identifies native sequence-structure

relationships than traditional definitions. By evaluating sequence-based contact prediction with this definition, it is shown that these traditional definitions inflate performance. By incorporating this notion of contact and other geometric descriptors into a more comprehensive statistical potential (which inherently discretizes these values) it is shown that these simple, geometric notions can be used to score sequence-structure relationships on par with well-validated, atomistic energy functions. Next, pair motifs built around these contacts are shown to be good objects for assembling protein backbones in a visual, interactive application. Finally, I demonstrate that these pair motifs generalize the notion of a contact potential; rather than conditioning amino acid pair statistics on a contact binary, these statistics can instead be conditioned on an ensemble generated around any pair motif of interest, resulting in energies specific to the contact geometry they come from. These structure-conditioned energies contain more information about native sequence preferences when compared to a contact potential, correlate highly with experimentally determined coupling energies (and more so compared to a contact potential), and can be used to show a general relationship between structural similarity and energetic similarity on a pairwise level, with similar contact fragments resulting in similar energies and, strikingly, vice-versa.

The focus not just on patterns in folding-sequence-and-designable-structure space, but specifically on discretized, geometric, and pairwise analysis has a long history in protein science and these kinds of techniques have a broader history in science as well.

First, discretization, the process of transforming a continuous or otherwise complex space into definite regions or categories, has been used to gain insight in myriad scenarios. For a historical example in biology, a Punnett square simplifies the complex genetic process of hybridization by discretizing phenotypes and categorizing them as dominant or recessive, providing reliable predictions while obviating detailed biochemical explanations or genetic analysis of the possible alleles involved. Looking at the history of protein science, the classic work of Monod, Wyman, and Changeux on allostery and oligomerization in proteins[10] provides a good model of how some inter-protein associations can alter the energetic cost of further associations (i.e., an allosteric partner reducing the cost of binding to an orthosteric one) by conceptualizing proteins as discrete, mostly rigid, geometric objects. The assignment of secondary structural

elements to particular regions of the Ramachandran plot is a widely used example of discretization which demonstrates that eliding some information (the exact backbone dihedral angle of a residue) can elucidate other things (the similarity between one alpha helical segment and another, despite slightly different angles, and the hydrogen bonds that make this similarity relevant). Rotamer libraries provide yet another example of the success of discretization, enumerating the most likely and relevant side chain conformations at a given position by exploiting the statistical patterns in native sidechain geometries[11,12]. Another notable instance of this comes from work on docking: the discretization of space employed by the fast Fourier transforms widely used in docking methods demonstrates that sufficiently small voxels in space can often adequately encode information about how a structure occupies the space it is in[13].

Second, geometric techniques have a long history of success as well. For a historical example, the study of geometrical optics was used to construct corrective lenses centuries before the physical nature of light was understood. Looking to proteins specifically, in addition to the work on allostery mentioned above, many models and tools of proteins conceptualize their structures geometrically at the residue or even fold level, rather than in terms of the atomistic physics driving this geometry. From structural classifications like CATH[14], to "periodic table" style analysis like that of protein-protein interactions[15], to geometric fragments being used to assemble larger structures[16], the geometric analysis of protein structure often detects patterns that lower-level analysis struggles to reproduce. A rotamer library constructed based on the statistics in the PDB, for instance, does not require a detailed model of the physics responsible for the useful degeneracy in sidechain conformation space, only the knowledge of this degeneracy needed to encode the most common conformations.

Finally, pairwise analysis, in particular the use of inter-residue contacts, is a widely used framework in the study of protein structure for a number of reasons. A model of structure comprising only self terms, without any higher order ones, is generally not useful because most features of structure cannot be explained by such simple models, as the prevalence of epistasis and non-zero coupling energies across many kinds of protein structures, among many other points of evidence, conclusively proves. While higher-than-pairwise features have certainly been used in a variety of structural models, there are

good reasons to focus on pairwise analysis in particular. From a practical angle, the combinatorics of higher order analysis quickly imposes data sparsity issues on any statistical analysis; there are just 400 amino acid pairs but 8,000 triplets and 160,000 quadruplets. Given the number of high quality, non-redundant structures in the PDB (tens of thousands) and the uneven amino acid distribution of pairs, triplets, etc., there are not enough data points to collect rich statistics on all 8,000 triplets and the problem grows multiplicatively worse as the degree of analysis increases.

But beyond technical issues like this, which may eventually be alleviated as more structural data are generated, more fundamental problems impinge on higher-order analysis in a way they do not on lower-order alternatives. A structural biologist familiar with contact potentials and other pairwise models can easily leverage the insights they contain and apply them to problems of interest. It is far harder to learn, visualize, or even examine 8,000 triplet-wise or 160,000 quadruplet-wise interaction preferences than it is for the 400 pairwise ones and therefore models that make conclusions in these higher-order regimes cannot be easily incorporated into intuitive models, which makes it much harder to detect faulty assumptions in the model or mistakes in the implementation, both essential steps in any scientific project. These psychological limitations become increasingly relevant as hyper-complex machine learning models become ingrained in research agendas; a black-box model with millions of parameters cannot be efficiently debugged if many of the parameters represent high order relationships that researchers have no prior knowledge of. Requiring complex models be reducible to pairwise analysis ensures that the results can be scrutinized and compared to prior structural knowledge. Even if higher-order analysis is unavoidable, building a strong base of self and pairwise knowledge that it can be checked against (e.g., by querying to what extent a higher-order model differs from the predictions of the pairwise ones and being critical if it differs in implausible ways) reduces the likelihood of mistakes or faulty assumptions going undetected.

More broadly, the study of lower order patterns complements that of higher order ones. The highest order models, such as a deep learning network whose input is a sequence and output is a structure, benefit from being interpretable using lower order approximations. For a simple example, if we have *a priori* knowledge of how pairs of

residues interact via a contact potential, we can begin validating a sequence-to-structure deep learning network by scoring its interactions with the potential and comparing those scores to those of known native structures. If the network's structures' scores are systematically less favorable than those of natives, then there is no need to perform more time consuming and expensive experimental validation; more work is evidently needed before the network's predictions can be considered realistic. For a practical and recent example of this, AlphaFold[17] leverages pairwise inter-residue distance analysis both to improve its structure predictions and to validate intermediate steps in the deep learning network.

Much of the work detailed below can be seen as testing the limits of discretized and pairwise geometric models. The replacement of traditional, distance-based definitions of contact with contact degree can be seen as an attempt at squeezing as much utility out of pairwise structural descriptors as possible by finding a definition that incorporates more context than plain distance. The statistical potential built around contact degree and conditioned against measures of backbone angles and environment effectively answers the question of how much predictive power can be encoded in such a simple model. The work on tertiary motif-based backbone design tests how feasible it is to interactively design backbones based around contact fragments. Finally, the work structure-conditioned amino-acid couplings is explicitly an extension of contact potentials that explores the information revealed by conditioning amino-acid statistics on particular pair motifs (i.e., the information contained within statistics centered around discrete, geometric, pairwise fragments).

2 Evaluating contact prediction with a novel definition of inter-residue contact

2.1 Introduction

Formation of tertiary structure in proteins is dependent on the establishment of close through-space interactions, often between amino-acid residues distant in sequence. Inter-residue contacts should impose constraints on evolutionary dynamics. Thus, mutations at

contacting pairs are expected to be coupled in the evolutionary record. Such compensatory mutational coupling in evolutionarily related proteins enables statistical methods to infer which positions in a multiple sequence alignment (MSA) of structurally homologous proteins may be in contact. The idea of using predicted inter-residue contacts, discovered by analyzing MSAs, to aid in structure prediction has been around for decades[18], but has experienced a resurgence recently due to the massively increased amount of available sequence data[19–22]. Several investigators have now shown that the large sequence datasets available today enable much more robust contact predictions than their smaller counterparts[23–26]. However, any successful contact prediction model must avoid inferring spurious couplings[27]. Indeed, pairs of mutations can co-occur by chance or appear to couple due to phylogenetic biases, unrelated to maintaining structure[28]. Trying to determine which apparent correlations correspond to contacts has been approached from a variety of angles, such as enforcing maximum entropy to remove spurious indirect couplings[4], using probabilistic graphical models to learn correlations from sparse statistics[19], and estimating evolutionary distance relationships to determine the significance of correlations[29]. Impressive precision rates upwards of 90% have been reported for the most confident few predicted contacts[19], which can be enough for practical structure prediction[30–32].

Several challenges in contact prediction remain to be addressed, however. For instance, accuracy drops considerably when more than a few contacts are predicted[33]. Additionally, current methods require large numbers of sequences in the right range of homology that are unavailable in many practical scenarios[34]. But perhaps more importantly, the high reported prediction rates are in relation to fairly loose definitions of contact between two residues—for instance, any two atoms being within 8 Å of each other in any available structure belonging to the family in question[4] or any two Cβ atoms being within 8 Å[35]. This aids in achieving a high precision rates, but such loose definitions may not be optimal for the purpose of making predictions about structure.

A reasonable quality measure for a contact definition is the amount of information, per contact, contributed towards discriminating correct from incorrect structural models. Guided by this idea, we propose a new contact definition, termed contact degree (CD), and show that the knowledge of a single CD-based contact

eliminates considerably more solution space in structure prediction than does knowledge of a contact defined via common distance-based criteria. On the other hand, we find that MSA-based contact prediction results in much lower precision for CD-based contacts as it does for traditional contact definitions. Thus, the remaining challenges in contact prediction are better revealed by adopting stricter definitions of contact that are ultimately more informative for structure prediction.

Motivated by these observations, and the need for both an informative contact definition and accurate prediction rates, we consider an additional source of information that can be used to supplement co-variation in contact prediction. In particular, we consider the fact that different amino-acid pairs have different a priori expectations of being in contact, based on observations in native proteins. These differential expectations are captured within so-called residue-level statistical contact potentials[36]. While contact potentials cannot encode all of the information required to fold a structure[37], they can be used to differentiate native structures from many varieties of decoys[38]. Thus, if a pair of MSA positions predicted to co-vary tends to be occupied by amino-acid pairs that do not score favorably by a residue-level contact potential, this should weaken our belief that the pair represents a true contact. On the other hand, if mutations at this pair of positions appear to compensate for each other in such a way as to produce consistently favorable contact potentials, this pair may be more likely to be a true contact. Based on this intuition, we propose a metric that combines a contact potential with a co-evolution score (from DCA or MetaPSICOV) and show it to improve the precision of both DCA and MetaPSICOV alone considerably.

The idea of using contact potentials in contact prediction has been put forth in recent work[35,39–41]. For example, Jones *et al.* include contact potential values as one of the many features in their neural network for predicting contacts[35]. In the analysis of the EPSILON-CP method developed by Stahl *et al.*[41], the mean contact potential energy is deemed an important feature in the neural net. However, to our knowledge, the isolated benefit of contact potentials towards improving contact prediction has not been studied extensively. Furthermore, it has been unclear to what extent the significant degradation in performance resulting from the utilization of more informative contact definitions can be mitigated by the incorporation of contact potentials. Here we show that the added benefit

of incorporating contact potentials can be quite significant, especially in conjunction with contact definitions that are difficult to predict but highly informative. Further, we find that averaging contact potential values across all sequences of an MSA (for a given pair of positions) produces significantly higher improvements in performance. Thus, in summary, this work both points out the significant room for improvement that remains towards accurately predicting informative inter-residue contacts and proposes a route towards attaining such improvement.

2.2 Results

2.2.1 Contact definition and interpretation

The best criterion for classifying a pair of residues as being in contact depends on the application—i.e., the meaning that a contact is interpreted to have. For many applications, including structure prediction and protein design, a reasonable interpretation of a contact would be a pair of residues that are capable of participating in a direct physical interaction in such a way as to have significant influence on each other's amino-acid identities. Such an interpretation would be particularly well aligned with the goal of predicting contacts based on mutational co-variation. It follows then that spatial proximity should be an important but not the sole determinant of a contact. The opportunity to establish an interaction, as determined by the surrounding structural environment, should also be a contributor. Traditional distance-based contact definitions capture the former but not the latter factors. Fig. 2.1 shows several examples of typical structural circumstances where a distance-dependent definition of contact does not agree with structural intuition. In particular, we consider three different commonly-used contact definitions: the one proposed by Morcos *et al.* in presenting the DCA method—i.e., two residues with at least one pair of non-hydrogen atoms within 8 Å of each other (hereafter referred to as the "any-heavy" definition)[4], the official CASP definition—i.e., two residues with Cβ (or Cα in the case of Glycine) atoms within 8 Å of each other (referred to as the "Cβ" definition)[35], and a definition based on a metric used in coarse-grained modeling—two residues with centroids within 6 Å of each other (referred to as the "centroid" definition)[42,43]. The top row in Fig. 2.1A–C shows situations where each of these definitions, respectively, would classify as contacting position pairs that, by

structural intuition, should not directly affect each other's amino-acid identity; even with stricter thresholds than stated above. For example, in Fig. 2.1A, the two positions involved are on opposite sides of a β-sheet. On the other hand, the bottom row in Fig. 2.1A–C demonstrates examples where each of the above definitions, respectively, would fail to classify as contacting residue pairs that would be expected to affect each other's amino-acid identities and, therefore, would be expected to co-vary even with more generous cutoffs that those typically used.

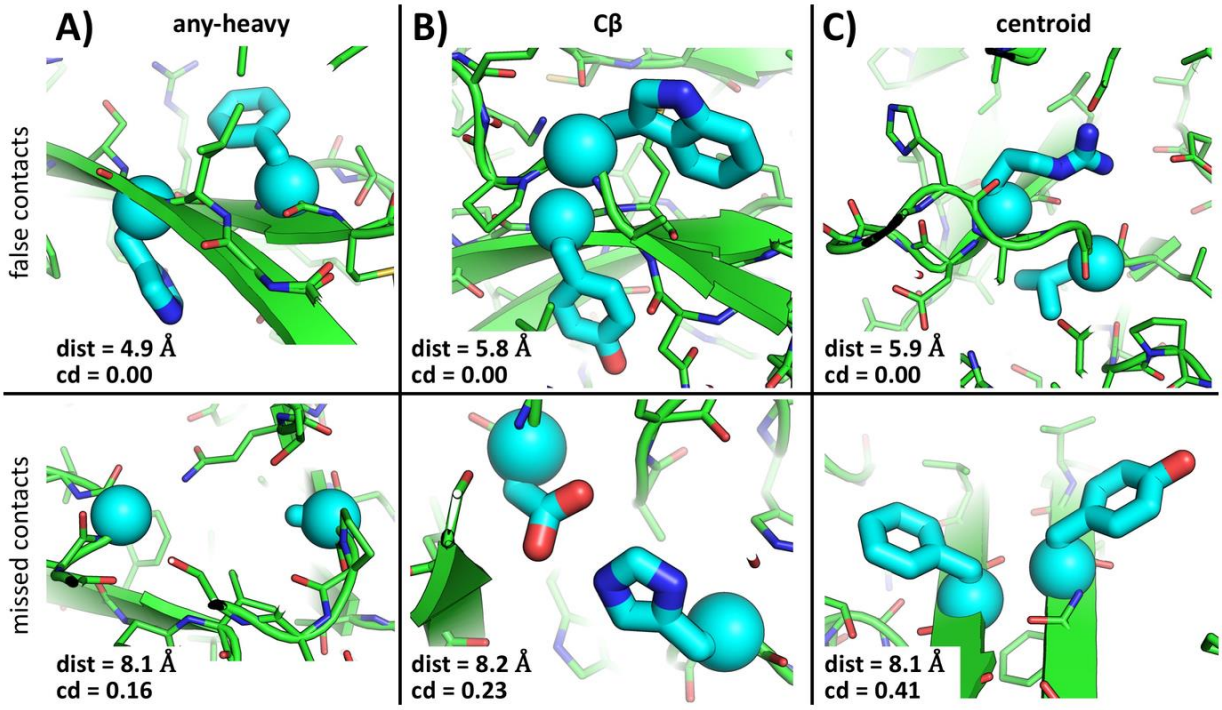


Fig. 2.1 Distance-based contact definitions can flag unreasonable contact geometries or fail to capture position pairs likely to co-vary. **A)**, **B)**, and **C)** correspond to any-heavy, Cβ, and centroid-based contact definitions, respectively. The top row show examples where residue pairs that would be classified as contacting, on the basis of a rather strict distance cutoff in each case, do not appear to have immediate influence on each other. Whereas the bottom row demonstrates cases where a rather loose distance cutoff, in each case, would miss an apparent contact (i.e., a pair of positions likely to co-vary). The value of the corresponding distance metric, along with the contact degree value, are shown at the bottom of each panel. Residue pairs of interest are highlighted in thick cyan sticks, with their Cα atoms shown with spheres. The contacts shown in the top row correspond to

position pairs (A126, A141), (A328, A344), and (V120, V128) from PDB structures 3JUM, 3JU4, and 1LM8 for **A)-C)**, respectively, and those in the bottom row correspond to position pairs (A55, A62), (C102, C201), and (B144, B153) from PDB structures 1JUH, 1JUH, and 4ACF for **A)-C)**, respectively. These illustrative cases were identified by manual inspection of a random set of 100 PDB structures. Molecular renderings created with PyMOL.

In order to overcome these flaws, we propose a more structurally informative definition of a contact, based on the metric of a contact degree, which we have used in prior work[5,7]. Rather than demarcate a contact based purely on distance, a contact degree considers all possible amino-acid and rotamer pair combinations for the position pair of interest and produces a value between 0 to 1 that represents the fraction of interfering rotamer pairs (i.e., those with non-hydrogen atoms within 3 Å of each other). More formally, the contact degree between two positions i and j, denoted $CD_{i,j}$, is defined as follows:

$$CD_{i,j} = \sum_{r_i \in R_i} \sum_{r_j \in R_j} C_{i,j}(r_i, r_j) \cdot \mathbb{P}_i(r_i) \cdot \mathbb{P}_j(r_j)$$

Eq. 2.1

Here, R_i is the set of every allowed rotamer from every amino acid at position i (based on some rotamer library) that does not clash with the backbone. $\mathbb{P}_i(r_i)$ is the probability of rotamer r_i at position i, taken from the rotamer library and normalized to unity over all non-clashing rotamers at i. $C_{i,j}(r_i, r_j)$ is unity if rotamer r_i placed at position i interferes with rotamer r_j placed at j (i.e., there are non-hydrogen atoms within 3 Å between the two rotamer side-chains) and zero otherwise. Thus, if none of the sterically possible rotamer pairs at the two positions interfere with each other, then $CD_{i,j} = 0$. At the other extreme, if all sterically possible rotamer pairs placed at i and j interfere, then $CD_{i,j} = 1$. To create a binary definition of contact, a cutoff c can be chosen so that all pairs of positions with a contact degree of at least c are considered to be in contact. In this study,

we use $c = 0.1$. This gives an average of 4.1 contacts per residue, which is in line with our structural intuition.

Contact degree addresses the limitations of the distance-based definitions discussed above. Obviously, spatial proximity contributes to the criterion because position pairs far apart in space cannot host mutually interfering rotamers. However, the opportunity to interact is also accounted for by means of assessing contact via allowable rotamers (i.e., rotamers that are compatible with the surrounding structural environment). For example, all of the cases in Fig. 2.1 are classified appropriately with a contact-degree cutoff of 0.1 (i.e., the top row is classified as non-contacting and the bottom row as contacting; corresponding contact degree and distance values are shown in each panel of Fig. 2.1). As an added benefit, because contact degree does not rely on the sidechain coordinates of a structure, it is sequence independent. That is, one can assess the possibility of a contact between two positions in a protein structural template, independent of the specific sequence associated with it (unlike, for example, with the centroid-based definition). This lends itself better to interpreting contacts as implying mutational co-dependence, especially within an evolutionary protein family.

2.2.2 Contact potential as a quality measure of contact definition

Given any geometric definition of inter-residue contact, one can derive a corresponding contact potential—a table of statistical pseudo-energies that reflect the relative propensity of different amino-acid types to be in contact within native-like protein structures[38,44,45]. We reasoned that a good quality metric for a contact definition would be the predictive power of the resulting contact potential. Of course, this is not the only quality metric, particularly given the fact that a contact potential alone is not sufficient to solve structure prediction[37]. Still, all else being equal, if the contact potential emergent from one contact definition systematically outperforms that emergent from another definition, it would seem reasonable to conclude that the former contact definition is better. Indeed, if a particular definition often classifies as contacting residue pairs that, in reality, do not significantly interact or influence each other, the resulting contact potential should have little meaning or predictive power. A similar argument would apply if a particular definition fails to classify many of the truly mutually influencing residues as contacting.

To evaluate the quality of our CD-based contact definition, we set out to compare the contact potential emergent from it relative to potentials emergent from several commonly used distance-based contact definitions (see Table 2.1). To isolate just the effect of the contact definition, we used the same simple reference-state model in all cases. This model assumes random redistribution of amino acids among contacts, such that the statistical potential associated with the contact between amino acids *a* and *b* is:

$$E(a,\ b)\ =\ -\log\left(\frac{N_c(a,\ b)}{\left(1\ +\ I_{a,b}\right)f(a)f(b)N_c}\right)$$

Eq. 2.2

Here $N_c(a, b)$ is the number of observed contacts between *a* and *b*, $f(a)$ is the frequency of amino acid *a* in the database, N_c is the total number of observed contacts (for all amino-acid pairs), and $I_{a,b}$ is an indicator variable that evaluates to unity if *a* and *b* are different and to zero otherwise. As the structural database, we used the PISCES set prepared by the Dunbrack Lab that included 8,106 structures, each with a maximum resolution of 2.2Å culled at 30% sequence identity[46]. Fig. 2.2 shows the pairwise contact-potential values for the CD-based and any-heavy-based potentials, which are generally well correlated (R = 0.81), but with non-negligible differences. For example, the mean absolute energy for the CD-based definition is 0.39, higher than the corresponding value of 0.23 for the any-heavy-based definition. This means that the degree of over/under-representations in amino-acid identities at contacting positions is generally higher for the CD-based definition, suggesting that it captures more of the underlying structural determinants of a true interaction. The same is also true when comparing the CD-based definition with Cβ and centroid definitions, which have mean absolute energies of 0.17 and 0.35, respectively. Hereafter, we will refer to the CD-based, any-heavy-based, Cβ-based, and centroid-based contact potentials as E_{CD}, E_1, E_2, and E_3, respectively (see Table 2.1).

		-2.75	-0.93	-0.79	-1.14	-0.84	-1.10	-0.72	-0.52	-0.30	0.29	-0.25	-0.07	-0.15	0.45	0.63	1.28	-0.25	0.84	1.25	0.16
		C	M	F	I	L	V	W	Y	A	G	T	S	N	Q	D	E	H	R	K	P
-1.46	C		-0.69	-0.71	-0.78	-0.74	-0.77	-0.66	-0.54	-0.42	0.03	-0.20	-0.12	0.17	0.16	0.37	0.49	-0.39	0.17	0.47	-0.14
-0.84	M	-0.38		-0.62	-0.60	-0.50	-0.62	-0.55	-0.49	-0.33	0.10	-0.02	0.05	0.30	0.25	0.52	0.54	-0.08	0.33	0.60	-0.03
-0.82	F	-0.53	-0.62		-0.74	-0.67	-0.67	-0.60	-0.57	-0.38	0.02	0.06	0.06	0.34	0.32	0.58	0.63	-0.08	0.21	0.51	-0.12
-0.70	I	-0.34	-0.47	-0.65		-0.86	-1.07	-0.51	-0.65	-0.65	0.20	-0.21	-0.01	0.34	0.22	0.51	0.43	-0.08	0.25	0.36	0.01
-0.62	L	-0.32	-0.43	-0.61	-0.57		-0.93	-0.56	-0.55	-0.61	0.26	-0.06	0.08	0.42	0.22	0.55	0.52	-0.03	0.23	0.49	0.01
-0.46	V	-0.28	-0.35	-0.52	-0.51	-0.45		-0.48	-0.53	-0.65	0.18	-0.25	0.01	0.39	0.22	0.49	0.42	-0.04	0.19	0.39	-0.01
-1.07	W	-0.57	-0.63	-0.78	-0.54	-0.57	-0.47		-0.49	-0.24	0.00	0.10	0.03	0.24	0.11	0.51	0.54	-0.24	-0.11	0.28	-0.39
-0.63	Y	-0.38	-0.47	-0.64	-0.46	-0.41	-0.34	-0.73		-0.24	0.07	0.08	0.08	0.23	0.28	0.53	0.59	-0.09	0.07	0.18	-0.24
0.08	A	0.07	-0.04	-0.11	-0.07	-0.10	-0.04	-0.18	-0.02		0.27	-0.02	0.16	0.42	0.46	0.48	0.67	0.18	0.48	0.77	0.15
0.29	G	0.04	0.08	-0.02	0.13	0.19	0.17	-0.18	-0.05	0.34		0.22	0.22	0.32	0.52	0.42	0.74	0.23	0.48	0.71	0.33
0.00	T	-0.04	-0.08	-0.20	-0.12	-0.05	-0.06	-0.30	-0.21	0.23	0.20		-0.07	0.02	0.06	0.05	0.18	0.03	0.24	0.33	0.14
0.15	S	0.00	0.05	-0.09	0.04	0.09	0.13	-0.21	-0.13	0.36	0.28	0.13		0.00	0.21	-0.09	0.19	-0.05	0.31	0.42	0.25
-0.10	N	0.03	0.03	-0.10	0.03	0.14	0.17	-0.27	-0.23	0.38	0.25	0.07	0.13		0.32	0.07	0.42	0.31	0.55	0.52	0.37
-0.08	Q	0.06	-0.04	-0.09	0.05	-0.01	0.12	-0.33	-0.17	0.24	0.32	0.10	0.17	0.07		0.53	0.83	0.38	0.51	0.64	0.33
0.31	D	0.22	0.17	0.04	0.16	0.22	0.24	-0.18	-0.12	0.40	0.35	0.18	0.24	0.13	0.19		0.97	0.09	-0.04	0.07	0.46
0.24	E	0.30	0.18	0.07	0.15	0.16	0.25	-0.12	-0.05	0.41	0.49	0.27	0.35	0.24	0.18	0.38		0.39	0.10	0.16	0.48
-0.54	H	-0.22	-0.22	-0.31	-0.09	-0.09	-0.02	-0.52	-0.35	0.21	0.12	0.00	0.05	0.04	0.01	0.01	0.10		0.35	0.70	0.19
-0.12	R	0.01	-0.07	-0.16	0.00	-0.08	0.03	-0.38	-0.25	0.16	0.17	0.07	0.12	0.08	-0.02	-0.13	-0.15	-0.08		1.31	0.52
0.16	K	0.28	0.17	0.05	0.09	0.16	0.23	-0.05	-0.10	0.45	0.41	0.25	0.30	0.11	0.19	0.04	-0.04	0.28	0.27		0.77
0.07	P	-0.02	0.00	-0.14	0.04	0.03	0.06	-0.37	-0.23	0.33	0.27	0.12	0.23	0.15	0.14	0.24	0.32	0.01	0.05	0.36	

Fig. 2.2 Statistical contact potential values for the CD-based definition of contact (upper right triangle and upper row for hetero- and homo-typic interactions, respectively) and the looser any-heavy-based definition (lower left corner and left column for hetero- and homo-typic interactions, respectively). Cells are colored blue to red in ascending order of statistical energies.

Name	Superscript	Description
CD-based	CD	contact degree greater than or equal to 0.1
any-heavy	1	at least one pair of non-hydrogen atoms within 8 Å of each other
Cβ	2	Cβ (or Cα in the case of Glycine) atoms within 8 Å of each other
centroid	3	residue sidechain centroids within 6 Å of each other

https://doi.org/10.1371/journal.pone.0199585.t001

Table 2.1 Contact definitions Definitions of the four types of considered contacts.

2.2.3 Comparison of contact potentials via decoy discrimination

To evaluate the predictive performance of each contact potential, we turned to decoy discrimination. A common benchmark experiment for structure-prediction scoring

functions, it tests whether the correct native (or a native-like) protein structure for a given sequence can be identified from a set that additionally includes incorrect/decoy structures. Specifically, we used two commonly employed decoy sets: the I-TASSER Decoy Set-II generated by the Zhang Lab[47] and the Rosetta decoy set by the Baker Lab[48]. These have been broadly used to test a variety of scoring methods[49–55]. The decoys in these two datasets were generated differently, and therefore represent different test cases for a scoring function. I-TASSER decoys were generated by refining I-TASSER ab initio predictions with the OPLS-AA force field in order to remove clashes and optimize torsion angles. The Rosetta decoys were generated by swapping native backbone dihedral angles with random ones from other native structures, filtering out structures with overly high radii of gyration or those with heavy atom clashes. The I-TASSER set contains 56 proteins, with 300-500 decoys for each, and the Rosetta set has 59 proteins with 100 decoys for each.

For each protein, the native structure and all of its decoys were scored using each potential. To evaluate performance, the rank of the native structure based on its score was determined for each protein in the sets. A rank of 1 means that the native received the most favorable score, whereas higher ranks indicate that some decoy structures scored better than the native. Table 2.2 shows the performance on the I-TASSER Decoy Set-II[56]. Among the four contact potentials considered, E_{CD} assigns the lowest rank to the native structure (or is tied for the lowest rank) in 37 cases, whereas E_1, E_2, and E_3 do so in 4, 10, and 10 cases, respectively. Overall, the ranks assigned by E_{CD} are well below those for all other potentials, and these differences in performance are highly statistically significant (see Table 2.2). Table 2.3 shows the performance on the Rosetta decoy set[48]. In this case, E_{CD} assigns the lowest rank to the native structure (or is tied for the lowest rank) in 27 cases, whereas the same is true for E_1, E_2, and E_3 in 7, 17, and 25 cases, respectively. The Rosetta decoy set appears to be a significantly simpler set than the I-TASSER one for all contact potentials, so differences in performance are less pronounced. Thus, although E_{CD} numerically outperforms all other potentials here as well, the difference is statistically significant only in comparison with E_1, whereas E_2 and E_3 perform similarly to E_{CD} (see Table 2.3).

Name	CD	any-heavy	C_β	centroid	Name	CD	any-heavy	C_β	centroid
1abv_	100	221	366	320	1mkyA3	87	267	234	151
1af7__	13	492	101	101	1mla_2	17	103	194	125
1ah9_	392	450	212	152	1mn8A	196	392	373	503
1aoy_	147	397	474	445	1n0uA4	171	269	266	277
1b4bA	3	322	52	6	1ne3A	76	498	537	503
1b72A	392	486	512	534	1no5A	2	36	2	84
1bm8_	3	208	10	40	1npsA	214	385	363	365
1bq9A	8	389	298	7	1o2fB_	4	248	246	19
1cewI	137	438	359	243	1of9A	1	507	432	31
1cqkA	2	282	23	76	1ogwA_	240	333	243	192
1csp_	220	305	195	255	1orgA	3	65	4	1
1cy5A	48	274	227	249	1pgx_	379	157	452	349
1dcjA_	72	2	289	69	1r69_	17	2	208	110
1di2A_	226	17	225	198	1sfp_	61	309	7	211
1dtjA_	18	284	90	282	1shfA	67	502	335	362
1egxA	83	156	20	13	1sro_	85	476	6	86
1fadA	95	391	337	430	1ten_	11	258	256	219
1fo5A	145	289	235	334	1tfi_	264	234	94	103
1g1cA	32	290	135	35	1thx_	4	228	40	6
1gjxA	32	474	283	256	1tif_	12	422	367	486
1gnuA	10	467	441	238	1tig_	201	478	466	397
1gpt_	56	383	316	343	1vcc_	9	550	414	398
1gyvA	12	229	5	60	256bA	335	445	336	335
1hbkA	172	265	234	178	2a0b_	219	234	221	218
1itpA	376	473	445	250	2cr7A	102	257	101	101
1jnuA	6	236	11	161	2f3nA	274	455	442	274
1kjs_	240	270	176	339	2pcy_	249	324	249	354
1kviA	455	475	298	540	2reb_2	45	91	309	337
Median	79.5	297.5	244.5	228.5					

https://doi.org/10.1371/journal.pone.0199585.t002

Table 2.2 Decoy discrimination on I-TASSER II Decoy-discrimination performance of E_{CD}, E_1, E_2, and E_3 potentials (in columns CD, any-heavy, Cβ, and centroid, respectively) on the I-TASSER II decoy set. Shown is the rank of native structure, in each sub-set, by the corresponding contact potential. The ranking of natives by E_{CD} is significantly better than the rankings using the other potentials, with the p-values from the Friedman test being $7.9 \cdot 10^{-10}$, $1.3 \cdot 10^{-5}$, and $4.5 \cdot 10^{-5}$ when comparing E_{CD} with E_1, E_2, and E_3, respectively.

Name	CD	any-heavy	C_β	centroid	Name	CD	any-heavy	C_β	centroid
1a19	8	26	14	22	1kpe	13	48	10	1
1a32	50	101	92	27	1lis	63	100	15	14
1a68	63	101	35	12	1lou	12	87	27	32
1acf	1	35	10	10	1nps	4	12	11	17
1ail	4	20	3	16	1opd	2	6	6	22
1aiu	61	101	67	64	1pgx	5	1	69	19
1b3a	16	80	48	38	1ptq	7	101	60	10
1bgf	35	76	15	11	1r69	38	1	54	37
1bk2	13	74	13	3	1rnb	1	18	1	22
1bkr	8	39	12	1	1scj	35	30	59	20
1bm8	1	34	10	1	1shf	25	68	22	38
1bq9	18	37	10	9	1ten	1	1	1	1
1c8c	15	49	34	13	1tig	5	48	2	25
1c9o	53	99	36	45	1tul	7	14	10	1
1cc8	29	35	8	17	1ubi	61	84	48	41
1cei	40	12	17	5	1ugh	4	46	33	57
1cg5	29	59	6	15	1urn	2	50	20	2
1ctf	53	1	14	4	1utg	100	101	101	100
1dhn	1	54	6	1	1vcc	6	94	20	9
1e6i	7	96	1	17	1vie	25	40	36	62
1elw	16	1	70	87	1vls	65	62	13	60
1enh	67	93	51	62	1who	1	10	1	1
1ew4	1	22	2	4	256b	62	1	28	76
1eyv	2	17	10	9	2acy	1	13	1	5
1fkb	1	14	4	1	2chf	23	87	36	72
1fna	19	33	27	14	2ci2	8	100	37	73
1gvp	6	76	41	15	2tif	1	1	1	1
1hz6	16	32	10	11	4ubp	1	33	1	1
1ig5	21	27	1	90	5cro	74	55	43	13
1iib	23	94	27	14					
Median	13	40	15	15					

https://doi.org/10.1371/journal.pone.0199585.t003

Table 2.3 Decoy discrimination on Rosetta Decoy-discrimination performance of E_{CD}, E_1, E_2, and E_3 potentials (in columns CD, any-heavy, C_β, and centroid, respectively) on the Rosetta decoy set. Shown is the rank of native structure, in each sub-set, by the corresponding contact potential. The ranking of natives by E_{CD} is significantly better than ranking by the all-heavy potential (E_1), and potentials E_2 and E_3 performing similarly to E_{CD} (Friedman test p-values are 10^{-7}, 0.17, and 0.78, respectively).

Because the only difference between these potentials is the definition of contact (the reference state is kept the same), the above results strongly suggest that CD is a more informative criterion for determining residue interactions. Thus, it would appear to be more advantageous for structural modeling to predict contacts defined via CD than the looser distance-based criterion. To test this claim more directly, we measured the amount of information contributed by each native contact to decoy discrimination. That is, we asked what fraction of decoys are eliminated (on average) by the knowledge of a single

contact in the native structure. We found that for the CD-based definition, an average contact eliminates 64% of the Rosetta decoys whereas this fraction is 48%, 48%, and 63% for the any-heavy-, Cβ-, and centroid-based definitions, respectively. Similarly, on average a CD-based contact eliminates 72% of the I-TASSER decoys compared to 47%, 44%, and 66%, respectively, for the other three contact definitions. This shows that it would be more advantageous, for the purposes of structure prediction, if evolutionary MSA-based methods predicted contacts under the CD-based definition.

2.2.4 Contact prediction using different contact definitions

We next asked how well the more valuable CD-based contacts are predicted from MSAs using the principle of co-evolution. As representative methods, we used 1) the Direct Coupling Analysis (DCA) approach by Morcos *et al*.[4], which has aided a number of structure prediction tasks[57–60]; and 2) MetaPSICOV by Jones *et al.*, a state-of-the-art consensus method that combines three different co-evolution calculations (PSICOV[25], mean-field DCA[61], and CCMpred[62]) with other features (e.g., predicted secondary structure, solvent accessibility, and others) into a neural network. MetaPSICOV has been among the best performers in the contact prediction category of recent CASP competitions[35,63]. In the DCA method, the direct information (DI) metric computed for all position pairs in an MSA is used to order the likelihood that each corresponds to a true contact, with a higher DI indicating a more likely contact. In MetaPSICOV's case, the output of the neural network produces a value between 0 and 1 termed the precision score, with a higher value indicating a more likely contact. Fig. 2.3 shows the performances of DCA and MetaPSICOV in the context of either the CD-based or the looser distance-based definitions of true contact. Shown is the positive predictive value (PPV) as a function of either the number of pairs predicted as contacting (*N*, Fig. 2.3A and 2.3C) or the length-normalized number (i.e., fraction) of predicted contacts (*f*, Fig. 2.3B and 2.3D), respectively.

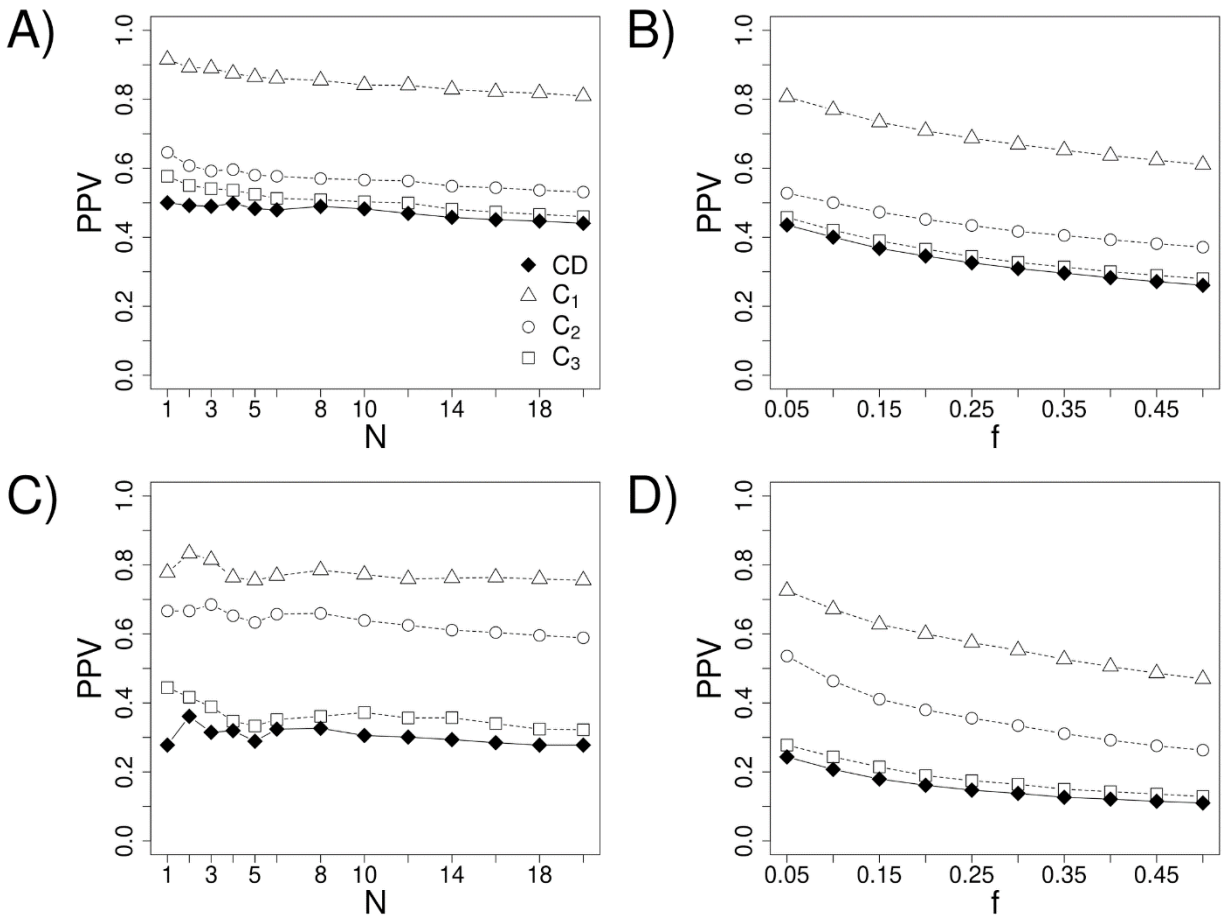


Fig. 2.3 Average PPV of contact prediction as a function of the number (*N*) or fraction (*f*) of predictions. Predictions labeled by *CD* refer to predictions when contacts are defined by contact degree and those labeled by C_1, C_2, and C_3 refer to predictions when contacts are defined by the other three definitions (see Table 2.1 for details). **(A, B)** Predictions of DCA on the Pfam dataset. **(C, D)** Predictions of MetaPSICOV on the CASP12 dataset.

Though different datasets are used to evaluate DCA and MetaPSICOV in Fig. 2.3 (thus, absolute results are not directly comparable between the two; see Methods), in all cases, the performance is lowest with the CD-based contact definition. Thus, although CDs are more informative, they appear harder to predict correctly. In general, unsurprisingly, contacts by looser criteria appear easier to predict. Indeed, ~20%, ~10%, and ~6% of position pairs are classified as contacting by the the any-heavy, Cβ, and centroid definitions, respectively, whereas only ~4% are in contact by the CD-based definition. This is consistent with contact prediction performance monotonically

increasing in the order of CD, centroid, Cβ, and any-heavy contact definitions (see Fig. 2.3). Based on the above contact frequencies, a randomly chosen position pair is, respectively, ~5.0, ~2.5, and ~1.5 times more likely to be a true contact by the any-heavy-, Cβ-, and centroid-based definition than by the CD-based one. On the other hand, the PPV for predicting CD-based contacts is reduced relative to that for other definitions by significantly lower fractions (see Fig. 2.3A). Thus, it would seem that predicting CD-based contacts may still provide more information. Notably, the greatest discrepancies in performance among the different definitions of contact occur for long-range contacts, defined as those with a sequence separation of at least 23. Given that long-range contacts tend to constrain the possible structure more than short-range contacts, these performance discrepancies are particularly important to address.

The above results suggest that contact degree captures useful information about structure, more so than other contact definitions, but the considerably lower precision of predicting it is not desirable, so we next seek ways of improving it.

2.2.5 A statistical contact potential aids in contact prediction

A statistical contact potential provides a convenient line of additional evidence towards predicting contacts, because it quantifies the a priori expectation that any two amino acid types would be in contact. Looking at a particular pair of positions (i, j) in an MSA, we can ask whether the amino-acid pairs found at these positions tend to correspond to favorable or unfavorable contact-potential values. Qualitatively, if the former is the case, this should strengthen our belief that (i, j) is a true contact, while the latter case would weaken this belief. To capture this quantitatively, one could (for example) look at the average value of a contact potential across all amino acid pairs at (i, j) in the MSA, which we will denote $\hat{E}^{i,j}$. This metric could then be used in combination with co-evolution scores (e.g., DI or precision score for DCA or MetaPSICOV, respectively) to make a call about a particular position pair. To test this concept, we propose a simple empirical metric:

$$\hat{S}^{i,j} = S^{i,j}\left(1 - \frac{\hat{E}^{i,j}}{S_{max}}\right)$$

Eq. 2.3

where $S^{i,j}$ is the MSA-based co-evolution score for the position pair (i, j) and S_{max} is the maximal value of the former for any pair of positions in the given alignment. The reasoning behind this combination is that contact potential values are on a fixed scale, whereas we have empirically found co-evolution scores to vary considerably from case to case, depending significantly on the depth and other properties of the MSA. Dividing $\hat{E}^{i,j}$ by S_{max} then serves to normalize the two metrics with respect to each other, across different MSAs. The negative sign in front of $\hat{E}^{i,j}$ reflects the fact that negative potential values correspond to favorable cases and the product ensures that $S^{i,j}$ and $\hat{E}^{i,j}$ jointly contribute towards scoring a potential contact. Note that much more sophisticated combinations of $S^{i,j}$ and $E^{i,j}$ are possible. In fact, MetaPSICOV includes the value of a statistical contact potential as one of the features that go into its neural network model[35]. However, our focus here is to establish and quantify the value of using contact potentials to augment co-evolution scores, under different contact definitions, so we chose a simple functional form for ease of interpretation.

We consider each of the contact definitions discussed above and derive four corresponding augmented S metrics, $S^{i,j}_{CD}$ and $S^{i,j}_{1}$, $S^{i,j}_{2}$, and $S^{i,j}_{3}$. Fig. 2.4 compares the performance of these combined metrics with that of unadjusted S towards predicted the corresponding contact types (i.e., how well $S^{i,j}_{CD}$ predicts CD-based contacts and how well each distance metric predicts the corresponding distance-based contacts). Encouragingly, the PPV for predicting CD-based contacts increases by as much as ~18% and ~12% for the first few predictions using DCA and MetaPSICOV, respectively (Fig. 2.4A and 2.4B). The performance also increases for the distance-based contact definitions (Fig. 2.4C–2.4H). These increases are smaller that with CD-based contacts, with the exception of the centroid definition in conjunction with MetaPSICOV improving PPV by a comparable amount (~14% for the first few contacts). The PPV using the any-heavy definition is close to perfect—over 90% for the first few contacts—but incorporating the

any-heavy potential still systematically improves the performance, demonstrating the general benefit of incorporating a contact potential.

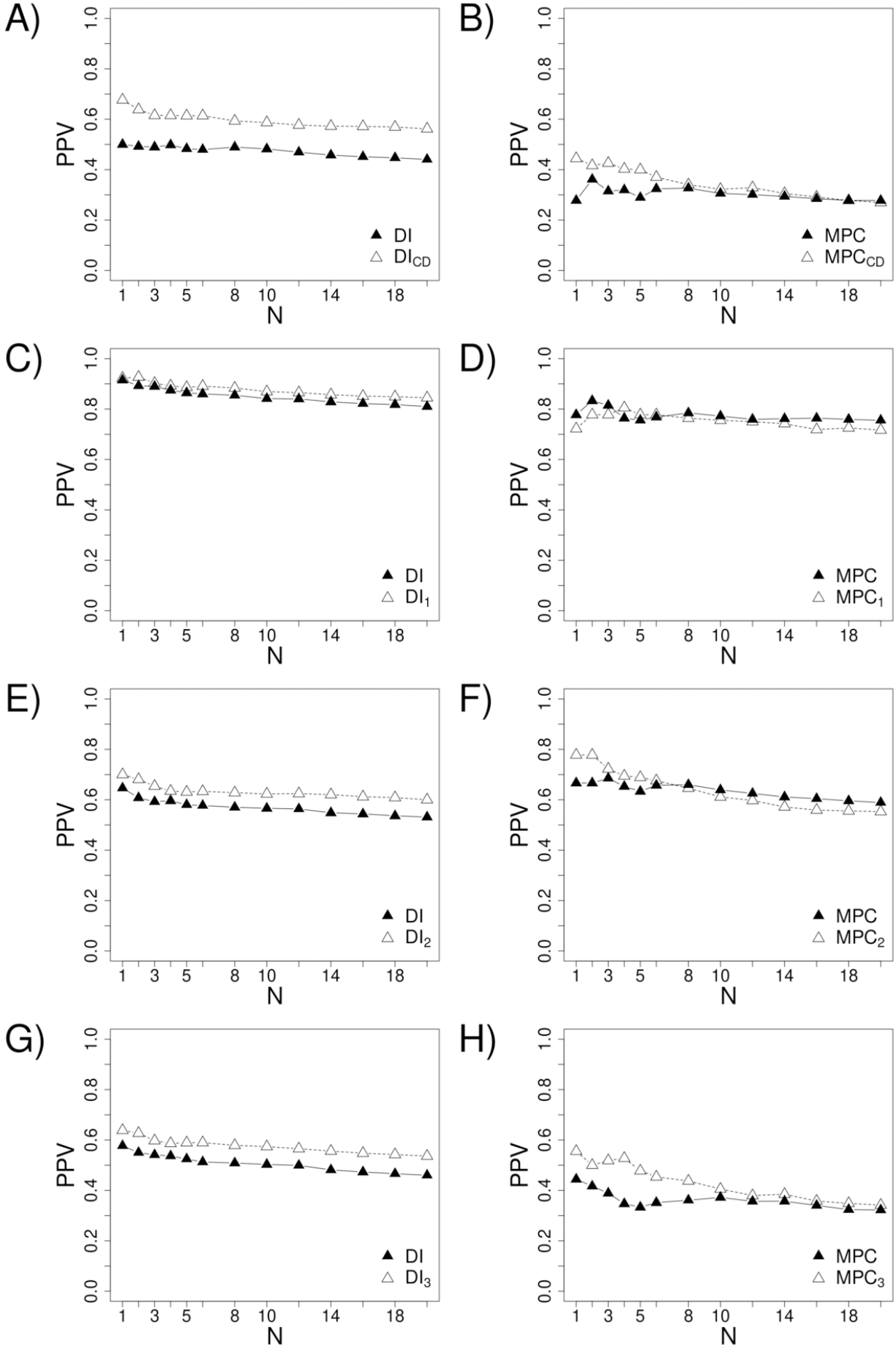

A)
B)
C)
D)
E)
F)
G)
H)
PPV
N
DI
DI_CD
MPC
MPC_CD
DI_1
MPC_1
DI_2
MPC_2
DI_3
MPC_3

Fig. 2.4 The effects of incorporating a contact potential into contact prediction. In plots **A)**, **C)**, **E)**, and **G)**, *DI* refers to predictions made using direct information alone. In plots **B)**, **D)**, **F)**, and **H)**, *MPC* refers to MetaPSICOV's predictions alone. DI_{CD} and MPC_{CD} respectively refer to DI and MPC's predictions augmented by contact degree (see Eq. 2.3). Similarly, for $n \in \{1, 2, 3\}$, DI_n and MPC_n respectively refer to DI and MPC's predictions augmented by contact definition n.

We next ask whether there is benefit in averaging the statistical contact potential values over all sequences of an MSA. That is, we ask whether comparable performance improvements are observed when the contact potential is computed only in the context of a single sequence (e.g., the sequence for which contacts are being predicted). To that end, Fig. 2.5 shows the performance improvement (averaged over five trials) when contact-potential energies are calculated in the context of only a single sequence randomly selected from the corresponding MSA. For DCA applied to the Pfam dataset (see Methods) incorporating these energies systematically improves the PPV (Fig. 2.5A). For MetaPSICOV applied to the CASP12 dataset (see Methods) the improvement is marginal at best (in fact, the performance drops slightly for larger N; Fig. 2.5B). This suggests that averaging contact potential values over the MSA does provide a significant benefit over evaluation in the context of a single sequence (compare Figs. 2.4A and 2.5A). On the other hand, average contact-potential values on their own do not provide sufficient information for effective contact prediction.

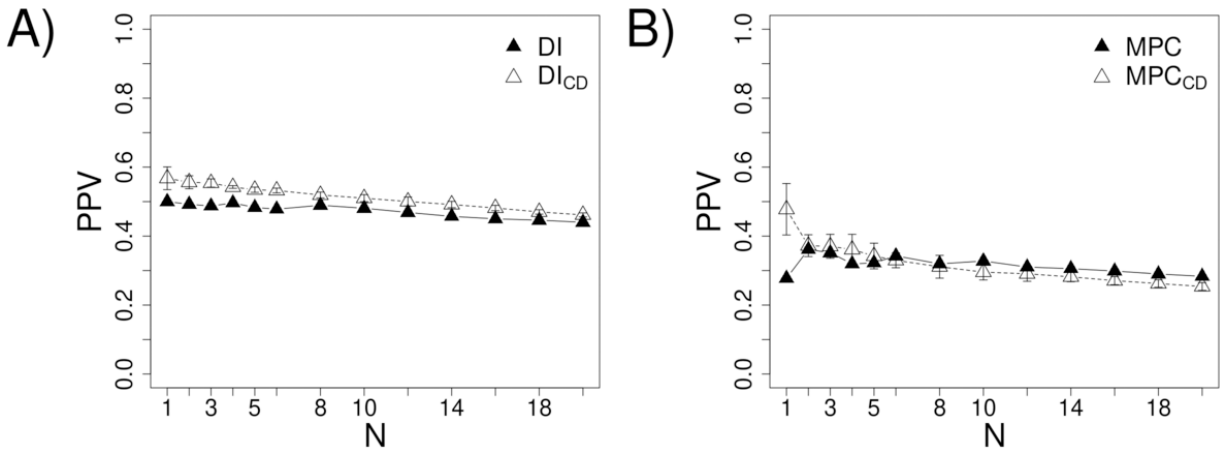


Fig. 2.5 Contact predictions made using **(A)** DCA and **(B)** MetaPSICOV alone are compared against predictions that combine co-evolution scores with the CD-based contact potential energies from a single randomly-chosen sequence in each alignment. This procedure was repeated five times. Each point displayed corresponds to the mean PPV and the error bars show the standard deviation.

We further test how the diversity of predicted contacts changes when different contact potentials are combined with co-evolution scores. Higher contact diversity is desirable because if a method's predicted contacts cover many regions in the contact map, each predicted contact can independently restrain the possible structures the sequence might fold into. To assess contact diversity, we adopted the definition used by He *et al.*, wherein the contact map of each target was divided into a 10 x 10 grid of equal-sized regions and the diversity D was quantified as the Shannon entropy of the distribution of the top N/2 contacts over these regions (where N is the length of the MSA)[64]:

$$D = -\sum_{i}^{100} p_i \log_2 p_i$$

Eq. 2.4

Here, p_i is the fraction of contacts that fall within region i. Table 2.4 shows the mean D over all targets when contacts are either ranked by co-evolution scores alone or by hybrid scores that combine the different contact potentials. Clearly, for both DCA and MetaPSICOV, diversity increases upon adding all contact potentials, but it increases the most when the CD-based contact potential is added.

	alone	with E_{CD}	with E_1	with E_2	with E_3
DCA	3.36	3.67	3.51	3.48	3.61
MetaPSICOV	3.38	3.65	3.54	3.49	3.61

https://doi.org/10.1371/journal.pone.0199585.t004

Table 2.4 Contact diversity The effect of incorporating contact potentials on contact diversity. Contact diversity was quantified by applying Eq 2.4 to the top $N/2$ contacts in

each alignment and then averaging over every alignment in the dataset (first row: DCA on the Pfam dataset; second row: MetaPSICOV on the CASP12 dataset, see Methods), where N is the length of an alignment. The "alone" column contains the diversities when no contact potential is applied (that is, when DCA or MetaPSICOV scores alone are used to rank contacts). The remaining columns contain the diversities resulting from ranking contacts by hybrid scores that combine the corresponding co-evolution score and a contact potential (based on the four contact definitions in Table 2.1, respectively).

2.3 Discussion

In this study we show that contact prediction performance depends critically on the underlying geometric definition of a contact. The previously reported high prediction rates have relied on relatively loose, distance-based definitions of contact. The definitions tested in this study—any heavy atoms within 8 Å, Cβ atoms within 8 Å, and centroid pseudo-atoms within 6 Å– respectively classify ~20%, ~10%, and ~6% of the residue pairs in a protein as contacting. Though this aids in achieving a high positive predictive rates, the looseness comes at the expense of information contributed towards structure prediction. This is evident when comparing these contact definitions to a stricter one we propose, based on the quantity of contact degree (CD, Eq. 2.1). Indeed, only ~4% of position pairs are classified as contacting based on CD (with the cutoff of 0.1 used throughout this study) and a single CD-based contact eliminates 5, 2.5, and 1.5 times more decoy structures than a contact defined by the any-heavy, Cβ, and centroid definitions, respectively. Also, a statistical contact potential corresponding to the CD-based contact definition exhibits a significantly better performance in decoy discrimination than do contact potentials derived from distance-based contact definitions.

Though more informative, CD-based contacts are also harder to predict (see Fig. 2.3). Encouragingly, however, we show that combining the co-evolution score of a given residue pair with the statistical contact potential energy for the pair, averaged over all sequences in the MSA, results in a significantly more predictive metric. The performance boost is particularly pronounced in the prediction of CD-based contacts. For example, the CD-based potential increases the precision of the DCA method by ~18% for the first few

contacts (see Fig. 2.4A). Such a performance increase is highly relevant given that the knowledge of only a few of contacts is often sufficient to aid structure prediction[65].

While the performance improvements were largest for CD-based contacts, incorporating a contact potential improved performance for every definition of contact using both methods, with the exception of the Cβ-based potential not improving the performance of MetaPSICOV. Notably, of the three distance-based contact definitions we have considered, the centroid-based definition exhibits considerable advantages: 1) it performs best (or tied for best) in decoy discrimination (see Tables 2.2 and 2.3), 2) contact-prediction improvement resulting from the incorporation of its corresponding contact potential is the highest (see Fig. 2.4H), 3) it eliminates the highest fraction of decoys based on a single contact, and 4) it leads to the highest contact diversity increase when augmenting a co-evolution score (see Table 2.4). It can be argued that these advantages, to some extent, are a result of the centroid-based definition using more information–i.e., the location of the side-chain. Indeed, side-chains positions must be known (or appropriately modeled) to even apply this definition of a contact. On the other hand, the CD-based definition achieves better performance in all of the above criteria without requiring side-chain information. Possible side-chain positioning is accounted for explicitly within the CD calculation procedure itself, in a sequence independent manner, resulting in a contact definition that can be applied to full-atom or backbone-only models alike.

2.4 Methods

2.4.1 Contact degree

CDs were calculated according to Eq. 2.1 using the 2010 backbone-dependent Dunbrack rotamer library[12]. Rotamers were labeled as clashing with the backbone (and removed from consideration) if at least one non-hydrogen atom in the rotamer sidechain was within 2.0 Å of any non-hydrogen backbone atom of the structure (except its own backbone). ConFind, a program that computes CDs, can be found at http://www.grigoryanlab.org/confind/.

2.4.2 Decoy discrimination

The I-TASSER II decoy set was downloaded from https://zhanglab.ccmb.med.umich.edu/decoys/decoy2.html[56]. The Rosetta decoy set was downloaded from https://zenodo.org/record/48780#.WqAU-HWnFhF[66].

2.4.3 DCA

As described by Morcos *et al*., 131 protein families were selected from Pfam's homologous sequence datasets based on the number of non-redundant sequences, fraction of sequences belonging to bacterial organisms, and the availability of high quality PDB structures[4]. This resulted in 856 corresponding PDB structures. DI for all residue pairs was calculated using Matlab code obtained from Dr. Morcos. To map the 856 PDB structures to their Pfam families, each PDB sequence was compared against all sequences in all of the above Pfam families. To account for point mutations introduced in PDB structures, a sequence-to-structure match was established if the sequence similarity was at least 95%. If no sequence was found to be a match for a particular PDB structure, the sequence that gave the highest sequence similarity score was considered as the match. In this way, each PDB structure in the list was mapped onto at least one of the 131 Pfam families. The MSAs and structures used for this analysis are exactly as those used in the original study, so the results in Fig. 2.3A for the loose contact definition reproduce the PPVs reported in that work.

2.4.4 MetaPSICOV

To evaluate MetaPSICOV's contact prediction, the sequences of each CASP12 target listed in Table 1 in Buchan *et al.* were submitted to the MetaPSICOV server (http://bioinf.cs.ucl.ac.uk/MetaPSICOV/) and the precision scores were extracted from the Stage 2 results[63]. Because not all CASP12 target sequences have publicly available structures, which are needed to determine which pairs of positions are in contact, only those sequences with corresponding PDB entries were considered, resulting in 19 sequences. Each sequence's PDB ID was taken from the CASP website (http://predictioncenter.org/casp12/targetlist.cgi) and the corresponding PDB file was downloaded from the PDB. To acquire the alignments used to produce MetaPSICOV's precision scores, MetaPSICOV was downloaded from

http://bioinfadmin.cs.ucl.ac.uk/downloads/MetaPSICOV/ and run locally. Due to technical difficulties, the alignment for target T0918 could not be computed, resulting in a dataset of 18 sequences: T0859, T0862, T0863, T0864, T0866, T0868, T0869, T0870, T0886, T0892, T0896, T0897, T0898, T0900, T0904, T0941, T0943, T0945.

2.4.5 Contact definitions

Contacts in each structure were identified using either the CD-based metric, with a cutoff of 0.1, or one of the three distance-based metrics specified in Table 2.1, C_1, C_2, and C_3. For C_1—"any-heavy"—a pair of positions was considered in contact if at least one non-hydrogen atom from the residue at one position was less than 8 Å of one non-hydrogen atom from the residue at the other position, backbone atoms included. For C_2—"Cβ"—a pair of positions was considered in contact if the Cβ atom from one position was less than 8 Å from the Cβ atom from the other position. For C_3—"centroid"—a pair of positions was considered in contact if a pseudo-atom located at the mean coordinates of one position's sidechain atoms was less than 6 Å from the corresponding pseudo-atom of the other position. For the Pfam dataset, a pair of positions in an MSA of a protein family was considered to be a true contact if the corresponding pair of positions was in contact within any PDB structure mapped to the family. For the CASP12 dataset, a pair of positions in an MSA was considered to be a true contact if the corresponding pair of positions was in contact in the PDB structure of the target sequence. To enable direct comparison between the results in this paper and those in Morcos *et al.*[4], a contact in the Pfam dataset was treated as a contact only if the two positions were separated in sequence by at least five positions. On the other hand, a contact in the CASP12 dataset was treated as a contact only if the two positions were separated in sequence by at least six positions, in accordance with CASP protocol (see http://predictioncenter.org/casp12/doc/rr_help.html).

2.4.6 Contact prediction

To predict contacts, all residue pairs separated by at least the minimum sequence separation (see the previous paragraph for details) were ranked in descending order of calculated co-evolution scores and top-ranking pairs were predicted as contacting. Top

pairs were selected either based on a fixed rank cutoff (i.e., the first N pairs predicted as contacting for each protein, as in Figs. 2.3A, 2.3C, and 2.4) or a length-normalized rank cutoff (i.e., for a protein of length N, the first $f \times N$ pairs predicted as contacting, with $f \in [0, 1]$, as in Fig. 2.3B and 2.3D). Positive predictive value (PPV) was assessed as the fraction of true contacts out of the predicted contacts. Since the set of true contacts depends on the geometric contact definition, PPV was a function of contact definition.

2.5 Acknowledgments

This chapter is adapted from Holland *et al.*[8]. This work was supported by the National Institutes of Health award P20-GM113132 (GG) and the National Science Foundation award DMR1534246 (GG).

3 Predicting native structures with a hierarchical, geometric, residue-based statistical potential

3.1 Introduction

The predictive success of the contact degree-based contact potential relative to other distance-based contact potentials (Chapter 2) suggests that with the right geometric descriptors, a residue-level statistical potential can encode valuable information about sequence-structure relationships. While the contact potential developed in Holland *et al.*[8] was deliberately simple, designed to highlight the efficacy of its underlying definition of contact, it seemed feasible to develop a more intricate version which sought to more fully describe the geometry of contacts and how it affects sequence preferences. The most successful statistical potentials condition their statistics on many geometric descriptors, such as distances and orientations, and such conditioning is essential to differentiate between various kinds of interactions and their varying sequence preferences.

In this project, I sought to construct a hierarchical statistical potential, conditioning the contact degree-based pair terms on self terms describing the geometry and environment of the residues involved in the contacts. That is, rather than estimate pair preferences

solely in the context of a database of contacting residue pairs, this statistical potential was based on the idea that to estimate pair preferences, the role of self preferences must be addressed first, adjusting the expectations of the pair terms by how often both amino acids would be expected based on their self preferences. More specifically, the first term in the potential was defined to be the φ- and ψ-dihedral angle preferences of each residue in the database of contacts. Having associated each bin of φ/ψ-space with its preferences, the ω-dihedral angle preferences were conditioned on these, estimating the favorability of each amino acid in each bin of ω-space given the already known φ/ψ preferences. The motivation to estimate preferences hierarchically like this was to avoid the data sparsity issues inherent in high-dimensional spaces. If dihedral angle space is jointly partitioned finely enough to accurately estimate preferences, the number of residues occupying all of these particular regions simultaneously falls precipitously. By first estimating the statistics of each bin of φ/ψ-space and then estimating the statistics of each bin of ω-space given each residue's φ/ψ preferences, there is no need for every region of this 3-dimensional space to be adequately populated—if residues with ω-dihedral angles in a particular region never have φ/ψ-dihedral angles in another particular region, then there is no need to estimate the statistics in this combined φ/ψ/ω region.

On top of these φ/ψ/ω backbone dihedral angle preferences, the preference for each amino acid in varying environments, encoded by a backbone-based, rotamer library-dependent metric we call "freedom", were estimated. Thus, this hierarchical potential encodes four layers of self preferences. The final layer of the potential is the pair preferences, encoded by partitioning contact degree into many bins and estimating the amino-acid pair preferences within each. For each contact considered, the self preferences of both of its constituent residues were considered using the four self preference layers and the expectation of the pair preferences was adjusted based on these. Thus, the preferences in each bin of contact degree reflect the extent to which each amino-acid pair is preferred given the known preferences of the residues involved in contacts in this bin.

The primary question being asked here was how much information about native sequence-structure preferences could be stored in a residue-level potential, and in particular, whether this information was sufficient to identify native structures among

decoys, a common challenge asked of energy functions that probes the limits of their predictions.

3.2 Results

3.2.1 Database and geometric descriptors

As with the contact potential developed in Chapter 2, a large database of non-redundant structures from the PDB was collected using PISCES[46]. For each structure in this database, the contact degree (Eq. 2.1) between each pair of residues was calculated. In addition, for each residue of each contact, the backbone dihedral angles (φ, ψ, and ω) were calculated as well as the "freedom" of each residue.

Just as contact degree measures the potential interaction between a pair of residues using only the backbone coordinates of the residues and a rotamer library, freedom measures the lack of crowdedness of a residue's environment using its backbone coordinates and a rotamer library. In particular, the freedom of a residue r is computed as:

$$F(r) = \frac{\sqrt{\frac{n_1^2(r) + n_2^2(r)}{2}}}{n(r)}$$

Eq. 3.1

Here, $n(r)$ is the number of rotamers of any amino acid that can be attached to the backbone of r (since the rotamer library is backbone-dependent, this number is variable). The terms $n_1(r)$ and $n_2(r)$ are the number of rotamers that pass the first or second collision threshold, respectively. Each rotamer is attached to r and then tested for clashes with contacting rotamers. If any non-hydrogen atom of the rotamer comes within 3 Å of any non-hydrogen atom of a rotamer t attached to a contacting residue, then it is considered clashing and the probability of t (according to the rotamer library) is added to the collision sum of the rotamer. If the collision sum of a rotamer is under 2.0, then 1 is added to $n_2(r)$. If the collision sum of a rotamer is under 0.5, then 1 is also added to $n_1(r)$.

Thus, $n_1(r)$ and $n_2(r)$ store how many rotamers are sufficiently "free" from other, contacting rotamers. This makes freedom a side-chain independent measure of a residue's (lack of) crowdedness, estimating how free or buried it is.

3.2.2 φ/ψ potential

To measure the preference of each region of φ/ψ space for each amino acid, the 2-dimensional space was divided into bins of 10° x 10°. For each bin and each amino acid, the number of times the amino acid was observed in the bin was calculated and divided by the total number of observations (of any amino acid) in that bin. Thus, the energy of an amino acid in a given φ/ψ bin measures the relative preference for it compared to other amino acids in the bin. Given the heterogeneous density of φ/ψ space, a small pseudocount was added to the numerator and denominator to avoid division of or by zero:

$$\mathrm{E}_{\varphi\psi}\left(a_i, b_k\right) = -\log \frac{\mathrm{N}_{\mathrm{obs}}\left(a_i, b_k\right) + \mathrm{f}\left(a_i\right) \cdot \varepsilon}{\mathrm{N}_{\mathrm{obs}}\left(b_k\right) + \varepsilon}$$

Eq. 3.2

Here, $\mathrm{N}_{\mathrm{obs}}(a_i, b_k)$ is the number of observations of amino acid a_i in bin b_k and $\mathrm{N}_{\mathrm{obs}}(b_k)$ is the total number of observations in bin b_k. Note that the pseudocount in the numerator is multiplied by $\mathrm{f}(a_i)$, the background frequency of a_i in the database as a whole. This ensures that in the absence of data in b_k, the energy reflects the background frequency. Furthermore, if b_k has no particular preference for any amino acid, the denominator ensures that the energy of each amino acid reflects is background frequency (since $\mathrm{N}_{\mathrm{obs}}(a_i, b_k) / \mathrm{N}_{\mathrm{obs}}(b_k)$ would then approximate $\mathrm{f}(a_i)$). This first potential thereby encodes the relative frequencies of each amino acid, refined by their relative frequencies in each region of φ/ψ space when data are available. Fig. 3.1 shows a few examples of what these energies look like.

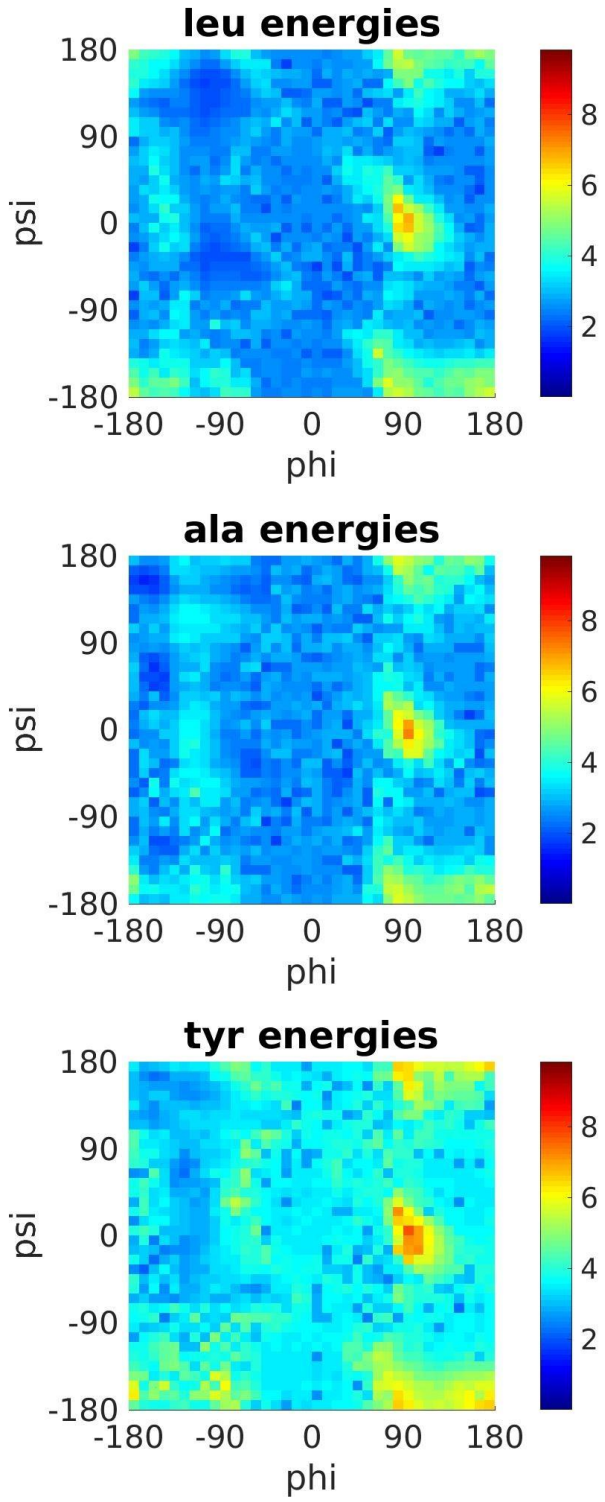


Fig. 3.1 Examples of φ/ψ energies for three amino acids, alanine, tyrosine, and leucine.

3.2.3 ω potential

As discussed, the ω potential is conditioned on the φ/ψ potential, adjusting the expectation of each amino acid in each bin by the φ/ψ preferences of each residue under consideration. Given the extreme sparsity of some regions of ω space (due to strict steric constraints), bins were apportioned not by fixed degrees but instead so that each bin had approximately the same number of observations. Within each bin, the relative favorability of each amino acid a_i was calculated given how often it should be expected given the φ/ψ preferences of each residue of type a_i:

$$\mathrm{E}_{\omega}\left(a_i, b_k\right) = -\log \frac{\mathrm{N}_{\mathrm{obs}}\left(a_i, b_k\right) + \mathrm{f}\left(a_i\right) \cdot \varepsilon}{\mathrm{N}_{\mathrm{exp}}\left(a_i, b_k\right) + \varepsilon}$$

$$\mathrm{N}_{\mathrm{exp}}\left(a_i, b_k\right) = \sum_{r \in R} \frac{\exp\left(-\mathrm{E}_{\varphi\psi}\left(a_i, \mathrm{b}_{\varphi\psi}(r)\right)\right)}{\sum_{j=1}^{20} \exp\left(-\mathrm{E}_{\varphi\psi}\left(a_j, \mathrm{b}_{\varphi\psi}(r)\right)\right)}$$

Eq. 3.3

Here, $\mathrm{b}_{\varphi\psi}(r)$ is the φ/ψ bin of residue r and R is the set of residues in ω bin b_k. What this means is that the expected number of observations, in the absence of sequence-structure relationships (i.e., the chosen reference state), is the sum of the φ/ψ energies of each residue of type a_i in b_k compared to the sum of the energies of every residue in b_k. For instance, if alanine residues in ω bin b_k have on average very favorable φ/ψ energies then the expected number of observations is increased accordingly. $\mathrm{E}_{\omega}(a_i, b_k)$ therefore represents how favorable a_i is in b_k conditioned on how favorable it is based φ/ψ energies. The total energy of a residue r of type a_r using these two potentials is then the sum $\mathrm{E}_{\varphi\psi}(a_r, \mathrm{b}_{\varphi\psi}(r)) + \mathrm{E}_{\omega}(a_i, \mathrm{b}_{\omega}(r))$. Fig. 3.2 shows examples of these energies using the same amino acids as in Fig. 3.1.

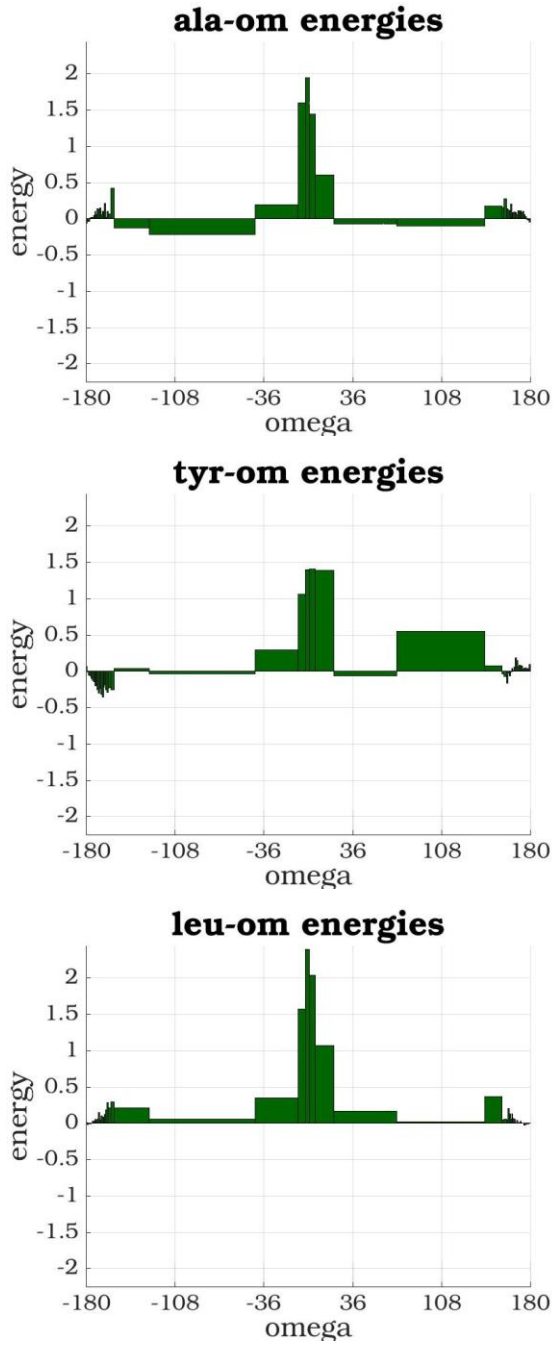


Fig. 3.2 Examples of ω energies for three amino acids, alanine, tyrosine, and leucine.

3.2.4 Freedom potential

The freedom potential is conditioned on the two previous potentials in the same way the ω potential was conditioned on the φ/ψ potential. That is, the expectation is a function of

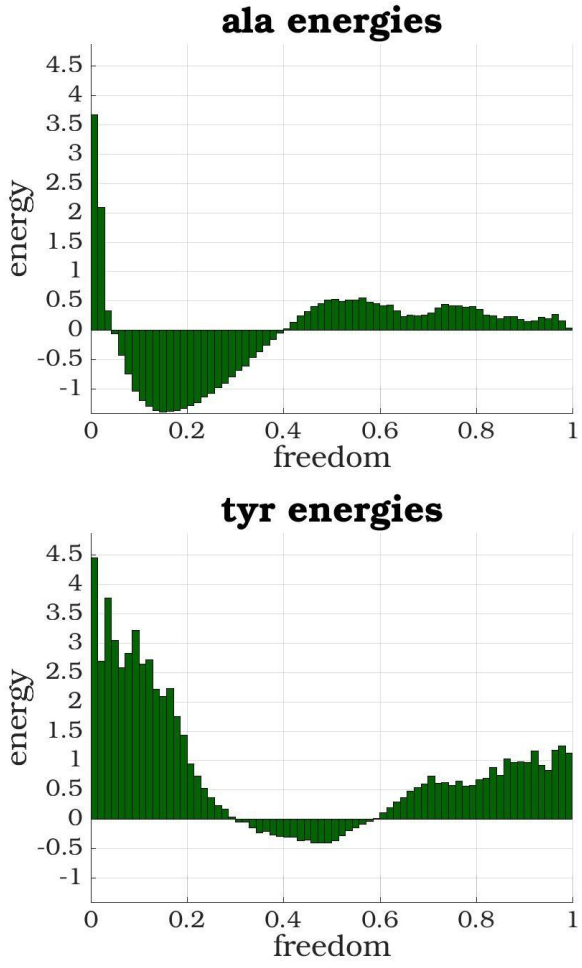
how favorable each residue under consideration is given the previously computed energies. Specifically, the expectation is calculated as:

$$\mathrm{E}_{\mathrm{free}}\left(a_i,\, b_k\right) = -\log \frac{\mathrm{N}_{\mathrm{obs}}\left(a_i,\, b_k\right) + \mathrm{f}\left(a_i\right)\cdot\varepsilon}{\mathrm{N}_{\mathrm{exp}}\left(a_i,\, b_k\right) + \varepsilon}$$

$$\mathrm{N}_{\mathrm{exp}}\left(a_i,\, b_k\right) = \sum_{r\,\in\, R} \frac{\exp\left(-\left(\mathrm{E}_{\varphi\psi}\left(\mathrm{a_i},\, \mathrm{b}_{\varphi\psi}(\mathrm{r})\right) + E_{\omega}\left(\mathrm{a_i},\, \mathrm{b}_{\omega}(\mathrm{r})\right)\right)\right)}{\sum_{j=1}^{20}\exp\left(-\left(\mathrm{E}_{\varphi\psi}\left(\mathrm{a_j},\, \mathrm{b}_{\varphi\psi}(\mathrm{r})\right) + E_{\omega}\left(\mathrm{a_j},\, \mathrm{b}_{\omega}(\mathrm{r})\right)\right)\right)}$$

Eq. 3.4

Fig. 3.3 shows examples of these energies using the same amino acids as in Figs. 3.1 and 3.2.



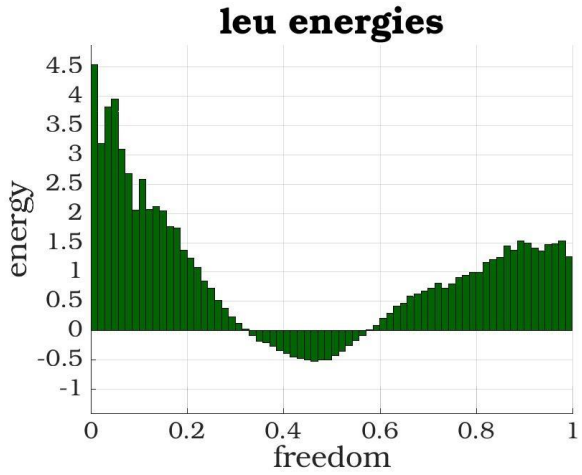


Fig. 3.3 Examples of freedom energies for three amino acids, alanine, tyrosine, and leucine.

3.2.5 Contact degree potential

The pair potential, estimating the preference for each amino acid pair in each bin of contact degree, conditions the expectation on all four computed self potentials. This is done in the same way as shown above but over each pair of residues in each contact degree bin, calculating how many observations should be expected in the reference state based on the self preferences. Because this potential is symmetric—$E_{CD}(a, b) = E_{CD}(b, a)$ by construction—the expectation for heterotypic pairs must be effectively doubled by considering both "directions", (a, b) and (b, a), since both directions contribute to the number of observations:

$$E_{CD}(a_i, a_j, b_k) = -\log \frac{N_{obs}(a_i, a_j, b_k) + f(a_i) \cdot f(a_j) \cdot \varepsilon}{N_{exp}(a_i, a_j, b_k) + \varepsilon}$$

$$N_{exp}(a_i, a_j, b_k) = \sum_{r_1, r_2 \in R \times R} \frac{N_{exp}(a_i, a_j, b_k, r_1, r_2) + (1 - \mathbb{I}(i, j)) \cdot N_{exp}(a_j, a_i, b_k, r_1, r_2)}{\sum_{m=1}^{20} \sum_{\kappa=1}^{20} N_{exp}(a_m, a_\kappa, b_k, r_1, r_2)}$$

$$N_{exp}(a_i, a_j, b_k, r_1, r_2) = \exp\left(-\left(E_{\varphi\psi}(a_i, b_{\varphi\psi}(r_1)) + E_{\omega}(a_i, b_{\omega}(r_1)) + E_{free}(a_i, b_{free}(r_1)) + E_{\varphi\psi}(a_j, b_{\varphi\psi}(r_2)) + E_{\omega}(a_j, b_{\omega}(r_2)) + E_{free}(a_j, b_{free}(r_2))\right)\right)$$

Eq. 3.5

Fig. 3.4 shows examples of pair energies for a few different amino acid pairs.

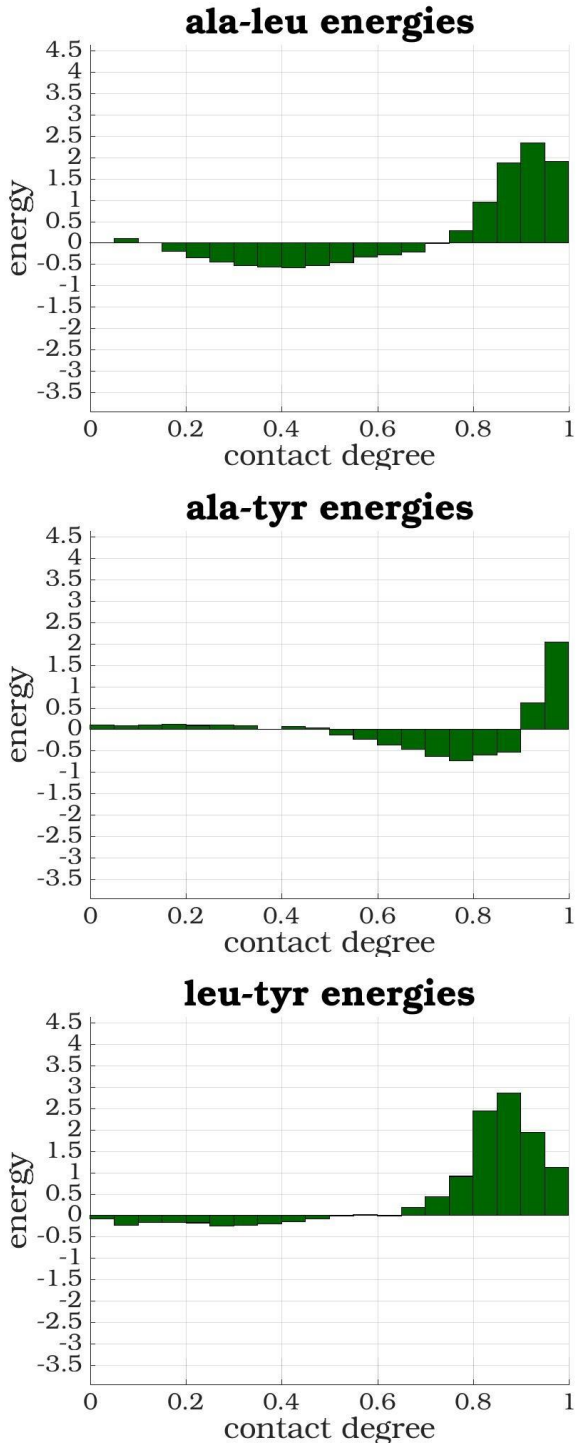


Fig. 3.4 Examples of contact degree energies for three amino acid pairs, alanine-leucine, alanine-tyrosine, and leucine-tyrosine.

3.2.6 Decoy discrimination

To test how much information is contained in this hierarchical potential, we collected several sets of decoy sets and performed decoy discrimination. Each decoy set comprised a set of native structures and hundreds of decoys, created to have the same sequence as the corresponding native but a different conformation. Each decoy set created decoys in a distinct way so that the collection of all of them contained a large diversity of decoy structures. If an energy function can accurately relate sequence and structure then it should be able to evaluate the compatibility of each structure's sequence and structure, scoring the native the most favorably and the decoys less so because they have, by construction, the wrong conformations for their sequences.

Each structure was scored by computing the contact degree between each pair of residues and then scoring the contact using the sum of the potentials. That is, for each contact, the amino acid pair was scored based on the bin its contact degree fell in, and both amino acids were scored using the four self potentials. The total score of the contact was the sum of all of these energies and the total score of the structure was the sum of the contact scores.

To put the performance in a larger context, we evaluated the same decoy sets using DFIRE[67] and FoldX[68], both of which are well-validated atomic-level energy functions, and compared the performance of this hierarchical potential to theirs. Table 3.1 shows a summary of the results, counting how many times each energy function correctly identified the native structure as the most favorable compared to its decoys.

Decoy set	**Hierarchical potential**	**DFIRE**	**FoldX**	**Max possible**
i-tasser[47]	43	44	53	56
moulder[69]	15	19	19	20
hg_structal[70]	24	12	16	29
ig_structal[70]	43	0	10	61
4-state[70]	7	6	7	7

fisa[70]	1	3	2	4
lmds[70]	6	7	8	10
lattice[70]	7	8	8	8

Table 3.1 Decoy discrimination comparison Decoy discrimination results of the hierarchical potential compared to DFIRE and FoldX. The first column lists each set of decoy sets. The middle three columns list the number of natives correctly identified by each method. The final column lists the number of native structures in each decoy set, which is the maximum possible number that could have been identified.

As can be seen, the hierarchical potential performs comparably to DFIRE and FoldX, sometimes outperforming and sometimes underperforming them. For almost all sets of decoy sets (i-tasser, moulder, etc.), the hierarchical potential identified the native correctly the majority of the time, suggesting it contains substantial information about native sequence-structure relationships. This is particularly striking because it is a residue-level potential and uses only information from the backbone, not the side-chains, which DFIRE and FoldX both take advantage of. While this project was not published, it did provide evidence that residue-level potentials could compete with atomic-level ones given the right geometric descriptors. The structure-conditioned contact potential in Chapter 5 can be seen as an extension of this, even conditioning pair energies on the same self terms (backbone dihedral angles and freedom), albeit in a slightly different formulation.

4 An interactive tool for building novel protein backbones

4.1 Introduction

The set of proteins evolved by natural selection offers an extremely large and diverse repertoire of functional macromolecules including enzymes, switches, clamps, sensors,

small molecule transports, and nanoscale morphological structures such as channels, vesicles, and elastic fibers. Yet natural proteins span only a miniscule set of points in the space of all designable structures—structures for which there is at least one amino acid sequence that folds into them—a space brimming with the possibilities of novel structures which lead to novel functions. Protein scientists have designed many novel structures over the last several decades[71–75], many of which comprise a *de novo* backbone—one which does not have a known counterpart in nature. Such *de novo* structures, while often taking inspiration and even parts from natural proteins, are unconstrained by the sequences, structures, and functions nature has happened to discover and are instead limited only by what their designers can come up with.

While many techniques for creating novel backbones have been presented[76–78,71,79–83,73,84,74,85], the problem remains challenging and there are many regions of designable structure space that have never been explored. One of the central problems underlying the creation of novel structures is the back-and-forth between sequence design and backbone creation needed to find a sequence energetically optimal for the backbone and a backbone optimal for the sequence. That is, the optimal sequence for a given backbone may itself be optimal for a different backbone, which in turn may have a different optimal sequence. This means that even if the desired function is known exactly, structures that can achieve this function must be discovered dynamically, which creates a need for rapid structure generation and modification as part of the design process. This is difficult to achieve experimentally, as the biophysical characterization of protein structures is time-consuming, expensive, and not guaranteed to work. These experimental difficulties have encouraged the development of faster feedback loops via computational predictions. While many developments have focused on sequence-dependent predictions, such as modeling tools that predict the most likely structure given the sequence and energy functions that predict the energetic stability of a structure given its sequence[67,47,86,87], there have also been many advancements in rapid structure generation itself, such as Blueprint Builder[81] and TopoBuilder[88].

While the fastest computational feedback loops often sacrifice accuracy for speed, their responsiveness can make them invaluable tools for rapidly testing many hypotheses, with the most promising hypotheses then subjected to more rigorous but time-consuming

tests, ultimately leading to experimental characterizations and tests of function. In the case of *de novo* backbone creation, it would be helpful to have some confidence in the designability of a backbone as quickly as possible during creation so that the focus can remain on backbones likely to be designable instead of the vast expanse of undesignability which comprises most of structure space. Tertiary motifs—structural fragments centered around likely inter-residue interactions—provide a promising framework for designability, as it has been shown that only a small number of them is required to cover a large portion of structure space[7]. A reflection of the degeneracy of structure space, the high coverage these motifs provide means a relatively small library of them could provide an expressive selection of structural building blocks.

Furthermore, given the myriad exciting possibilities for novel structures and functions, it would also be helpful to leverage human creativity and visual intuition in the process of creation. Considering jointly the desiderata of designability and visual intuition, an interactive application that allows a user to create any backbone they can imagine, but with their attention focused on only those backbones most likely to be designable, would offer protein designers a fruitful and interesting way to propose new structures. Motivated by this idea and inspired by other interactive protein structure tools such as Foldit[89] and Suns[90], we present a tool, Protein Builder, that offers its user a way to interactively create backbones by assembling native fragments piece-by-piece in ways empirically known to occur, fusing each fragment with the rest of the assembly after each step. This leads to a design process that ensures native-like interactions locally without constraining the user to globally native geometries. The abilities to add and remove fragments and undo/redo actions, among other useful features, facilitates rapid testing of structural hypotheses and can be used to create entirely new structures whose inter-residue interactions are often known to occur in native structures. We expect Protein Builder to serve as a useful complement to other protein design tools in the creation of novel structures.

4.2 Results

4.2.1 Motivation and overview

As outlined in the introduction, Protein Builder is designed to satisfy a number of criteria. First, the user should be able to build structures piece-by-piece to create a novel backbone. To achieve this, we compiled a database of representative structural fragments from the PDB comprising a combination of linear and discontinuous tertiary and quaternary motifs. We then created a database of overlaps specifying how each fragment in the database is known to spatially and topologically overlap with each other fragment (including itself). This enables the user to assemble a backbone one fragment at a time, with each added fragment known to be consistent with the fragment it overlaps (the first fragment can be chosen arbitrarily). While the backbone can be assembled in a linear fashion, adding an overlapping fragment that is discontinuous (i.e., comprises multiple disconnected segments) can create multi-segment assembles. Such assemblies can be left multi-segment, resulting in a complex with multiple chains, but two or more segments can also be bridged, resulting in a single segment, by adding fragments that overlap with multiple segments simultaneously. This allows for interaction-centric design strategies which involve the addition and bridging of many segments in order to satisfy particular contacts or geometries. Second, the user should be able to create a single, coherent backbone out of the assembly of fragments. We do this through a process we call *fusion*, which finds a backbone (the "fused structure") that best satisfies the geometries of the underlying fragments. Third, the user should be able to choose where to extend the assembly and which (known-to-overlap) fragment should be added. This is achieved through the user interface, which allows the user to select one or more residues of the fused structure from which to extend the backbone, and the search process, which uses the overlap database to filter the possible fragments, only returning to the user candidates that are known to spatially and topologically overlap with at least one of the fragments in the assembly underlying the selected residue(s). Putting this all together, the user is able to assemble a new backbone fragment-by-fragment, choosing where and how the fragments are added but constrained by known fragment overlaps, and receive a fused backbone structure at each step of the assembly.

While the above outlines the general design process, there are other features included to achieve additional goals. One, while Protein Builder is capable of building structures from scratch, it can also start from existing structures by breaking them down

into an assembly of fragments from the fragment database, a process we term *alphabetization*. In combination with the ability to remove residues from the fused structures (by removing the underlying fragments), users can redesign interfaces or any other part of an existing structure. The alphabetization feature can also be used on the fused structure itself, which generates a richer underlying fragment assembly from which more overlaps can be found during the search process. Two, because of the pseudo-discrete nature of designable structure space, only a small number of fragments is needed to cover most parts of most native structures to a high degree of accuracy, but it takes an enormous number of fragments to completely cover every part of every structure. Thus, no matter how many representative fragments are included in the database, there will always be some structural configurations that are difficult to achieve with only a set of discrete components. To address this, we included a bridging feature that, given a gapped pair of residues, searches a database of full structures for fragments that bridge the gap. These bridging fragments can be chosen just like overlapping fragments and are useful for closing loops. Three, in order to aid in the rapid hypothesis testing we believe essential to creating new structures, each step of the build process is marked as a distinct state, and undo/redo buttons can be used to flip through the states. When the user saves their session, all of the states are saved, preserving the entire build process, not just the current fused structure.

It is worth emphasizing that the contact-centric, discontinuous nature of most of the fragments greatly impacts the build process. For instance, searching for a fragment to extend an alpha helix often brings in candidates that not only extend the current helix, but introduce a neighboring one (or two) as well. The spatial information contained in the overlaps ensures that new segments interact with existing ones in reasonable ways and encourages users to build backbones around interactions rather than in a linear N-terminus-to-C-terminus fashion. The combination of introducing new segments via contact-centric fragments and joining multiple segments into one with linear fragments and bridging leads to interesting and unexpected directions of structure extension, with intermediate structures often comprising many segments which can ultimately be linked into a single structure by “clicking” or “slotting” fragments into any remaining gaps.

Below, each of the listed features and how they fit into the creation process are

described in more depth. For technical details, see Methods.

4.2.2 Fragment database

To collect the database of fragments, PISCES[46] was used to select a non-redundant subset of structures from the PDB. This set of structures (StructDB) was then decomposed into fragments of varying topologies (Fig. 4.1). Because interactions are central to structure design, most of the chosen topologies comprise multiple segments centered around a contact, but linear fragments are important for tasks like closing loops and extending termini, so linear topologies were also included. In particular, six topologies were used to decompose the structures into fragments: 5-mers, 7-mers, 9-mers, 3x3-mers, 5x5-mers, and 7x7-mers. For the 5-, 7-, and 9-mers, all contiguous stretches of residues of length 5, 7, and 9, respectively, were extracted from StructDB. The 3x3-, 5x5-, and 7x7-mers are each a topology centered around a contact and comprising the contacting residues and their flanking residues. A 3x3-mer comprises a pair of contacting residues and one residue on each side of both contacting residues (i.e., two segments of three residues each, with the middle residues in contact). A 5x5-mer includes an additional flanking residue on each side of both contacting residues, and a 7x7-mer includes two additional residues.

For each topology, the set of fragments found in StructDB was then clustered using an in-house greedy method. This method subsamples the set of fragments, marks the fragment with the lowest average best-fit root-mean-square deviation (RMSD) to the other sampled fragments as the cluster representative, and then forms a cluster by including every fragment (not just those subsampled) within a chosen best-fit RMSD cutoff to the cluster representative. These fragments are then removed from consideration and the remaining fragments are subjected to the same procedure. For 5-mers and 3x3-mers, this was repeated until 90% of fragments were clustered. For the other topologies, this was repeated until 1000 clusters were generated. The fragment database was defined as every cluster representative from every topology for a total of around 16,000 fragments.

To make it easier to create a structure from scratch, the top 6,000 fragments from a previous study (Top6000), which sought to cover as much of native structure space

using as few fragments as possible, were added to the fragment database, forming the final database used (FragDB). These 6,000 fragments are mostly higher order, comprising many segments, and often form cores or other complex assemblies. They serve as useful starting fragments or bulk extensions.

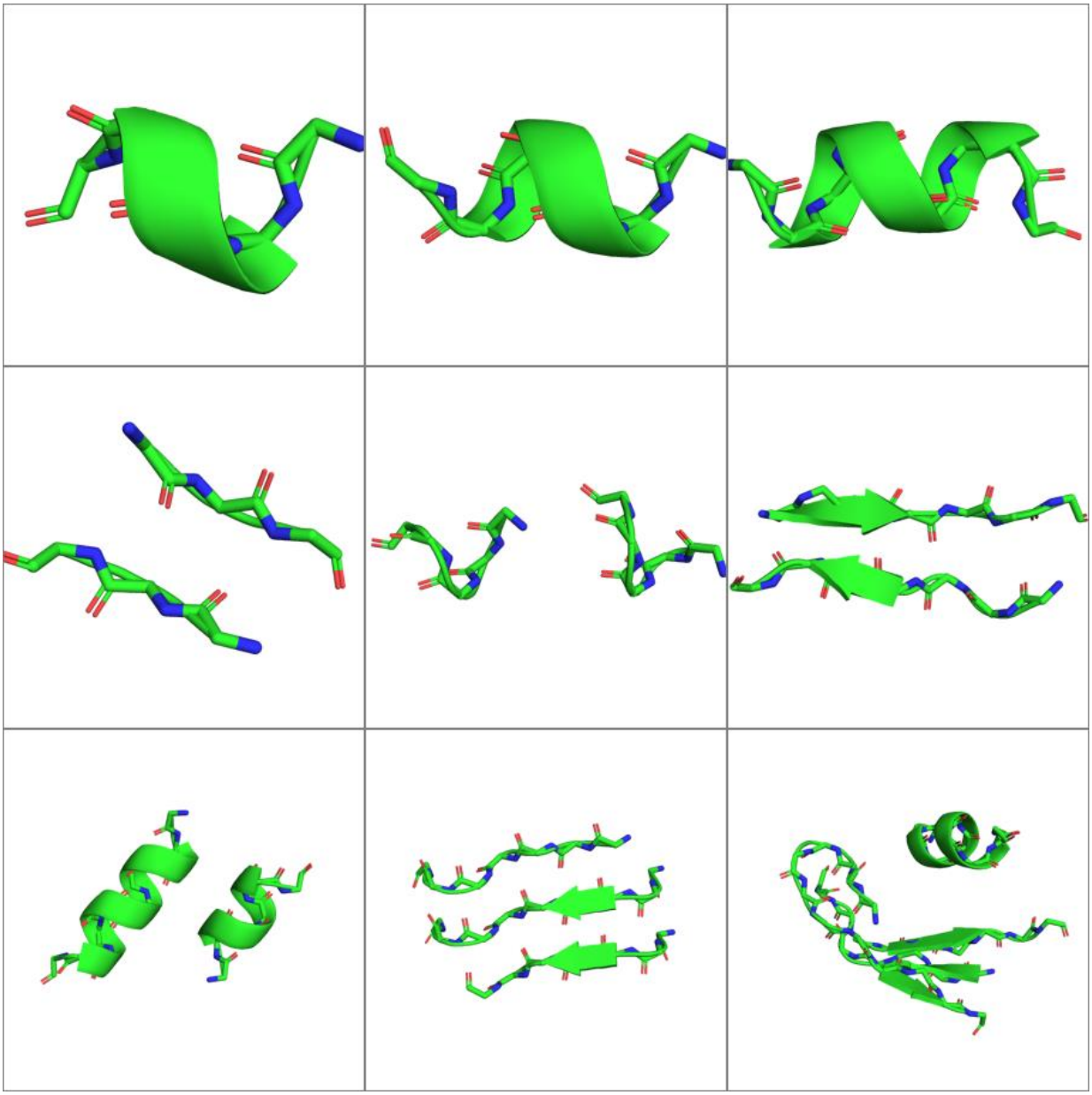


Fig. 4.1 Examples of fragment topologies. From Top-to-bottom, left-to-right: 5-mer, 7-mer, 9-mer, 3x3-mer, 5x5-mer, 7x7-mer, and three higher order fragments from Top6000.

4.2.3 Overlap database

In order to determine how fragments from FragDB can be put together, we used StructDB to empirically discover which fragments are known to overlap. Using a structural search tool similar to MASTER[6], each fragment in FragDB was used as a query and StructDB was used as the database of structures to find matches in. Each match was recorded and, for each pair of fragments in FragDB, the respective pairs of matches that came from the same structure and shared at least *N* residues were examined. If, for a pair of matches, the two fragments, aligned based on their matches, had a sufficiently low in-place RMSD, they were marked as overlapping. Both the transformation matrix describing one match's translation and rotation relative to the other and the overlapping residue indices were stored with the overlap (Fig. 4.2). The overlap database (OverlapDB) was defined as the set of all overlaps between all pairs of fragments and specifies every allowed way for fragments to be placed on top of each other.

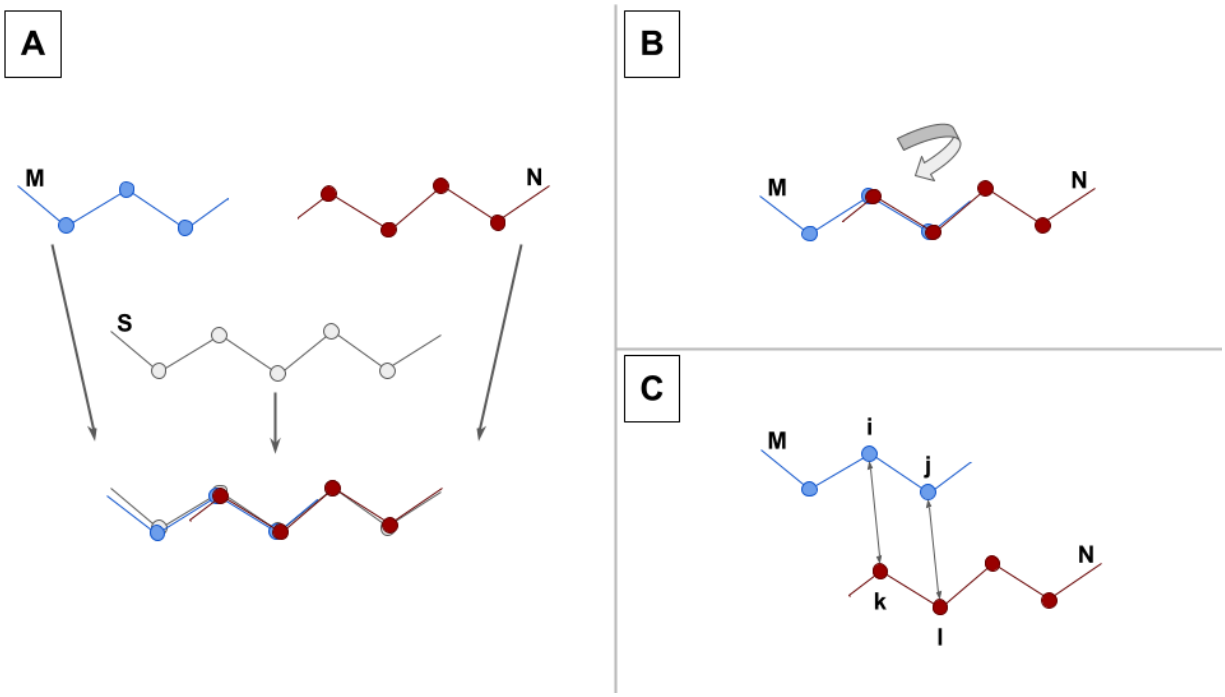


Fig. 4.2 How fragment overlaps are determined and stored. **A**: Fragment *M* and *N* from FragDB both fit onto structure *S* from StructDB and thus are considered overlapping. **B**: The transformation matrix encoding the translation and rotation required to optimally align the overlapping residues of *M* to *N* when fit onto *S* is stored. **C**: The topological mapping

between *M* and *N* when fit onto *S* is stored. Here, residue *i* from *M* corresponds to residue *k* from *N* and residue *j* from *M* corresponds residue *l* from *N*.

4.2.4 Creation process and operations

With StructDB, FragDB, and OverlapDB in hand, Protein Builder can perform all of its operations. The creation process can start from scratch, in which case the user begins by selecting a starting fragment from FragDB. This fragment is the initial fused structure. The process can also start with an existing structure, which gets alphabetized into an assembly of fragments from FragDB, and then fused using this assembly.

Search: Whether starting from an initial fragment or an alphabetized structure, the user advances by selecting one or more residues to extend the backbone structure from. The overlaps of fragments in the assembly underlying the selected residue(s), taken from OverlapDB, are examined and those that match the criteria are returned to the user to select from. The exact criteria depend on the settings chosen by the user, but in the general case, an overlapping fragment must overlap a fragment underlying each of the selected residues (the selected residues can be satisfied by a single overlap or multiple consistent ones), be geometrically consistent with each of the overlaps involved, and be free of clashes with neighboring backbone atoms. All possible overlaps are found at once, but the user receives only the specified number at a time, with paging buttons available to see the rest.

Fragment placement: When the user selects an overlapping fragment to add, the fragment is added to the assembly and then the entire assembly is fused (Fig. 4.3). The new fused structure is then sent to the user, replacing the old one. The user can then select residues of this fused structure to add additional overlaps to, repeating the process. Finding bridging fragments works similarly to finding overlapping ones, with the user selecting at least two gapped residues to bridge. The fragment comprising these selected residues is searched for in StructDB but with the gaps required to be filled so that the returned fragments are each a single segment overlapping the selected residues (geometrically via an RMSD cutoff, not via OverlapDB) and bridging the gap.

Bridges: Just as with overlapping fragments, all bridging fragments are found at once, but the user receives only a specified number at a time. When the user selects a

bridging fragment to add, it is added to the assembly, the assembly is fused, and the new fused structure is returned.

Residue removal: When the user selects one or more residues to remove, any fragment underlying the residue(s) is removed from the assembly, the assembly is then fused, and the new fused structure is returned.

Realphabetization: When the user realphabetizes the fused structure, each fragment from FragDB is used as the query and the fused structure is searched for all matches. Each time a fragment can be mapped to the fused structure with a sufficiently low RMSD, it is added to the assembly. After all fragments have been added, the assembly is fused, and the new fused structure is returned.

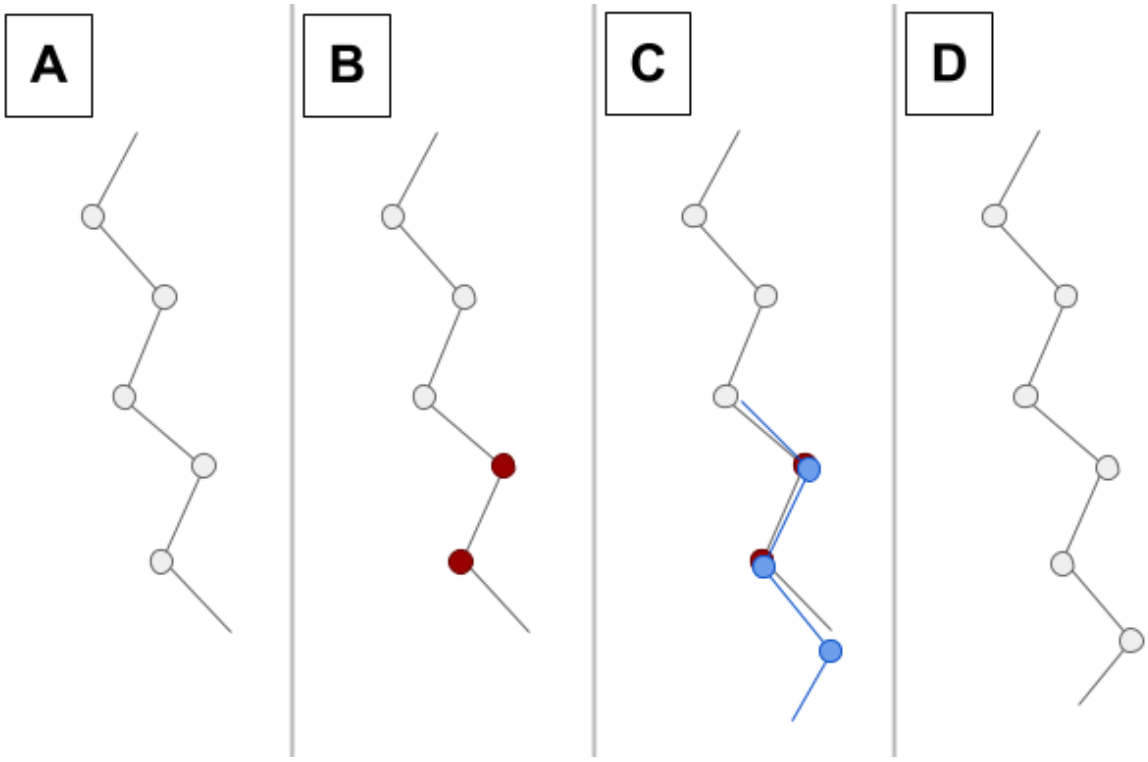


Fig. 4.3 How a fragment is chosen and incorporated into the fused structure. **A**: The current fused structure. **B**: The user selects two residues (red) to find overlaps with. **C**: After choosing from the search results, the user selects an overlapping fragment to add to the assembly (blue). **D**: The new fused structure after fusing the blue fragment with the rest of the assembly.

4.2.5 User interface

The user interface for Protein Builder is a PyMOL[91] plugin (Fig. 4.4). PyMOL's built-in selection ability is used to select the fused structure's residues and a window provides buttons for each feature (search, bridge, undo/redo, etc.), widgets to configure the settings, and a panel that displays matches (starting, overlapping, or bridging fragments). When the user hovers over a match, a transparent version of the fragment is shown where it will be placed, helping the user predict the effect of adding the fragment. Clicking the match places a solid version of the fragment color-coded to indicate its topology, and double clicking the match adds it to the assembly. The plugin serves as the client; the server is a separate program, implemented as a Flask server[92], which can be run locally or over the internet. It contains the databases and performs the needed calculations (search, fusion, etc.), allowing the client to be a lightweight plugin with no dependencies outside of what PyMOL already requires.

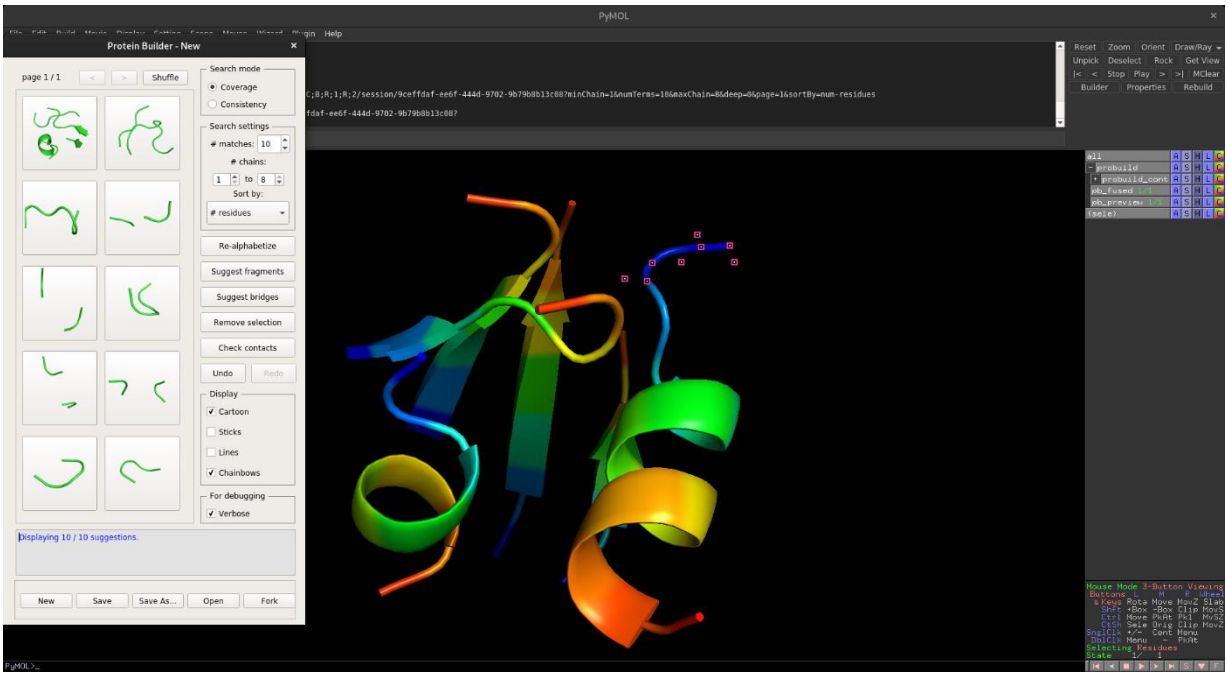


Fig. 4.4 The user interface of Protein Builder.

4.2.6 Examples of de novo backbones built with Protein Builder

As a simple proof of principle, Fig. 4.5 shows a collection of novel backbones

created with Protein Builder. As can be seen, the structures are diverse and encompass a range of shapes, secondary structure elements, and topologies. The question at this point is what else can be imagined and created.

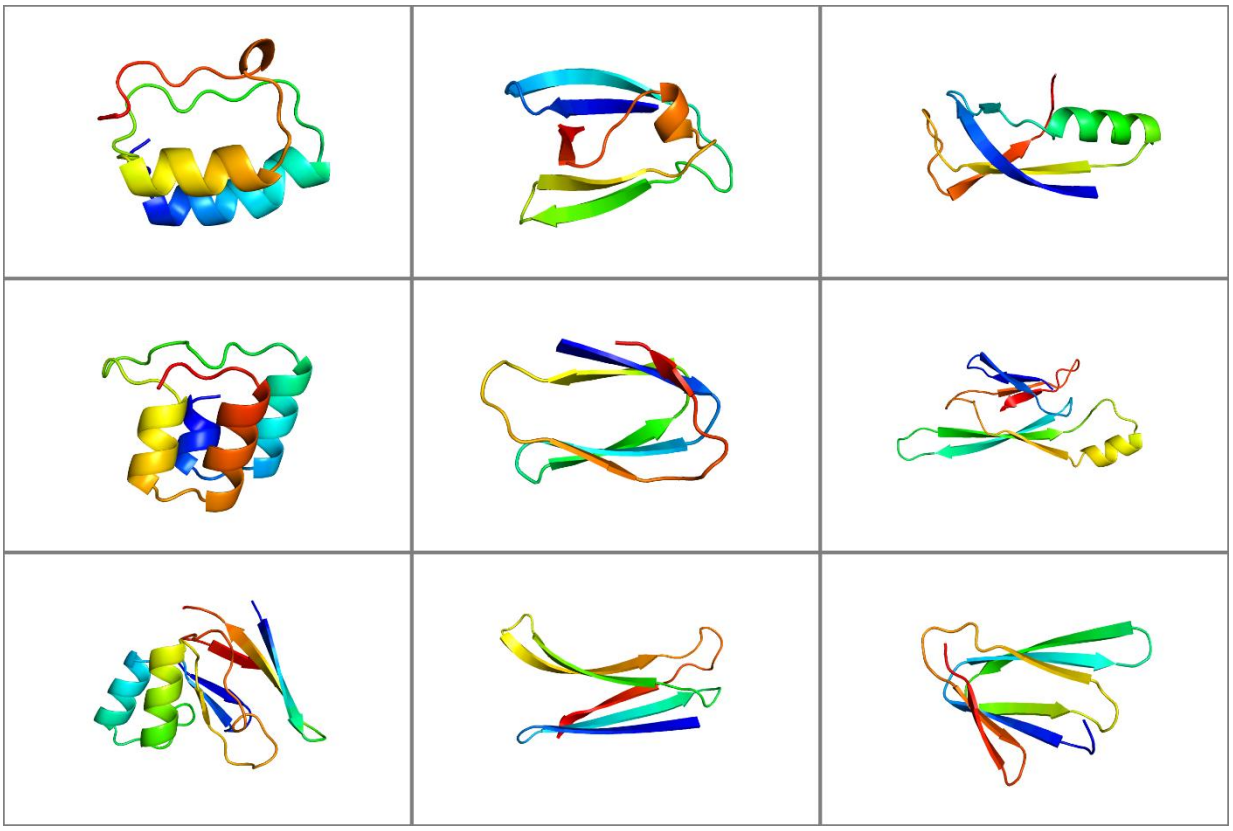


Fig. 4.5 Examples of novel backbones created with Protein Builder.

4.3 Methods

4.3.1 Clustering

The RMSD cutoff chosen for cluster membership was 0.5 Å for linear fragments (5-mer, 7-mer, 9-mer) and 1.3 Å for tertiary fragments (3x3-mer, 5x5-mer, 7x7-mer). In all cases, the number of fragments subsampled per round was 1,000.

4.3.2 Overlap database

Each fragment in FragDB was used as a query in a FASST search using StructDB as the database. A match was accepted if the RMSD was within the cutoff specified by the

formula used by dTERMen (see Eqs. 25 and 26 in the supplementary information of Zhou[9]) using a maximum RMSD of 1.1 Å and a persistence length of 15. The structure and residues involved in the match were then stored.

For each pair of fragments (*f*, *g*) in FragDB, the two sets of matches (M_f, M_g) were compared. Each instance in which a match $m_f \in M_f$ and a match $m_g \in M_g$ came from the same structure *s* in StructDB and shared at least one residue, and in which m_f or m_g included at least one residue not included in the other's match (i.e., one extended the other), was considered an overlap. The transformation matrix encoding the translation and rotation required for *g* to structurally align onto *s* (i.e., the transformation associated with the best-fit RMSD of *g* onto the backbone of *s*) was recorded so that when the user searches for overlaps with *f* (i.e., searches for overlaps with residues in the fused structure which involve an instance of *f* in the underlying assembly), *g* can be optimally aligned with respect to it when added to the assembly. The shared residues between f and g were also stored so that when the user searches for overlaps with *f*, *g* is only returned if its overlap involves the residues of *f* selected by the user. For each overlap found, the overlap of *g* with respect to *f* and of *f* with respect to *g* are stored in the database since the overlaps are bidirectional. Since *f* and *g* can overlap in multiple structures and/or using distinct sets of shared residues, there can be many overlaps between *f* and *g*.

4.3.3 Searching for overlaps

In any given state of creation, there is a fused structure *s* and an assembly *A* comprising a set of fragments which the fused structure derives from. For each fragment $f \in A$, each of its residues corresponds to some residue in *s* ("covers" some residues in *s*), and for each residue in *s*, there is at least one fragment in *A* which covers it. When the user selects one or more residues and clicks the "search" button, each of the fragments in *A* which cover the selected residues are collected in a set *C*. For each fragment $f \in C$, each of its overlaps is considered. For an overlap to be accepted as a match, it must pass five filters.

First: It must have an appropriate number of segments given the minimum and maximum specified by the user (via the graphical user interface).

Second: Its topology must be consistent with *s*. Linear fragments are always

consistent, but a multi-segment fragment may overlap in a way that would require a residue in *s* to play the role of two residues in the overlap, which is not allowed (e.g., if an overlap extends a pair of helices, but in *s* the helices are already joined together by a loop). For a match comprising multiple overlaps (see "Fifth" below), the topology of all of the overlaps must be consistent (e.g., if one overlap's topology specifies that residue *i* in the fragment maps to residue *j* in *s*, then any other overlap of the fragment involving *i* must also map to residue *j*).

Third: The RMSD between the residues in *f* that overlap with the residues in *s* must be within the cutoff determined by the maximum configured RMSD, *m*. The RMSD cutoff is determined by the same empirical formula used in the creation of the overlap database (see the "Overlap database" section above), with the maximum RMSD set to *m*, and is defined in Eqs. 25 and 26 in the supplementary information of Zhou *et al.*[9].

Fourth: The residues in *f* that do not overlap with *s* must not clash with nearby residues. A clash is defined as a non-hydrogen backbone atom in *f* coming within 2 Å of any non-hydrogen backbone atom in *s*, excluding residues in *s* that overlap with *f* or are sequential neighbors of a residue that does.

Fifth: If some of the residues selected by the user are covered by an overlap of *f* with *g* and other residues are covered by an overlap of *f'* with *g* (where *f* may be the same as *f'*), then *g* is accepted as a match if each overlap satisfies the first four filters. Any number of overlaps can be combined in this manner, allowing large selections to come from overlaps with any number of fragments in the assembly. If a match involves multiple overlaps with *g*, each overlap contributes an instance of *g* to the assembly, with each transformation matrix coming from the respective overlap.

4.3.4 Fragment assembly and backbone topology

When a new fragment *f* is placed into the assembly by the user (by selecting a match upon searching or bridging), the topology of the fused backbone *s* must be updated so that each residue in each fragment of the assembly is given the appropriate index based on its position in *s*. As discussed in the "Overlap database" section above, for each overlap in the database between fragments *g* and *h*, there is a mapping between the overlapping residues of *g* and *h*, specifying for each overlapping residue in *g* the

corresponding residue in *h*. Additionally, the residues of the fragment(s) that *f* overlaps with in the assembly have already been given an index in the topology since the topology is updated after each fragment is placed. Therefore, updating the topology upon the placement of *f* is a matter of assigning its overlapping residues the same indices as the residues they overlap with and determining whether its non-overlapping residues correspond to existing residues in *s* or new ones, in which case new indices must be assigned. If every residue of *f* overlaps with a residue in *s*, then *f* is simply added to the assembly, mapping each of its residues to its corresponding residues in *s*. If a residue *r* in *f* is not part of its overlap(s), there are two possibilities. One possible case is that based on the topology of *f*, *r* corresponds to an existing residue in *s*. For instance, if *f* is a single segment of four residues, three of which are known overlap with three non-terminal residues in a segment of *s*, then the fourth residue in *f* must correspond to an existing residue in *s* (the residue just before/after the overlapping ones). In this case, *r* is given the same topological index as the residue in *s* that it corresponds to. The other possible case is that *r* does not correspond to an existing residue in *s*. For instance, if *f* is again a single segment of four residues, but this time its three most N-terminal residues overlaps with the three most C-terminal residues in a segment of *s*, then the fourth residue in *f* (the C-terminal one) does not correspond to any residue in *s*—in fact, this residue extends *s*. In this case, a new topological index is created to accommodate it, which means adding a residue to the C-terminal end of the overlapping segment of *s*. Note that *f* may extend one or more existing segments in *s* and/or add new segments. If *f* overlaps with two segments and joins them together (e.g., by closing a loop), the two segments will be merged into one before fusion.

4.3.5 Fusing the fragment assembly

Fusion seeks to find a backbone that minimizes the average in-place RMSD to each fragment in the assembly while obeying known bond lengths, angles, and dihedral angles. It does so by local optimization, using gradient descent to minimize the following function:

$$\mathrm{F}(s,A) = \sum_{f \in A} \mathrm{RMSD}(s_f, f)^2 + c_l \sum_{a_1,a_2 \in s} \mathrm{distPen}(a_1,a_2) + c_a \sum_{a_1,a_2,a_3 \in s} \mathrm{anglePen}(a_1,a_2,a_3) + c_d \sum_{a_1,a_2,a_3,a_4 \in s} \mathrm{dihedPen}(a_1,a_2,a_3,a_4)$$

Eq. 4.1

Here, s is the fused structure and A is the set of fragments in the assembly. s_f is the set of residues of s that overlap with f so that RMSD(s_f, f) is the best-fit RMSD between f and the part of the fused structure it corresponds to. The force constants c_l=10, c_a=0.02, and c_d=0.001 weight the bond length, angle, and dihedral angle penalties, respectively. All penalties—distPen, anglePen, and dihedPen—are harmonic, penalizing bond lengths, angles, and dihedral angles, respectively, by the square of the deviation if they lie outside of the accepted ranges. These penalties are applied to all backbone atoms except oxygen. distPen is applied to all pairs of sequentially neighboring atoms (i.e., N-Cα of residue 1, Cα-C of residue 1, C-N of residues 1 and 2, etc.), anglePen is applied to all triples of sequentially neighboring atoms, and dihedPen is applied to all quadruplets of sequentially neighboring atoms. The acceptable ranges are determined by examining the distances, angles, and dihedral angles of the input fragments so that, e.g., the minimum N-Cα distance considered acceptable for the (N, Cα) atom pair in residue r of s is set to the minimum N-Cα distance found in the (N, Cα) atom pairs of the fragments that overlap r. Formally, the penalty terms are defined as follows:

$$\mathrm{distPen}(a_1,a_2) = \begin{cases} 0 & \text{if } b_{min} \le \mathrm{dist}(a_1,a_2) \le b_{max} \\ \left(\mathrm{dist}(a_1,a_2) - b_{min}\right)^2 & \text{if } \mathrm{dist}(a_1,a_2) < b_{min} \\ \left(\mathrm{dist}(a_1,a_2) - b_{max}\right)^2 & \text{if } \mathrm{dist}(a_1,a_2) > b_{max} \end{cases}$$

$$\mathrm{anglePen}(a_1,a_2,a_3) = \begin{cases} 0 & \text{if } n_{min} \le \mathrm{angle}(a_1,a_2,a_3) \le n_{max} \\ \left(\mathrm{angle}(a_1,a_2,a_3) - n_{min}\right)^2 & \text{if } \mathrm{angle}(a_1,a_2,a_3) < n_{min} \\ \left(\mathrm{angle}(a_1,a_2,a_3) - n_{max}\right)^2 & \text{if } \mathrm{angle}(a_1,a_2,a_3) > n_{max} \end{cases}$$

$$\mathrm{dihedPen}(a_1,a_2,a_3,a_4) = \begin{cases} 0 & \text{if } \mathrm{ccwDiff}(d_{min},d) \ge \mathrm{ccwDiff}(d_{max},d) \\ \left(2\pi - \mathrm{ccwDiff}(d_{max},d)\right)^2 & \text{if } \mathrm{ccwDiff}(d_{min},d) > 2\pi - \mathrm{ccwDiff}(d_{max},d) \\ \mathrm{ccwDiff}(d_{min},d)^2 & \text{if } \mathrm{ccwDiff}(d_{min},d) \le 2\pi - \mathrm{ccwDiff}(d_{max},d) \end{cases}$$

Eq. 4.2

Above, dist(a_1, a_2) is the Euclidean distance in Ångstroms between atoms a_1 and a_2, angle(a_1, a_2, a_3) is the angle in radians between atoms a_1, a_2, and a_3, and d is short for dihed(a_1, a_2, a_3, a_4), the dihedral angle in radians between atoms a_1, a_2, a_3, and a_4. The constants b_{min} and b_{max} are the minimum and maximum observed bond lengths across all input fragments, n_{min} and n_{max} are the minimum and maximum observed bond angles, and d_{min} and d_{max} are the minimum and maximum observed bond dihedral angles. ccwDiff(d_1, d_2) is the counter-clockwise difference between dihedral angles d_1 and d_2, which can be computed as follows (with % representing the modulo operation):

$$\text{ccwDiff}\left(d_1, d_2\right) = \left(\left(d_1 \,\%\, 2\pi\right) - \left(d_2 \,\%\, 2\pi\right)\right) \,\%\, 2\pi$$

Eq. 4.3

Thus, if a given bond length is beyond the observed range of lengths, its deviation from this range is harmonically penalized. Similarly, if a given bond angle is beyond the observed range of angles, its deviation from this range is harmonically penalized. Because dihedral angles lie in a circular space, calculating the deviation a given dihedral angle may have from the observed range of dihedral angles requires considering the counterclockwise difference from the minimum (via ccwDiff(d_{min}, d)) and the clockwise difference from the maximum (via 2π-ccwDiff(d_{max}, d)); if there is a deviation, it is harmonically penalized.

The optimization is over the Cartesian coordinates of the fused backbone atoms, halting after either 100 iterations or when the difference in RMSD between the previous iteration and the current one is less than 10^{-4} Å.

4.3.6 Alphabetization

Alphabetization transforms an arbitrary input structure into an assembly of fragments from FragDB and then fuses this assembly as it would any other, enabling the user to start from any structure they prefer while still taking advantage of the database of fragments and overlaps. The assembly is created by using each fragment f in FragDB as a query in a FASST search whose database comprises only the input structure. Each match

for the query *f* thus corresponds to a site on the input structure onto which *f* can be placed. Whether the fragment fits well enough is determined by the configured RMSD cutoff. The configuration file enables each fragment topology (5-mer, 3x3-mer, etc.) to have its own cutoff, with the defaults being 0.5 Å for linear fragments and 1.3 Å for tertiary fragments, identical to the cutoffs used to cluster them. Topologies not specified in the configuration are given cutoffs of 1.3 Å. The resulting assembly is the set of all instances of all fragments in FragDB which can be placed onto the input structure according to the specified RMSD cutoffs. This assembly is then fused to become the new fused structure.

Realphabetization works identically to alphabetization, using the fused structure as the input. Note that after realphabetization, an entirely new set of fragments comprise the assembly, not necessarily the ones the user originally selected (because other fragments may have shifted the fused structure so that a previously valid overlap no longer falls within the RMSD cutoff, or because the configured RMSD cutoff for alphabetization is chosen to be stricter than that for overlaps).

4.3.7 Session state management

A user's session is stored as a sequence of states, allowing the "undo" and "redo" buttons to flip between them. A new state is created upon each of the following operations: when the user starts a session; when the user requests starting fragments; upon searching for overlapping or bridging fragments; upon shuffling the suggested starting, overlapping, or bridging fragments upon alphabetizing a new structure; upon realphabetizing the fused structure; upon placing a new fragment in the assembly; and upon removing a residue from the fused structure. When the user clicks the "save" button, the information needed to reproduce all of the states is saved (i.e., the entire session is saved). When the user clicks the "open" button and selects a previously saved session, the file is uploaded to the server and a new session is created with all of the same state information, which obviates requiring the server to save sessions indefinitely. The maximum number of states can be set with the "num_save_states" option in the configuration. When the maximum number of states has been exceeded, the earliest are overwritten.

4.3.8 Implementing the client as a PyMOL plugin

The client application is a set of python files which can be loaded as a plugin for PyMOL 2.0+. The window containing search results, operation buttons, and configurations was designed with PyQt5 and communication with the server was performed with the Requests library. Both PyQt5 and Requests come pre-installed with PyMOL so the plugin does not require any additional dependencies. Two parameters must be defined by the user: the host URL and port number. If the server is running locally, these can be set to 127.0.0.1 and 5000, respectively. Otherwise, the host should point to the address running the server application and whichever port it is listening on. On a Unix machine, these options can be set by running the interactive "configure" script. On a non-Unix machine, these options can be manually set by settings the contents of the "config" file to the following two lines: "host = X" and "port = Y".

4.3.9 Implementing the server as a Flask application

The server application uses Flask 1.1.2 to receive communications from clients, with additional Python 3.9 code to manage sessions, and Boost.Python 1.75 to interface with C++11 code, which does the actual computations (search, fusion, etc.). The server is designed to run on Unix machines and can be configured by running the interactive "configure" script. After configuration, running "make libs" will use gcc (tested with version 11.2.1) to make the shared object containing the required C++ code. After compilation, running the "scripts/startServer" script will start the flask server, which can be run as a background process.

4.4 Acknowledgements

This chapter is adapted from a paper that will soon be uploaded to bioRxiv.

5 Structure-conditioned amino-acid couplings: how contact geometry affects pairwise sequence preferences

5.1 Introduction

For many decades now, the wealth of structural information in the Protein Data Bank (PDB) has enabled protein scientists to infer relationships between the amino-acid sequence of a protein and its native structure based on statistical patterns. A classic example of how even simple structural statistics can provide useful information is the contact potential, which encodes the relative interaction preferences for each amino-acid pair based on simple log-odds ratios of observed versus expected occurrences in a large set of contacts, sometimes partitioning the set of contacts into bins according to distance or other geometric parameters[1,42,50,67,93–99]. Contact potentials in varying forms have provided insight into sequence-structure relationships since the 1970s and have been incorporated into numerous effective predictive models over the years (e.g., Rosetta[48,86], RaptorX[100], PoPMuSiC[101,102], and many others[9,103–106]). The continued efficacy of contact potentials, despite their apparent simplicity, suggests that elaborations or extensions to the core concept may also serve as useful bridges between native sequence and structure elements. Multiple extensions have already been proposed, seeking to condition amino-acid pair preferences on more detailed structural circumstances[107,67,50]. For example, there are potentials that incorporate the relative orientation between residue pairs[108–113], condition on residue depth to capture the effects of polarity and hydrophobicity[114,115], include additional terms from pseudo-physical force fields[116], alter the definition of contact[117,118], and optimize parameters by contrasting the statistics of native structures and decoys[110,119].

Given the complex geometry of an inter-residue contact and its surrounding structural context (e.g., a pair of contacting residues and their flanking residues comprises 24 backbone atoms and thus 72 spatial coordinates), there is a potentially large and high dimensional interaction space throughout which amino-acid pair preferences might vary. In order to explore how pairwise sequence preferences might depend on interaction geometry, a more general formulation of structure conditioning is required. Such a formulation should be able to quantify the pairwise sequence preferences of any type of interaction. Just as a contact potential quantifies which amino-acid pairs prefer to interact over a set of contacts in general, a structure-conditioned potential (SCP) could quantify which amino-acid pairs prefer to interact over a particular set of structurally similar

contacts. Here, we define such a potential as one that takes an additional argument, an input fragment centered around a pair of contacting residues, which determines the contact geometry that the resulting statistics are conditioned on. In particular, a given input fragment is used as the query motif of a structural search which returns an ensemble of structurally similar motifs whose amino-acid pair statistics are then used to compute the statistical energies. By conditioning on an ensemble of similar interaction motifs, there is no need to determine which geometric parameters (distances, orientations, etc.) are best suited for capturing interaction preferences, instead letting the statistics of the PDB inform the preferences via an ensemble of motifs. This process is made feasible not just by the growing size of the PDB but by the recent availability of structural search tools[6], which enable us to specify a query motif (i.e., a structural fragment) and efficiently search for all structurally similar fragments in a database of structures.

The primary purpose of this work is to establish a framework for understanding modular sequence-structure relationships on a pairwise level. First, we show that structure-conditioned coupling energies (SCEs, 20x20 per motif) converge to similar energies as those encoded in a traditional contact potential when many interaction instances are averaged. Having established this link, we then show that SCE matrices encode more information linking the structures of pair motifs to their associated amino-acid pairs and better reflect experimentally determined inter-residue coupling values. Looking more deeply into the link these SCE matrices provide between structure and sequence, we show that structurally similar pair motifs are more likely to be energetically similar and the reverse, that energetically similar pair motifs are also more likely to be structurally similar. This link reveals the modular and context sensitive link between sequence and structure elements and we demonstrate a general relationship between the two. Turning to structure modeling, we compare how well SCEs can evaluate the structural quality of CASP models to how well contact potentials do so, highlighting how much additional information is encoded by conditioning on structure. As simple, interpretable objects that still offer valuable information on second-order contributions to sequence-structure relationships, we find the SCP to be a convenient tool for dissecting the sequential and energetic effects of interaction geometry.

Recent breakthrough successes in structure prediction, with end-to-end deep learning models now able to give accurate predictions of a sequence's native contacts[120] and structure[17,121], suggest that the PDB contains many generalizable patterns that relate sequence and structure. The SCP is an example of a PDB-derived generalization, being simple to describe and understand while capturing important aspects of pairwise contributions to sequence-structure relationships in a context-sensitive way. Better understanding these and other generalizations will be helpful for guiding novel prediction and design methods, especially in areas where improved performance is much needed, such as the prediction of protein-protein interfaces[122].

5.2 Results

5.2.1 Definitions of the contact potential and structure-conditioned potential

A contact potential infers pseudo-energies associated with amino-acid pair interactions from observations of amino-acid contacts in a structural database. In its simplest form, a contact potential measures the extent each amino-acid pair is over- or under-observed relative to the number of observations expected if there were no pair preferences[1] (e.g., the number expected based on the product of the marginal distributions of both amino acids comprising the pair). If the database from which the amino-acid pair statistics arise includes a large diversity of contacts, the resulting pseudo-energies reflect how disproportionately each amino-acid pair interacts in native structural contexts. For example, cysteine-cysteine contacts are observed much more frequently than expected based on the low background frequency of cysteine residues in native structures and therefore cysteine-cysteine contacts have a strongly negative (favorable) energy according to a traditionally derived contact potential. There are alternative formulations of the reference state—the framework for computing the expected number of observations[67,97,110]—but most assume an independence between sequence and structure elements and therefore estimate cysteine-cysteine interactions favorably. Eq. 5.1 describes the traditional formulation of a contact potential, where $N_{obs}(a)$ is the number of observations in the database of pairs of amino acid type a, $N_{obs}(a, b)$ is the number of

observations of amino acid types *a* and *b* in contact in either order, and *N* is the total number of amino acids in the database of pairs. The term H(*a*, *b*) adjusts the expectation for heterotypic pairs (i.e., amino-acid pairs for which *a* is not *b*) as the potential is directionless. The pseudocount ε ensures sparse statistics do not result in a division of or by zero. Taking the negative log of the ratio transforms the values into an additive pseudo-energies metric. See the "Contact Potential" section of Methods for details.

$$E(a, b) = E(b, a) = -\log \frac{N_{obs}(a, b) + \varepsilon}{N_{exp}(a, b) + \varepsilon} = -\log \frac{N_{obs}(a, b) + \varepsilon}{N_{obs}(a) \cdot N_{obs}(b) \cdot H(a, b)/N + \varepsilon}$$

Eq. 5.1

The SCP computes a contact potential for a given interaction motif (e.g., a specific pair of residues together with their surrounding backbone fragments). The resulting interaction matrix encodes the pseudo-energetic preferences for each amino-acid pair in this given structural context. Like a contact potential, these pseudo-energetic preferences are calculated from amino-acid pair statistics, encoding to what extent each amino-acid pair is over- or under-observed relative to the number expected if the two positions in question did not influence each other. Unlike a contact potential, the amino-acid statistics do not come from a generic database of contacts, but instead from a structural ensemble of fragments that share a similar geometry with the interaction motif in question (i.e., constrained by a maximum RMSD over backbone atoms to the input motif). The resulting SCEs reflect the extent to which each amino-acid pair disproportionately interacts in the specific structural context of the interaction motif. The formulation of SCEs is therefore similar to Eq. 5.1 but computed over a constrained set of statistics:

$$SCE(a, b) = -\log \frac{N_{obs}(a, b) + \varepsilon}{N_{exp}(a, b) + \varepsilon}$$

Eq. 5.2

Here, (a, b) is the amino-acid pair, $N_{obs}(a, b)$ is the number of occurrences of pair (a, b) in the interaction motif's ensemble of matching fragments, $N_{exp}(a, b)$ is the number of occurrences of pair (a, b) that would be expected if there were no pair preferences, and ε is a pseudocount. See section "Structure-conditioned potentials" in Methods for details. In order to measure the impact of incorporating additional structural context, here we consider three types of interaction motifs that incorporate an increasing number of flanking residues around a pair of interacting positions. Specifically, 1x1, 3x3, and 5x5 motifs comprise a pair of contacting residues with no flanking residues, one flanking residue on each side, or two flanking residues on each side, respectively (Fig. 5.1). Larger motifs may carry important contextual information, which can offer certain advantages, but may also be associated with decreased statistics in a limited database. Comparing the results with these different motif types measures the impact of increased structural context and potentially decreased database statistics.

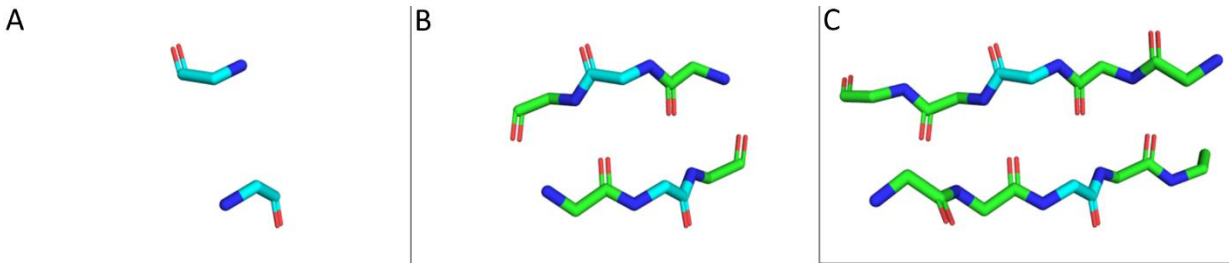


Fig. 5.1 Visualization of an interaction motif. **A**-**C**: The same pair of interacting residues is shown with increasing structural context: 1x1 (A), 3x3 (B), and 5x5 (C). In each case, the pair of interacting residues is colored in cyan. The pair of interacting residues is (A108, A188) from the structure with PDB ID 4G1Q and was visualized in PyMOL[91].

5.2.2 Averaging SCEs over many structural contexts converges to a traditional contact potential

If each SCE matrix encodes the amino-acid pair energies in a particular context, then the average of each energy over many contexts should encode similar information to a generic contact potential. To test this, a database of 200,000 contacts from a nonredundant subset of the PDB (DB200K, see "Contact database creation" in Methods)

was used to compute a contact potential and, for each of their corresponding 3x3 interaction motifs, a 20x20 matrix of SCEs. For each amino-acid pair, the average SCE over all contacts involving the pair in the database was computed. Fig. 2 plots contact potential energies (CEs) against corresponding mean SCEs, showing a linear correlation coefficient of R=0.88. In contrast, SCEs at a particular pair of sites generally correlate quite poorly with contact potential energies (R=0.20 on average, see Fig. 5.3A). This suggests that while SCEs and CEs capture similar effects, and converge on average, SCEs are much more context sensitive and thus have the potential to capture many details that CEs may miss.

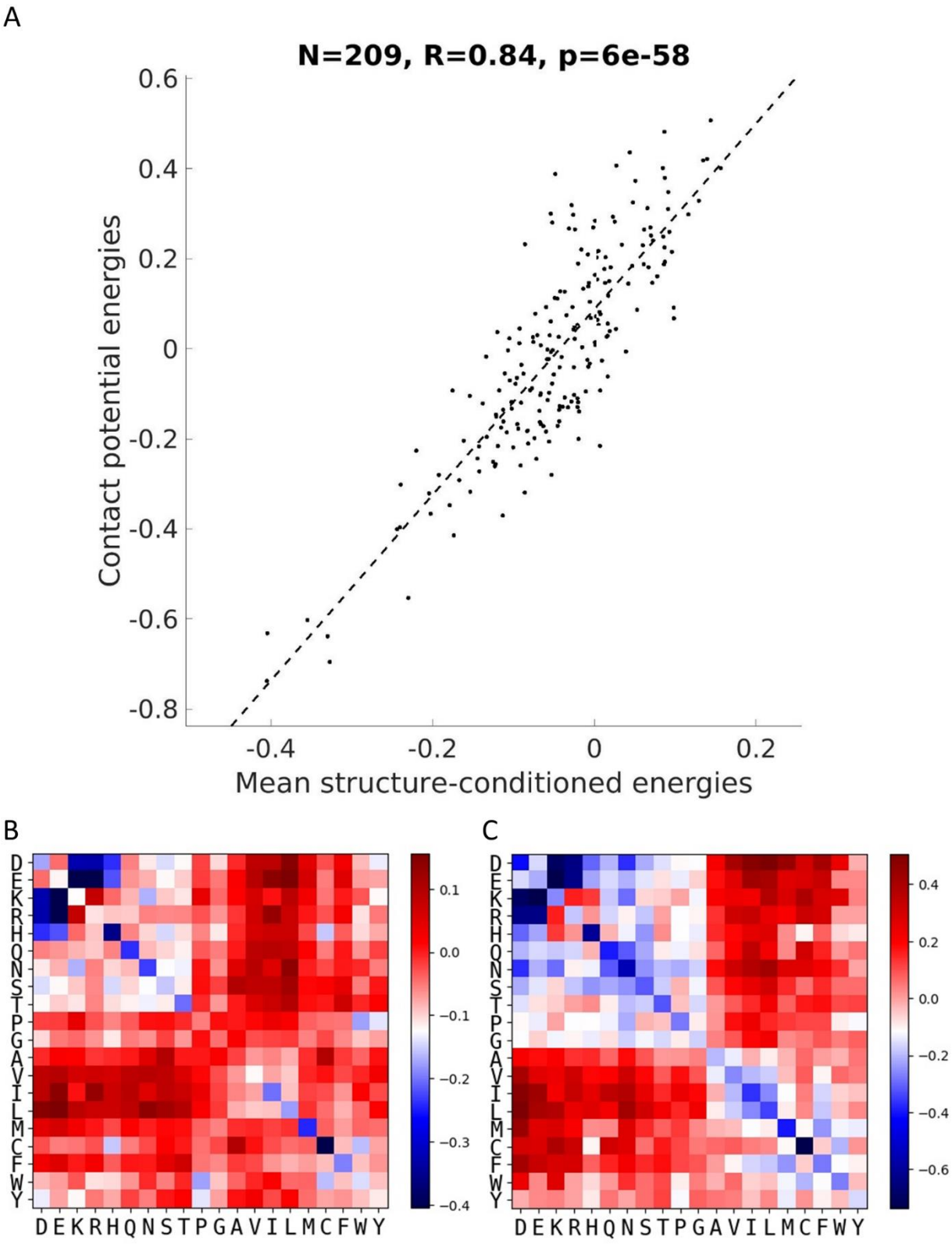
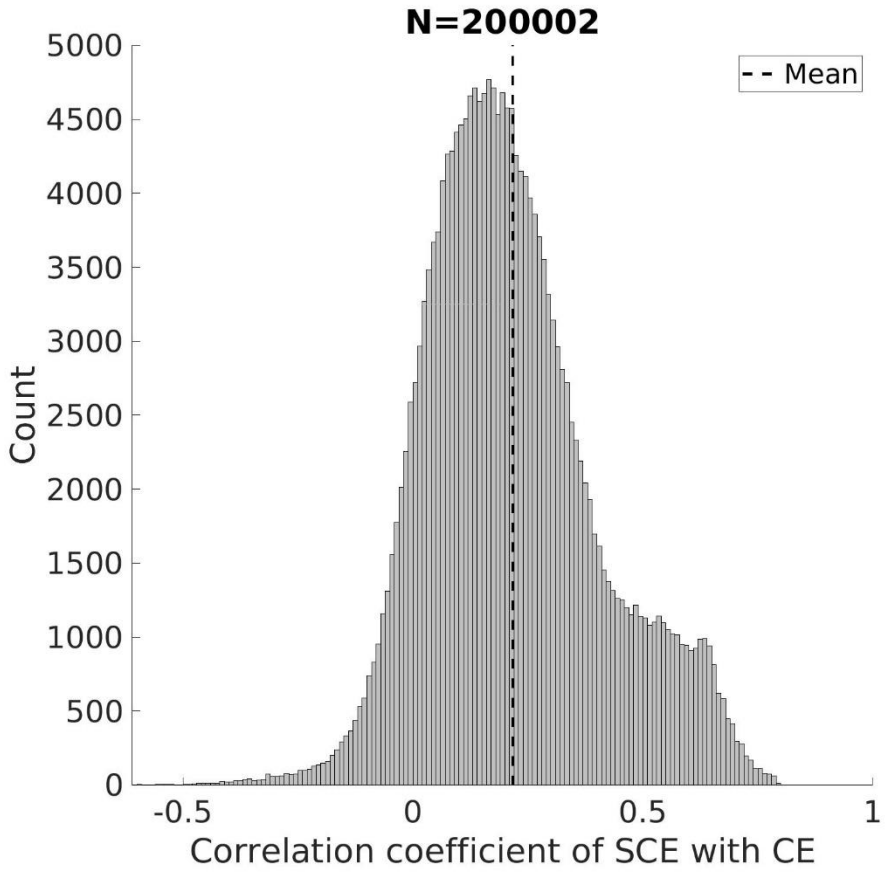
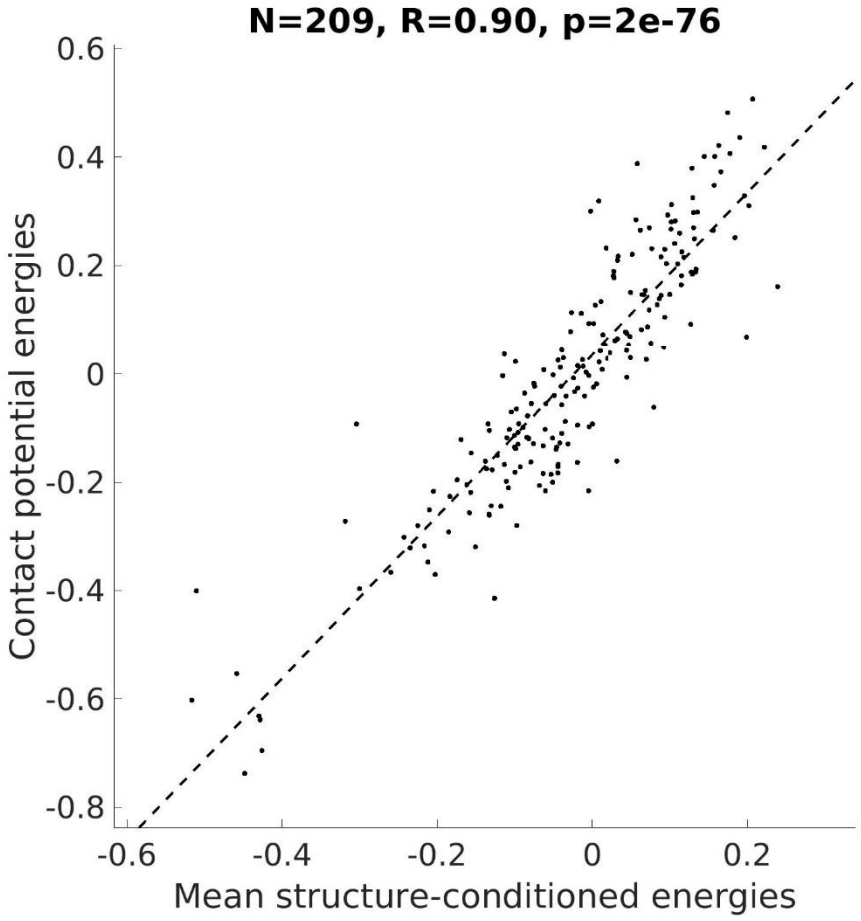


Fig. 5.2 Correlation between each SCE from 3x3 motifs, averaged over many contexts, and

the corresponding energy according to a contact potential. **A**: Mean SCEs plotted against corresponding contact-potential energies. Pair Cys-Cys is not shown as it occupies a point far to the bottom-left (contact potential of -1.78 and mean SCE of –1.10), though its inclusion increases the correlation to R=0.88. The dotted line indicates the best linear fit of the data. **B**: Heatmap of the mean SCEs. **C**: Heatmap of contact-potential energies. For (B) and (C), the energy scale was capped at the most favorable energy except for Cys-Cys for easier visualization.

Interestingly, the strong correlation between mean SCEs and CEs also holds for 1x1 and 5x5 motifs, but does show a decline with increased context length (i.e., R=0.91 and R=0.84 for 1x1 and 5x5, respectively; Fig. 5.3B,C). This decline is expected as more averaging would be needed to integrate out the influence of more detailed structural context. Overall, these results show that this SCP can be thought of as a more elaborate and context-sensitive counterpart of the traditional contact potential.


N=200002
Mean
Count
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
-0.5
0
0.5
1
Correlation coefficient of SCE with CE


N=209, R=0.90, p=2e-76
Contact potential energies
Mean structure-conditioned energies
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-0.6
-0.4
-0.2
0
0.2

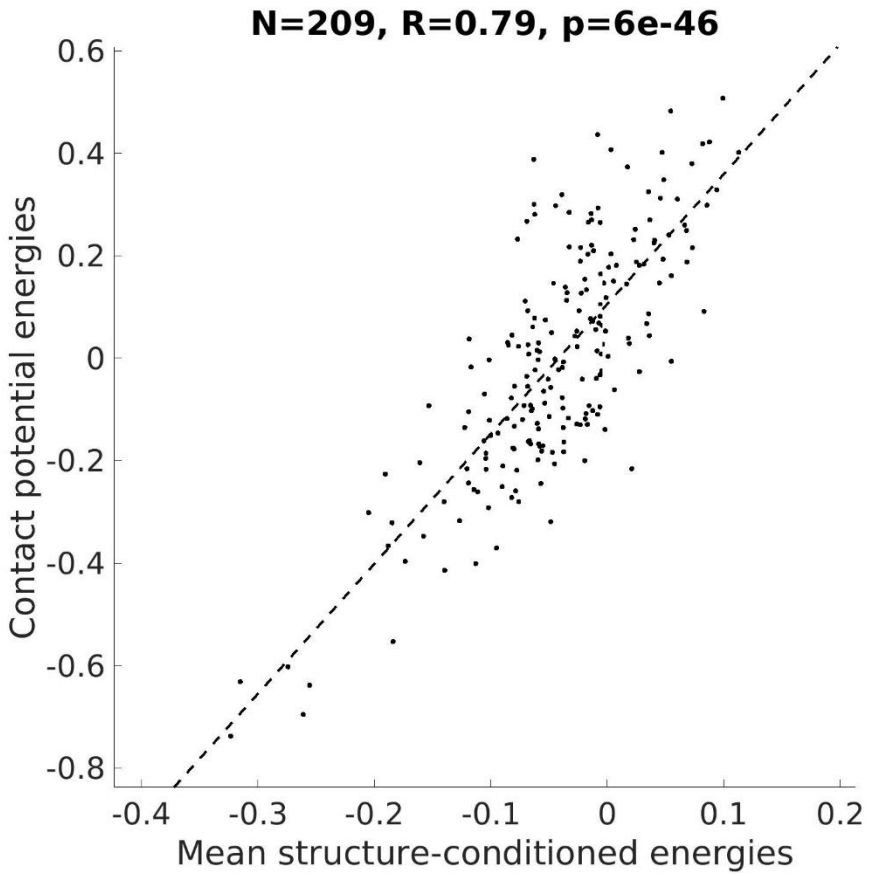


Fig. 5.3 Additional information about correlations between SCEs and CEs. **Top**: Distribution of correlation coefficients between each set of SCEs in DB200K and the CEs. The dotted line indicates the mean correlation of R=0.20. **Middle**, **Bottom**: Plot of the averaged, symmetrized SCEs from 1x1- (Middle) or 5x5-motifs (Bottom) vs those from a contact potential. As with Fig. 2A, Cys-Cys is not shown as it occupies points to the far bottom-left (mean SCE of -1.70 (1x1) or -0.88 (5x5)), though its inclusion increases the correlation to R=0.91 (1x1) or R=0.84 (5x5).

5.2.3 Conditioning on structure encodes more accurate sequence information

A simple test of the information gained by conditioning on structural context is to measure how reliably the SCP favors the native amino-acid pair of an interaction motif it corresponds to, compared to how reliably the contact potential favors it. Note that one may not expect very high performance in such a test, as the native choice of amino-acid

pairs is guided not only by second-order effects, but also (and perhaps more importantly) by first-order effects. Nevertheless, as a vehicle for discerning the effects of structure conditioning and context, the test is still fair—i.e., a more accurate second-order potential *should* more frequently pick out the native pair.

With the same set of contacts used to create the statistical potential and averaged energies in Fig. 2, the SCP of each contact's interaction motif was used to predict the residue type pair of that contact. Success was measured both by identification (whether the amino-acid pair with the most favorable pair energy is the native residue pair) and score (by how much the energy of the native residue pair differs from the energy of the most favorable pair, as measured by a modified Z-score; see Eq. 5.3 in the "AA pair identification" section of Methods), both of which are shown in Fig. 5.4. For comparison, CEs, as encoded in the aforementioned contact potential, were used to compute the same metrics. To further understand the role that structural context plays in encoding sequence preferences, the predictive performance of SCEs was computed for 1x1, 3x3, and 5x5 interaction motifs.

As shown in Fig. 5.4, conditioning energies on structural context substantially improves the information they contain about native sequence preferences, with energies from 3x3 interaction motifs about four times as likely to identify the native residue pair as by chance (1/400), compared to the below-chance performance of general amino-acid pair preferences (contact potential). The poor sequence-identification performance of the contact potential occurs because it predicts every native pair to be the most favorable one, cysteine-cysteine (Cys-Cys), while in actuality Cys-Cys pairs make up less than 1/400th of the total pairs (~0.17%). Cystines are rare but very frequently occur in pairs (as either disulfide bonds or in functional sites), and for this reason the Cys-Cys pair gets an anomalously favorable contact potential as the most over-represented pair over expectation. In so far as statistical potentials represent interaction strength, this is not entirely wrong—Cys-Cys pairs can form covalent bonds, which are much stronger than non-covalent interactions of other amino-acid pairs. However, it is not the case that disulfide bonds can be made across any proximal residue pair, and in fact strict geometric requirements have been identified for Cys-Cys bond formation[123]. The traditional contact potential has no choice but average out the effect of Cys-Cys contacts across all

geometric circumstances, yielding still a very strong effective contact energy. On the other hand, the SCP is able to discern where disulfides are likely and better apportion the high favorability of Cys-Cys pairs to just the relevant geometric contexts. In fact, for 1x1, 3x3, and 5x5 motifs, Cys-Cys is predicted as the most favorable contact in 36.9%, 4.3%, and 2.6% of pairs, respectively, a progressive narrowing of the geometries considered permissible for Cys-Cys as additional structural context is incorporated.

The contrast in performance between 1x1 and 3x3 motifs (Fig. 5.4) highlights the effect of incorporating structural context further. Because 1x1 motifs lack flanking residues, the structural similarity of the ensembles used to calculate their energies lacks information about and therefore conflates a variety of structural contexts; a 1x1 ensemble implicitly constrains simple features like overall inter-residue and relative orientation but not how the surrounding structure frames the contact. In contrast, the extensive context of 5x5 motifs leads to a distribution of native energies that is reliably more favorable than the median energy for each pair, likely because the choice of amino acid types in such specific contexts is even more constrained than for 3x3 motifs.

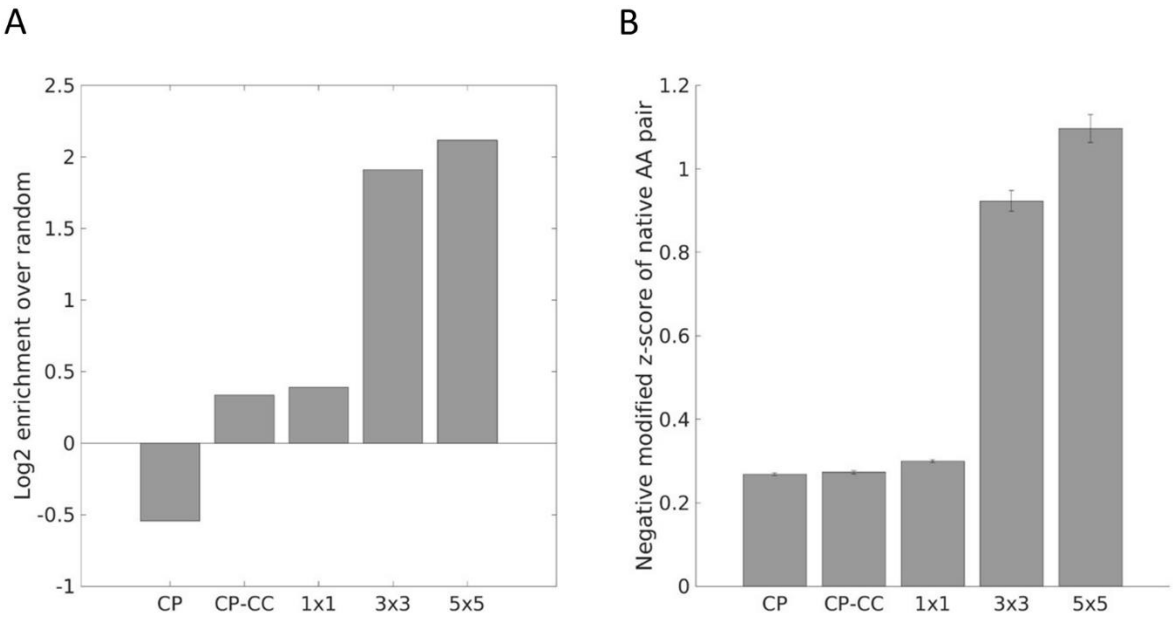


Fig. 5.4 Sequence information contained in SCEs. **A**: Enrichment over random choice of native amino-acid pair identification for DB200K comparing energies for each native residue pair using a contact potential (CP), a contact potential without cysteine-cysteine

(CP-CC), and SCEs from three types of interaction motifs, 1x1, 3x3, and 5x5. **B**: Negative modified z-score of the energies used in (A). Modified z-scores were computed using medians and median absolute deviations rather than means and standard deviations. Error bars indicate the standard error of the mean.

As shown in Fig. 5.4, conditioning energies on structural context substantially improves the information they contain about native sequence preferences, with energies from 3x3 interaction motifs about four times as likely to identify the native residue pair as by chance (1/400), compared to the below-chance performance of general amino-acid pair preferences (contact potential). The poor sequence-identification performance of the contact potential occurs because it predicts every native pair to be the most favorable one, cysteine-cysteine (Cys-Cys), while in actuality Cys-Cys pairs make up less than 1/400th of the total pairs (~0.17%). Cystines are rare but very frequently occur in pairs (as either disulfide bonds or in functional sites), and for this reason the Cys-Cys pair gets an anomalously favorable contact potential as the most over-represented pair over expectation. In so far as statistical potentials represent interaction strength, this is not entirely wrong—Cys-Cys pairs can form covalent bonds, which are much stronger than non-covalent interactions of other amino-acid pairs. However, it is not the case that disulfide bonds can be made across any proximal residue pair, and in fact strict geometric requirements have been identified for Cys-Cys bond formation[123]. The traditional contact potential has no choice but average out the effect of Cys-Cys contacts across all geometric circumstances, yielding still a very strong effective contact energy. On the other hand, the SCP is able to discern where disulfides are likely and better apportion the high favorability of Cys-Cys pairs to just the relevant geometric contexts. In fact, for 1x1, 3x3, and 5x5 motifs, Cys-Cys is predicted as the most favorable contact in 36.9%, 4.3%, and 2.6% of pairs, respectively, a progressive narrowing of the geometries considered permissible for Cys-Cys as additional structural context is incorporated.

The contrast in performance between 1x1 and 3x3 motifs (Fig. 5.4) highlights the effect of incorporating structural context further. Because 1x1 motifs lack flanking residues, the structural similarity of the ensembles used to calculate their energies lacks information about and therefore conflates a variety of structural contexts; a 1x1 ensemble

implicitly constrains simple features like overall inter-residue and relative orientation but not how the surrounding structure frames the contact. In contrast, the extensive context of 5x5 motifs leads to a distribution of native energies that is reliably more favorable than the median energy for each pair, likely because the choice of amino acid types in such specific contexts is even more constrained than for 3x3 motifs.

5.2.4 Energies conditioned on structure correlate more highly to experimental coupling energies

Another way to test the SCP is to evaluate how well its pair energies correspond to experimentally determined preferences. Most useful would be experimental measurements that focus on pair interaction strength (i.e., the second-order effect). Detailed thermodynamic measurements or high-throughput deep mutational scans usually focus on point mutations[124]. For instance, while ProThermDB[125] contains measurements of how point or double mutations change the stability of their structure, there are no position pairs with measurements for more than a few residue pair combinations. However, some systems have been well studied using the double-mutant coupling energy approach[126], which attempts to isolate solely the second-order effect on stability. For example, Vinson and co-workers have produced high-quality coupling energy measurements for ~100 amino-acid pairs at two inter-chain site pairs for the dimeric parallel coiled coil system (a set of 81 for a-a' core interactions[127] and 16 for interfacial g-e' interactions[76]). These coupling energies measure the change in the free energy of folding when a pair of interacting residues is simultaneously mutated to alanine, relative to when each is individually mutated to alanine[76]. Because folding and dimerization are concomitant in this system, these coupling energies cleanly isolate just the contribution of the residues interacting in the folded state (as these interactions are absent in the unfolded/dissociated state). This an ideal scenario for comparing to SCEs, which report on the relative second-order preferences for different amino-acid pairs in a specific structural context.

The advantage of SCEs for capturing coupling energies is that their underlying statistics come from an ensemble of similar interaction motifs. In fact, the specific ensemble represented by the native structure would be most appropriate to use when

trying to estimate true coupling energies. Of course, we do not know the native ensemble or even what RMSD neighborhood around the native interaction motif would be most appropriate to represent it. For this reason, we chose to consider multiple ensemble sizes (corresponding to a range of RMSD cutoffs) to investigate the impact on the correspondence between SCEs and experimentally determined coupling energies. Results are shown in Fig. 5.5A and 5.5B for a-a' and g-e' interactions, respectively, with the performance of CEs shown with a dashed line. Interestingly, the correlation for both the a-a' and g-e' interactions is maximized when the ensemble is very tight (maximum RMSDs of 0.24 and 0.38, respectively), suggesting that the coiled-coil system used to measure these coupling energies may occupy a relatively narrow native ensemble at equilibrium. But while the maximum correlations occur at low RMSDs, the correlation remains high over a broad range. For the a-a' interactions, the correlation exceeds that of the contact potential no matter the size of the ensemble. For the g-e' interactions, while the correlation is about equal to the contact potential's for larger ensembles, it is much higher for small ensembles except when the statistics are sparse enough that the number of matches from which SCEs are derived becomes considerably lower than the number of amino-acid pairs (left-most point in Fig. 5.5B). Note that our "default" setting for computing SCEs in this work is to require a minimum ensemble of 1000 matches to ensure that data sparsity does not come into play (see the "Structure-conditioned potentials" section in Methods). The correlations for the energies computed with these default settings are shown with asterisks in Figs. 5.5A and 5.5B and in both cases, the correlation using SCEs is higher than when using CEs.

Figs. 5.5C and 5.5D compare experimental coupling energies to SCEs whose ensembles achieved maximal correlation for the a-a' and g-e' interactions, respectively. This is a striking contrast to the correlations produced by using contact potential energies (R=0.65 versus R=0.31 for a-a' energies and R=0.85 versus R=0.66 for g-e' energies; see Fig. 5.6A,B for plots of CEs vs experimental energies). It is clear that amino-acid statistics from structural ensembles resembling native interaction geometries give better insights into the thermodynamic coupling between positions than the more generic preferences of a contact potential do.

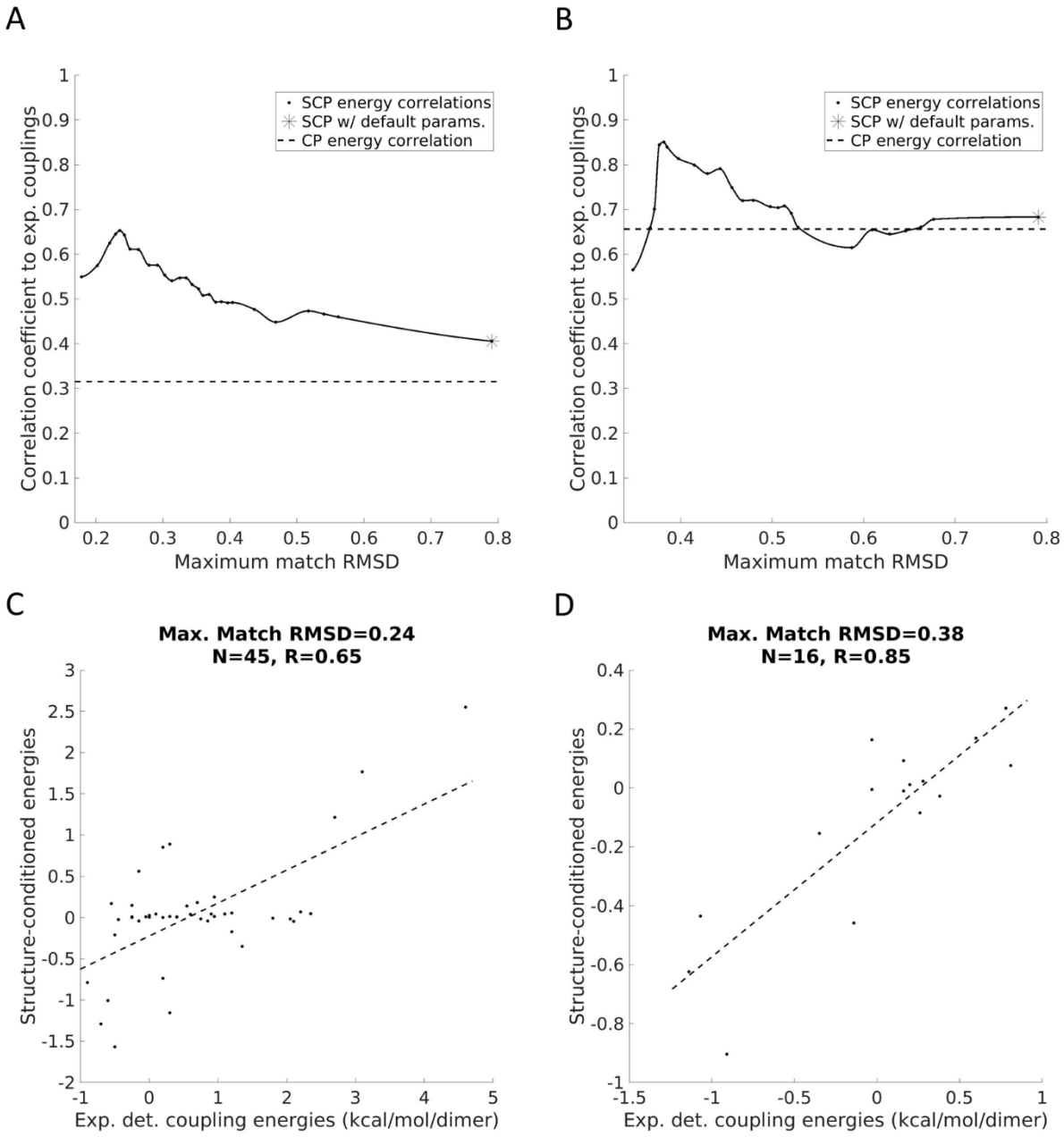
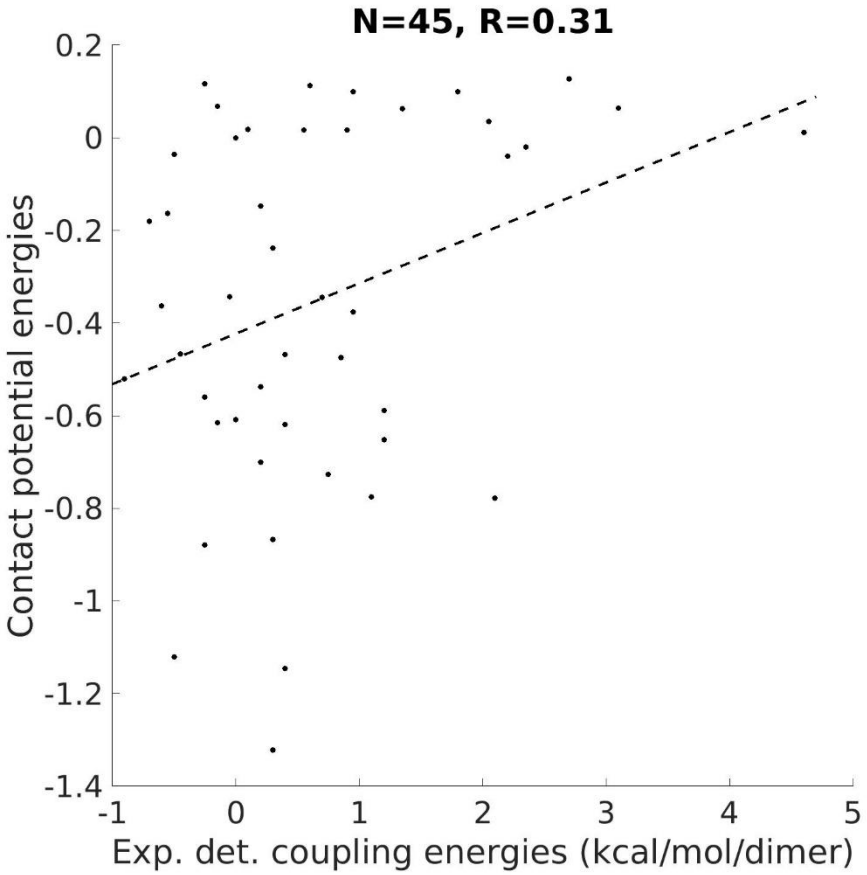


Fig. 5.5 SCEs vs experimentally determined coupling energies. **A**, **B**: Correlation between experimentally determined energies vs SCEs over a range of ensembles for a-a' (A) and g-e' (B) interactions. The dotted line in each plot indicates the correlation achieved by contact potential energies. The right-most points of each plot, labeled with an asterisk (*), correspond to the energies using the default parameters. The curve in between points was computed using the 'pchip' function of MATLAB and is for visualization purposes only. **C**, **D**: Correlation between experimentally determined energies vs optimal SCEs for a-a' (C) and g-e' (D) interactions. The dotted line indicates the best linear fit of the data.


N=45, R=0.31
Contact potential energies
Exp. det. coupling energies (kcal/mol/dimer)
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
-1
0
1
2
3
4
5

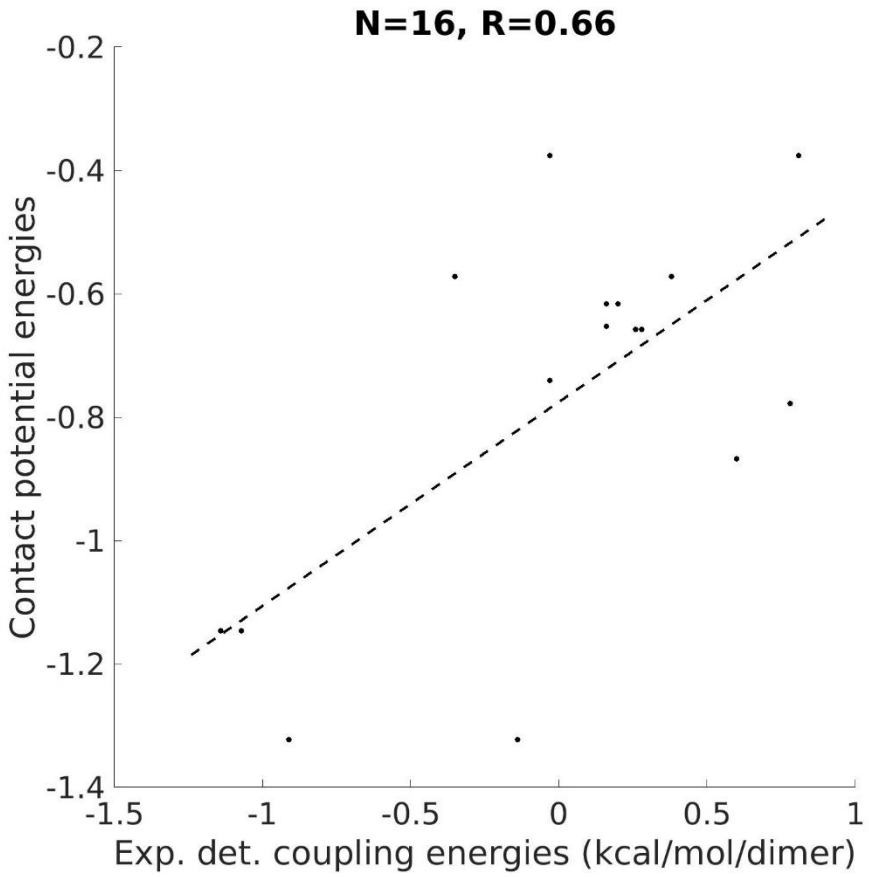


Fig. 5.6 CEs vs experimentally determined coupling energies. **Top**, **Bottom**: Correlation between experimentally determined energies vs CEs for a-a' (Top) and g-e' (Bottom) interactions. The dotted line indicates the best linear fit of the data.

5.2.5 Similar SCE patterns correspond to similar structural motifs and vice versa

Given that the amino-acid pair statistics the SCP uses to compute its energies come from an ensemble of motifs structurally similar to the input motif, we expect pairs of structurally similar input motifs to generate similar energies. We next ask if the reverse is true, whether pairs of motifs with similar SCE matrices are structurally similar as well. The additional information about native sequence preferences contained in these energies relative to those of the contact potential (see Fig. 5.4) suggests this may be the case, as this information gain is likely driven by the distinct amino-acid pair statistics imposed by particular contact geometries. This potential relationship—the mutual information

between contact geometry and SCEs—can be examined by clustering a large set of interaction motifs by both their structures and SCE matrices and examining the structural and energetic similarities within the resulting clusters. Taking a random subset of 50,000 motifs from DB200K, motifs were clustered by both structure (via RMSD) and energy (via r_E, a function of the linear correlation of the energies as 400-vectors; see the "Clustering" section in Methods). Clustering was done greedily, with clusters defined by their radius (in RMSD space for structures, and in correlation space for energies; see the "Clustering" section in Methods).

Fig. 5.7 shows a summary of these clustering results. Fig. 5.7A shows visualizations of the structures of three representative clusters in each case, Fig. 5.7B shows the corresponding mean SCE matrices, and Figs. 5.7D and 5.7E display the distributions of RMSDs and energetic distances (r_E) over each case's first 100 clusters. Remarkably, the structural similarity of motifs clustered by energy is nearly as high as when clustered by structure; similarly, the energetic similarity of motifs clustered by structure is nearly as high as when clustered by energy. More quantitatively, the mean RMSD to the medoid (cluster representative) for the first 100 clusters is 0.36 Å when clustering by structure and 0.55 Å when clustering by energy, compared to 3.70 Å when clusters were assigned randomly (with the sizes of the random clusters chosen to match the sizes when clustering by structure). Moreover, the mean energetic distance to the medoid for the first 100 clusters is 0.16 when clustering by structure and 0.18 when clustering by energy, compared to 0.93 when clusters were assigned randomly (with the sizes of the random clusters chosen to match the sizes when clustering by energy), which is close to the r_E=1.0 value corresponding to uncorrelated matrices. Note that in all three cases, the number of motifs being clustered is approximately the same in order to ensure the comparison is fair. As can be seen, SCE matrices contain sufficient information about the structures they were derived from that similar SCE matrices usually correspond to similar structures. Even in the cases in which the structures of SCE-based clusters are not as mutually similar as in structure-based clusters, the clusters are far more similar than expected by chance, with higher RMSDs usually indicative of multiple sub clusters rather than unrelated motifs, and consistent with the idea that distinct interaction geometries can induce similar energetic preferences. An example of this is shown in Fig. 5.7C, which

splits the energy-based cluster marked by the asterisk into eight subclusters (by visual analysis in PyMOL), revealing a set of particular beta-sheet geometries which evidently all share similar energetic preferences (see Fig. 5.9 for additional visualizations of clusters).

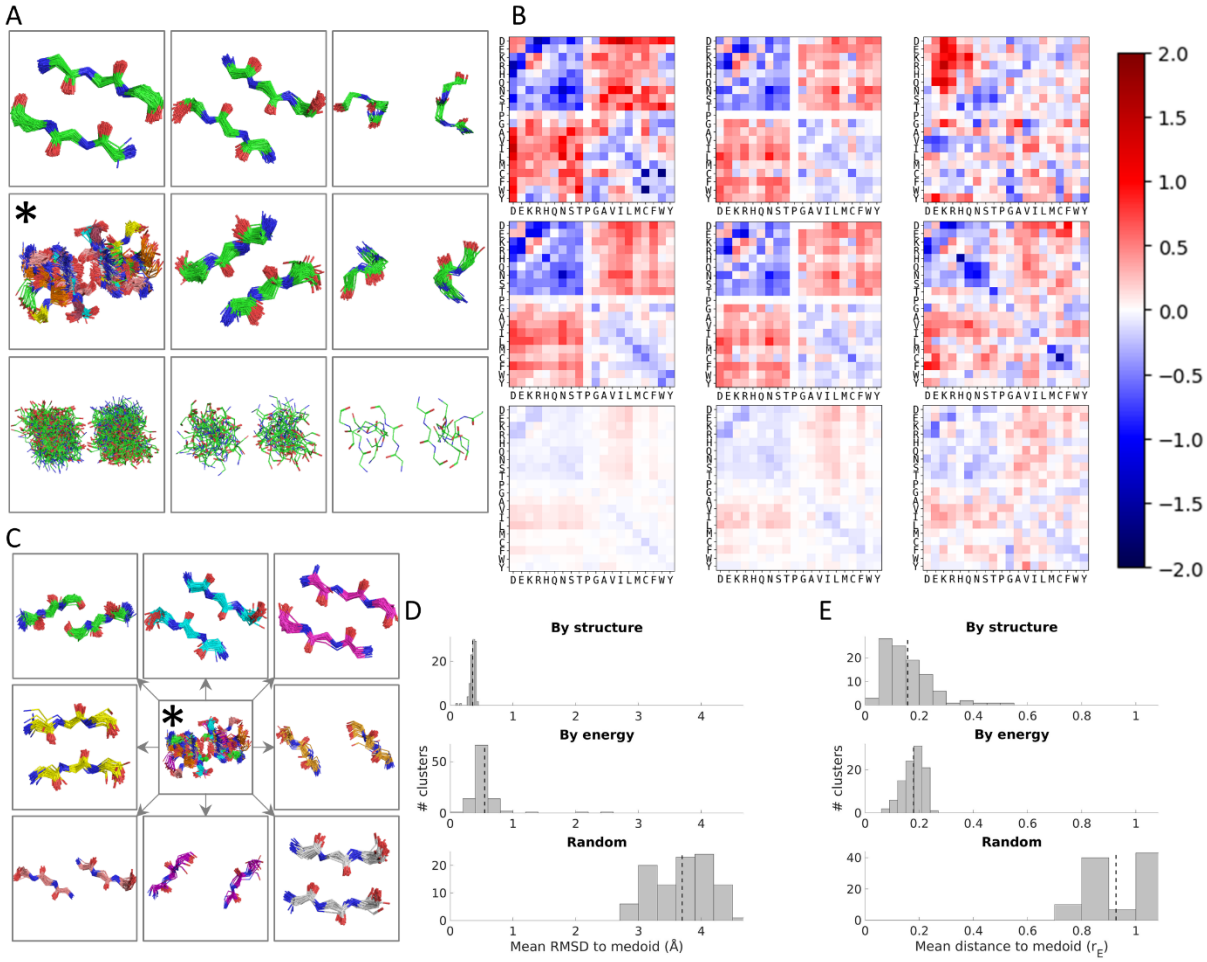


Fig. 5.7 Structurally similar motifs have similar SCEs and vice versa. **A**: Fragment ensembles of the top three clusters by RMSD when clustering by structure (top row), by energy (middle row), and randomly (bottom row). **B**: Mean SCE matrices for the clusters shown in (A). **C**: The subclusters of the cluster marked by the asterisk (*). The subclusters are sorted by size, descending, starting with the top left and going clockwise. **D**, **E**: Distributions of RMSD and energetic similarity (r_E) over the first 100 greedily obtained clusters when clustering by structure, clustering by energy, and by random assignment. For

each distribution shown, the vertical dotted line indicates the mean value. Fragments were visualized with PyMOL.

To look in more detail at how the similarity between structure and energy behaves and where it diverges, we considered a large number of pairs of interaction motifs and computed both their structural similarity (via RMSD) and energetic similarity (via r_E). Specifically, the RMSD and r_E between each pair in a set of 20,000 motifs was computed, the pairs were partitioned into bins based on RMSD, and the average r_E for each bin was calculated (Fig. 5.8). While there is some noise in the relationship, there is a clear pattern of structural similarity implying energetic similarity, with pairs of 3x3 motifs with an RMSD in the range [0, 0.2) having an average r_E of 0.14, in contrast to pairs with an RMSD in the range [1.8, 4) having an average r_E of 0.94, which is about what would be expected for unrelated motifs. Furthermore, an inter-motif RMSD value of ~1.0 Å appears to be roughly where structural and energetic similarity diverge. That is, motif pairs that are closer to each other than 1.0 Å RMSD tend to exhibit various degrees of similarity in their energetic preferences in a way that strongly correlates with structural similarity, while beyond 1.0 Å energetic preferences tend to be mostly unrelated and in a way that does not strongly depend on the specific structural distance.

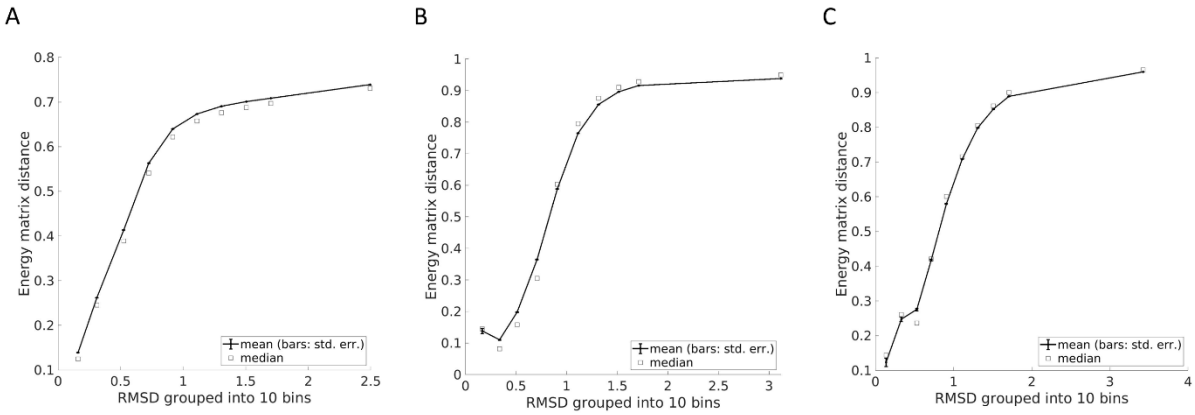
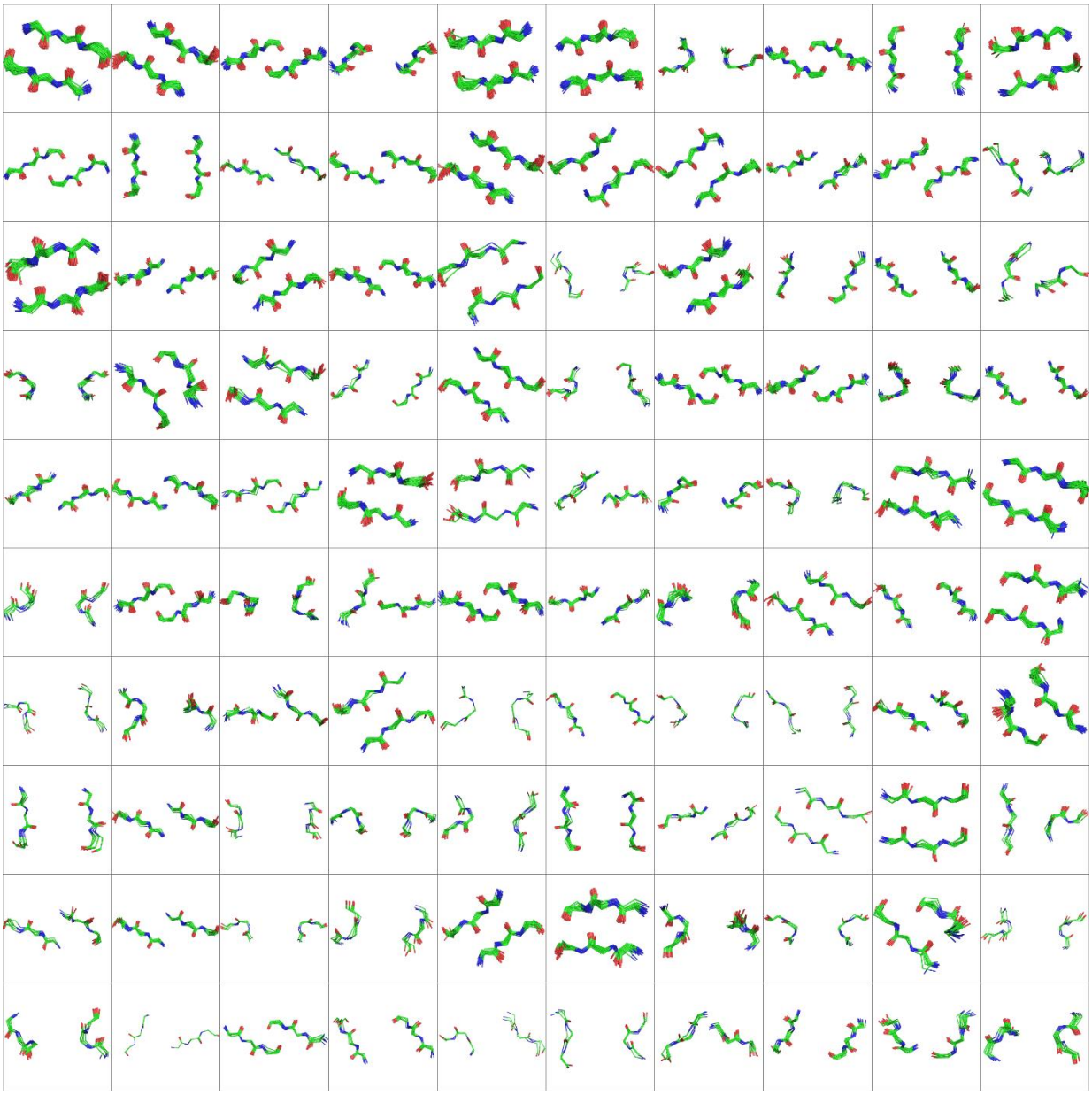
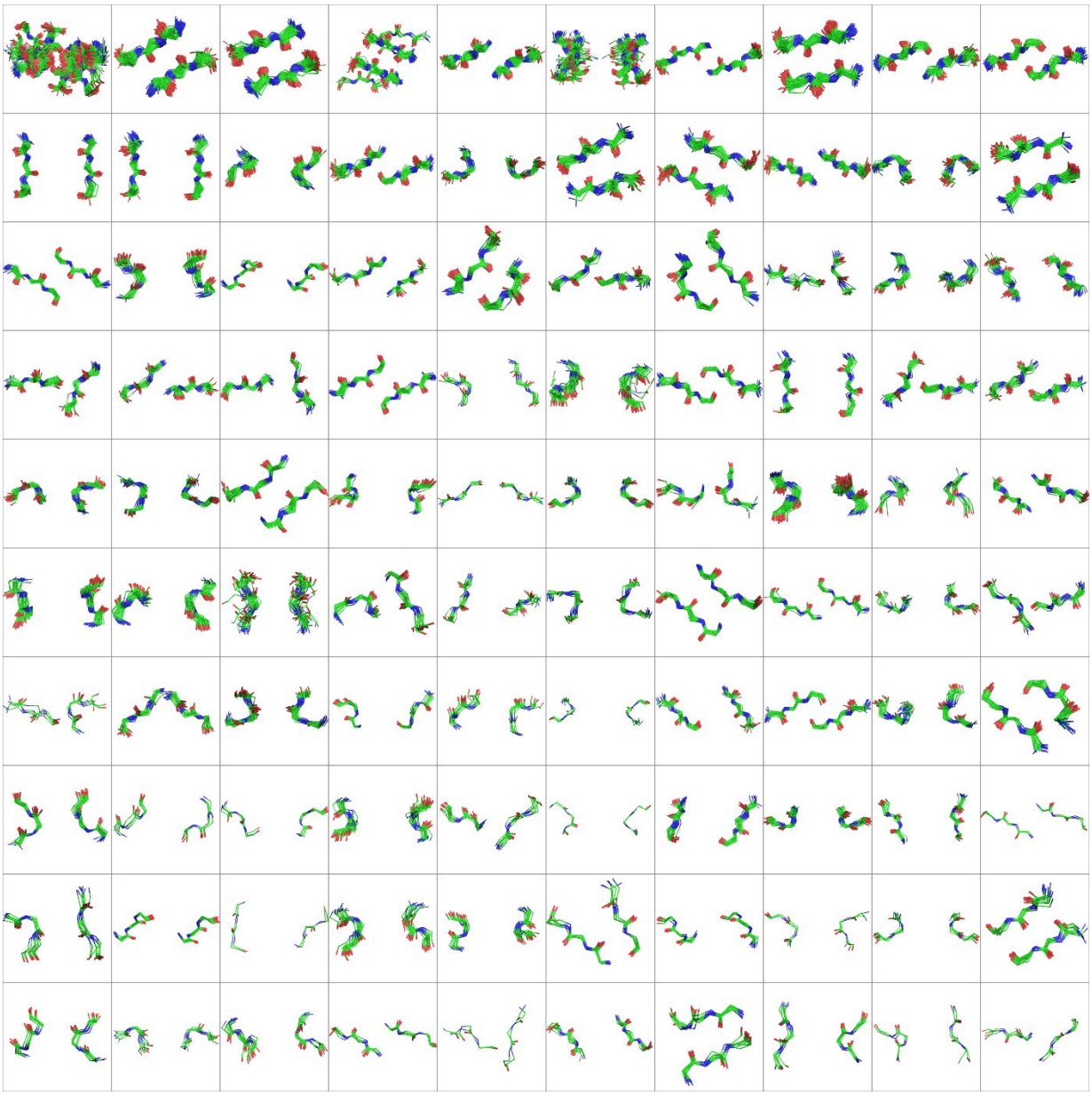
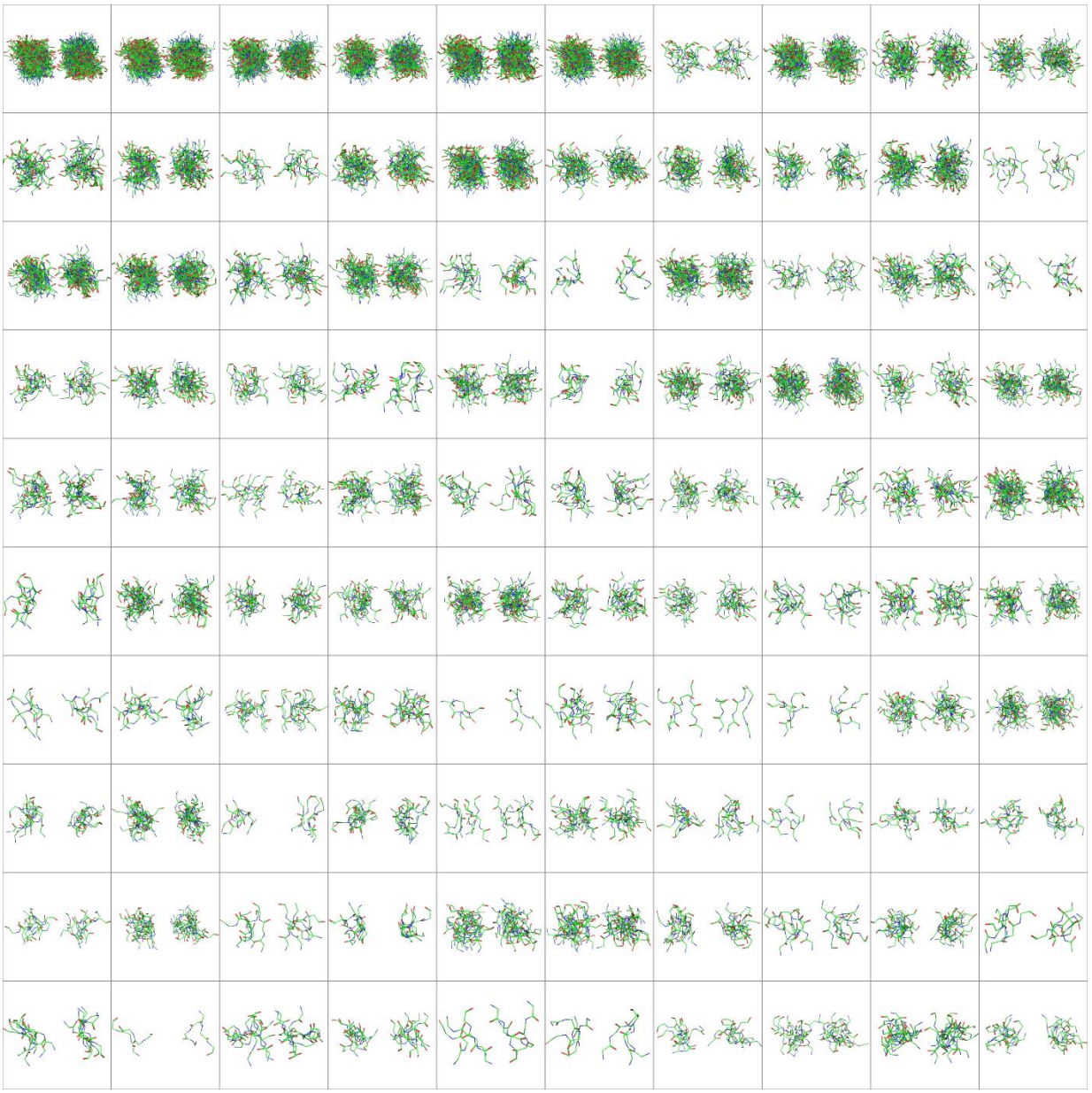
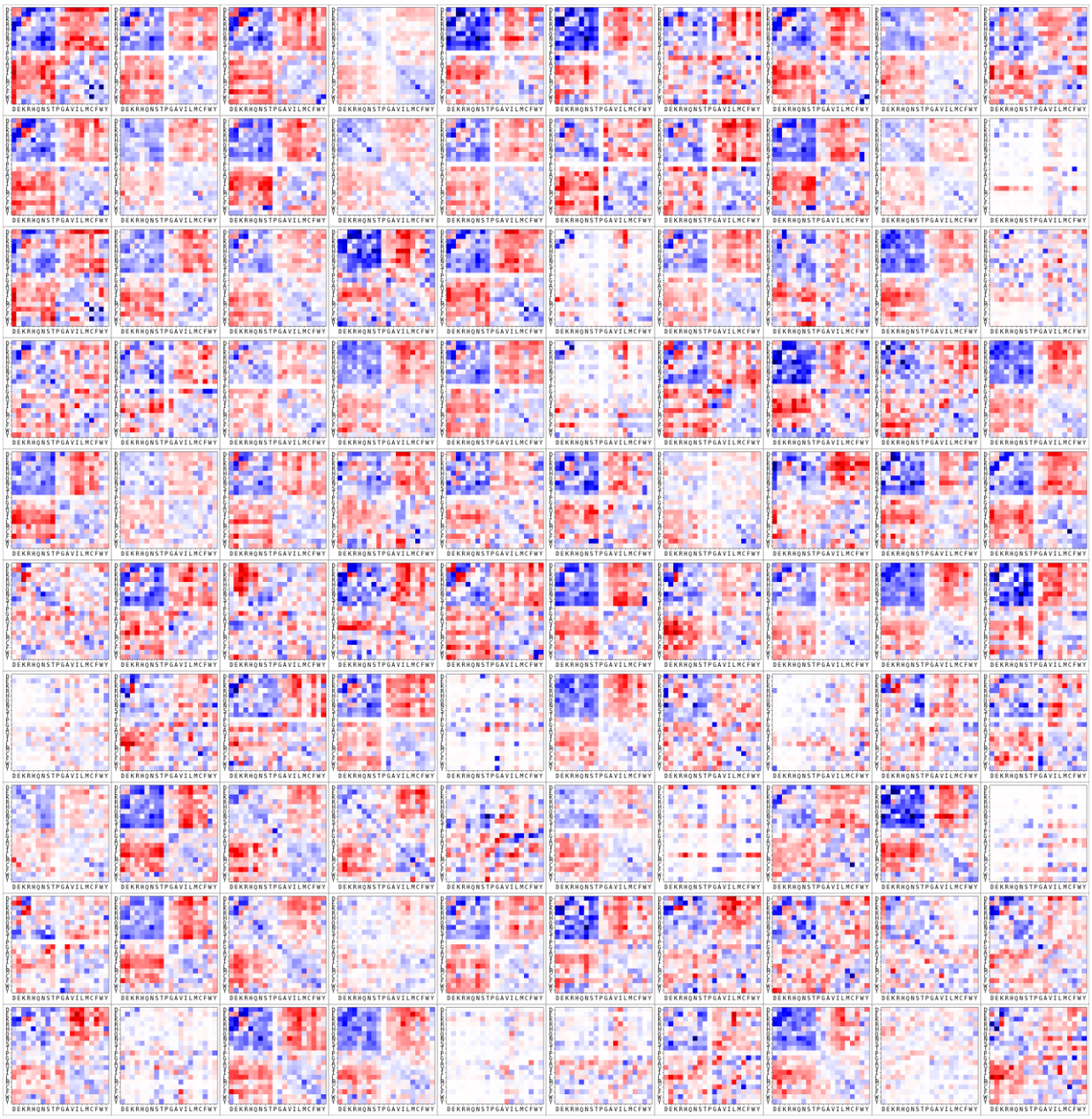


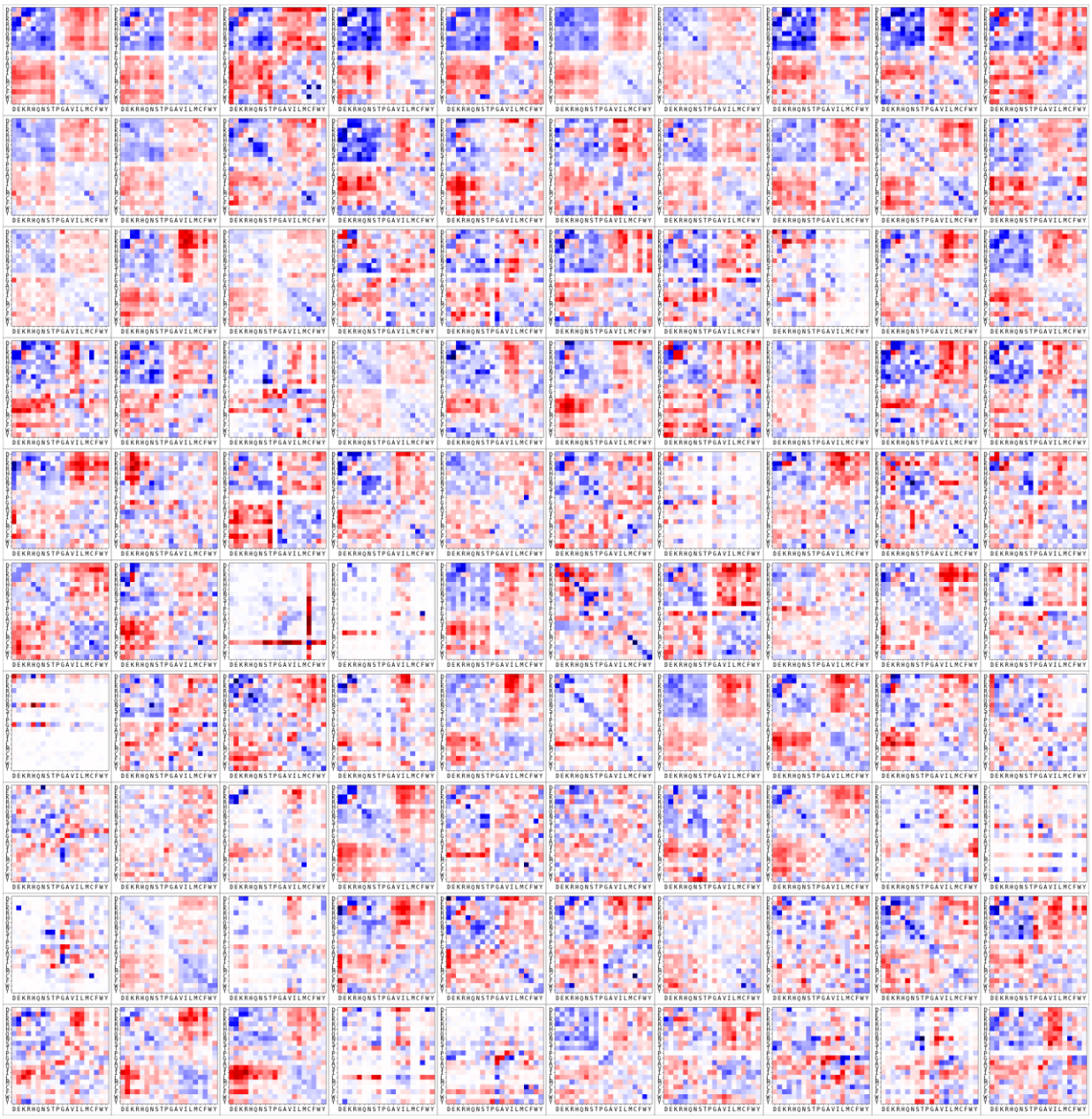
Fig. 5.8 Relationship between structural similarity and energetic similarity. **A**-**C**: 1x1, 3x3, and 5x5 motifs, resp. Circles represent the mean, squares the median, and error bars the standard error.











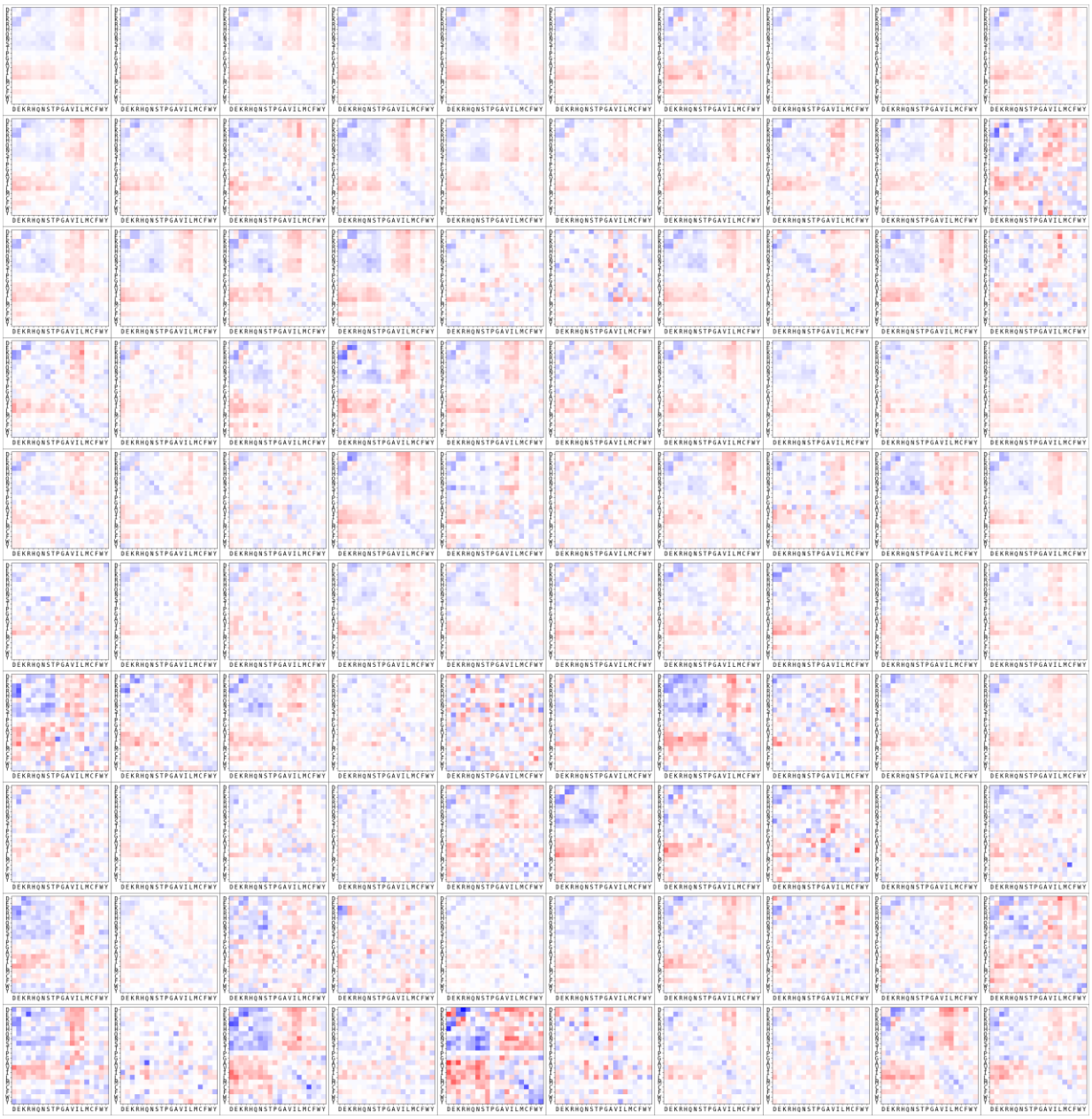


Fig. 5.9 Additional clustering visualizations. The first three figures show the fragment ensembles of the top 100 clusters when clustering by structure, energy, or randomly, respectively. The bottom three figures show the mean SCE matrices for these respective clusterings. The color scale is the same as shown in Fig. 5.7.

5.2.6 SCEs outperform traditional contact energies in native structure discrimination

If SCE matrices reflect how much each interaction motif prefers each possible amino-acid pair, then scoring a structural model by the SCEs of its interacting residues should reflect how compatible the structure is with its sequence. While additional terms would be needed to fully measure sequence-structure compatibility (e.g., self energies as well as the pairwise SCEs), the pair energies should be informative enough to identify native-like interactions. To test this hypothesis, data from previous Critical Assessment of protein Structure Prediction (CASP) competitions, in particular from the refinement challenges[128–133], were collected, resulting in a large set of structures, both native and submitted models—134 targets and ~2500 structures. For each of the 134 refinement targets from CASP9-14, the native structure and 20 models submitted under the refinement category were scored by calculating the mean SCE of the native amino-acid pair over all contacts (using 1x1, 3x3, and 5x5 motifs). As a control, each of these contacts was also scored using CEs and averaged. Following the convention of CASP, GDT_TS[134] was used to measure the quality of each model, and this structure quality was compared to the mean SCE or CE per structure, plotting ROC curves for several GDT_TS thresholds. These curves plot the false positive rate vs. the true positive rate and indicate how well each scoring function can differentiate between models higher vs. lower quality models. Fig. 5.10 shows the curve for differentiating models with a GDT_TS of at least 50 vs. those with a GDT_TS less than 50. Predictive success can be measured with the area under the curve (AUC), which quantifies how likely a scoring function is to correctly rank models and ranges from 0.5 (the expected AUC for a random classifier) to 1.0 (achieved by perfectly ranking the models). The AUCs when models are scored with the mean SCE using 1x1, 3x3, and 5x5 motifs is 0.64, 0.80, and 0.69, respectively. In contrast, the AUC when models are scored with the mean CE is 0.54, indicating the mean CE essentially ranks models randomly with respect to GDT_TS. The increased AUCs achieved by SCE-based scores holds when using other thresholds of GDT_TS (60, 70, 80, 90; see Fig. 5.11).

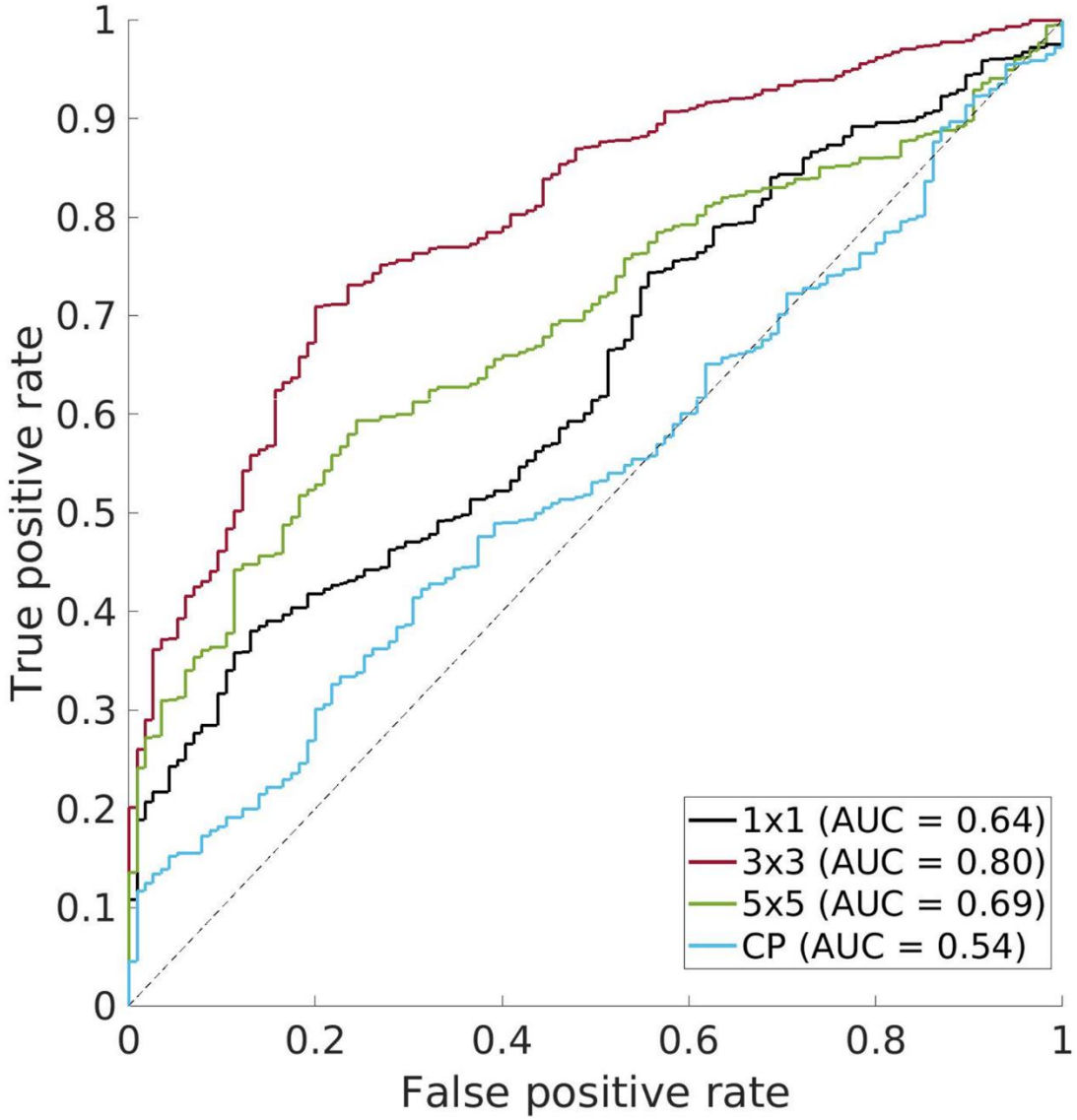


Fig. 5.10 The performance of SCE-based and CE-based scoring functions when differentiating between low- and high-quality CASP models. Each scoring function's ROC curve plots the false positive rate vs. the true positive rate achieved when predicting whether or not each structure in the set of CASP models has a GDT_TS of at least 50. 1x1, 3x3, and 5x5 refer to the SCE-based scoring functions using the respective motif sizes and CP refers to the CE-based scoring function.

The substantially larger AUCs achieved by the SCE-based scores, in particular the 0.8 achieved using 3x3 motifs, is another piece of evidence for the SCP encoding

additional information about native sequence preferences compared to the CP and shows this information can be exploited to evaluate the structural quality of predicted models. The improved predictions using the 3x3-based scores over the 1x1-based scores is expected, but the lack of systematic improvement by the 5x5-based scores may result from limitations associated with the sparser and less robust statistics of larger structural motifs. To see whether the relationship between SCE-based scores and GDT_TS generalizes to other measures of structural quality, the same experiment was performed using TM-score[135] (Fig. 5.12), another common measure of structure quality, and RMSD (Fig. 5.13). In both cases, the same pattern was observed, with SCEs predicting structural quality more effectively than CEs across all tested thresholds of TM-score and RMSD.

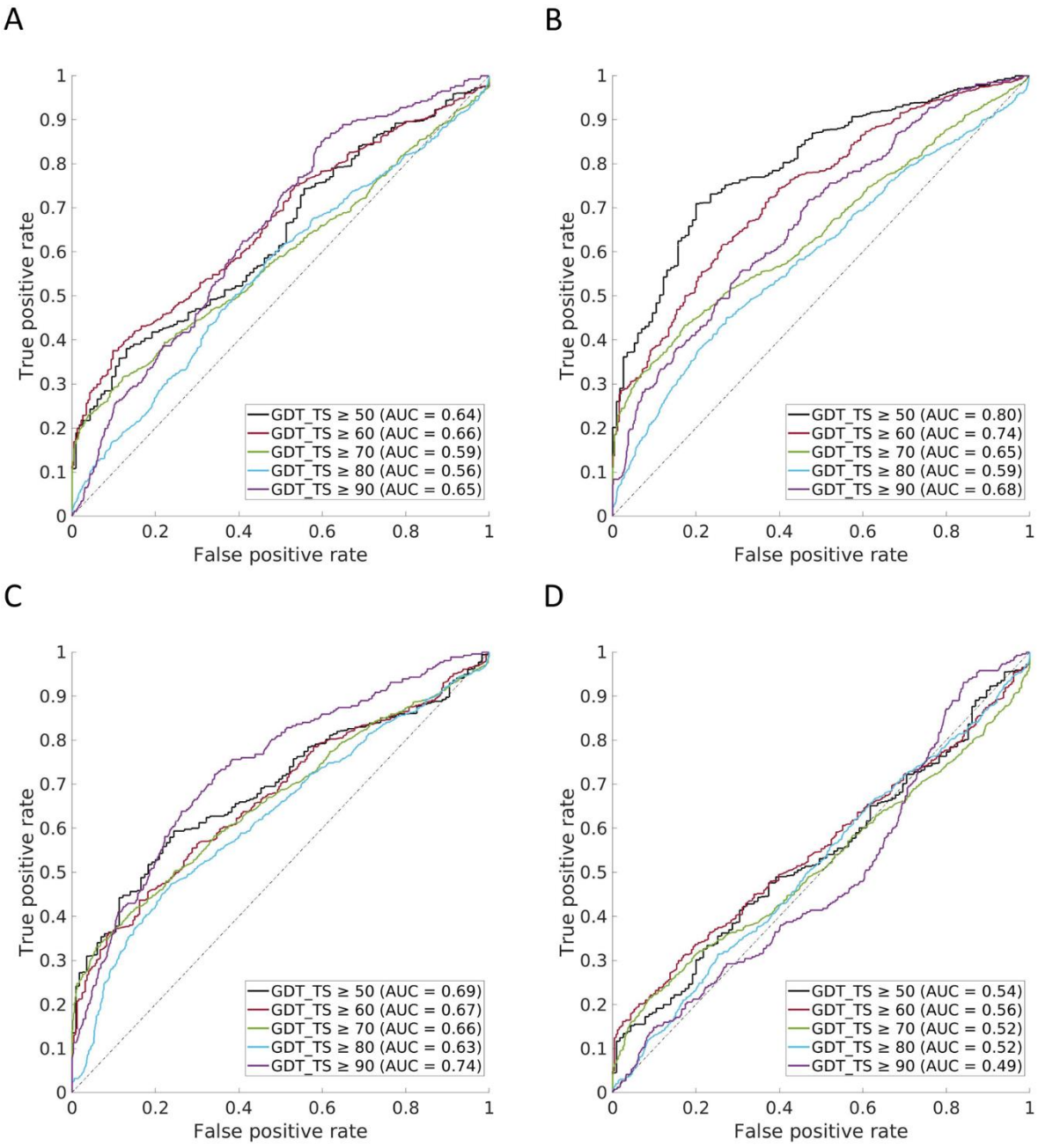


Fig. 5.11 Relationship between GDT_TS and statistical energies over a set of predicted CASP models and their corresponding native structures via ROC curves. **A-C**: SCEs vs GDT_TS. (A), (B), and (C) correspond to 1x1, 3x3, and 5x5 SCEs. **D**: CEs vs GDT_TS.

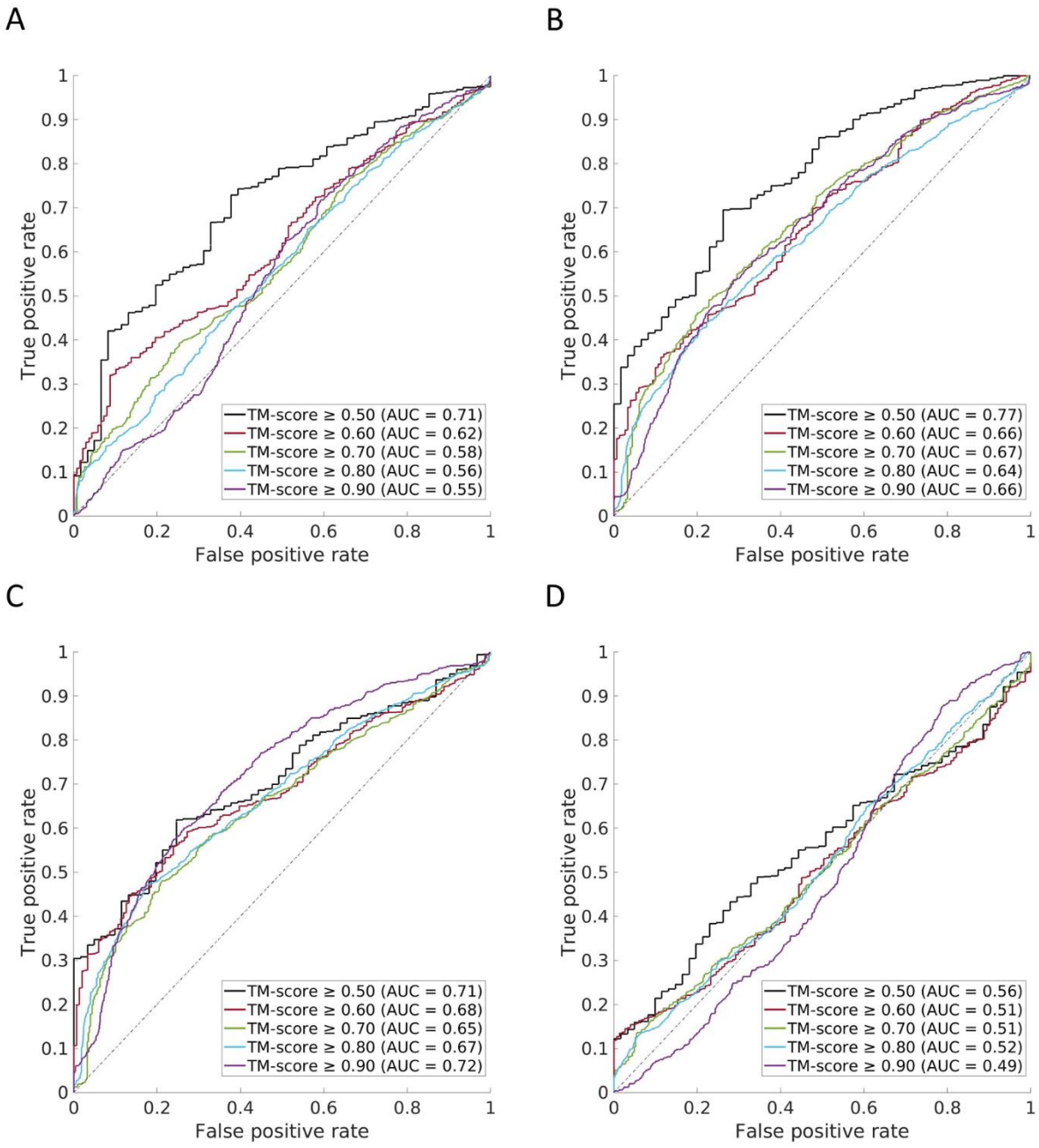


Fig. 5.12 Relationship between TM-score and statistical energies over a set of predicted CASP models and their corresponding native structures via ROC curves. **A-C**: SCEs vs TM-score. (A), (B), and (C) correspond to 1x1, 3x3, and 5x5 SCEs. **D**: CEs vs TM-score.

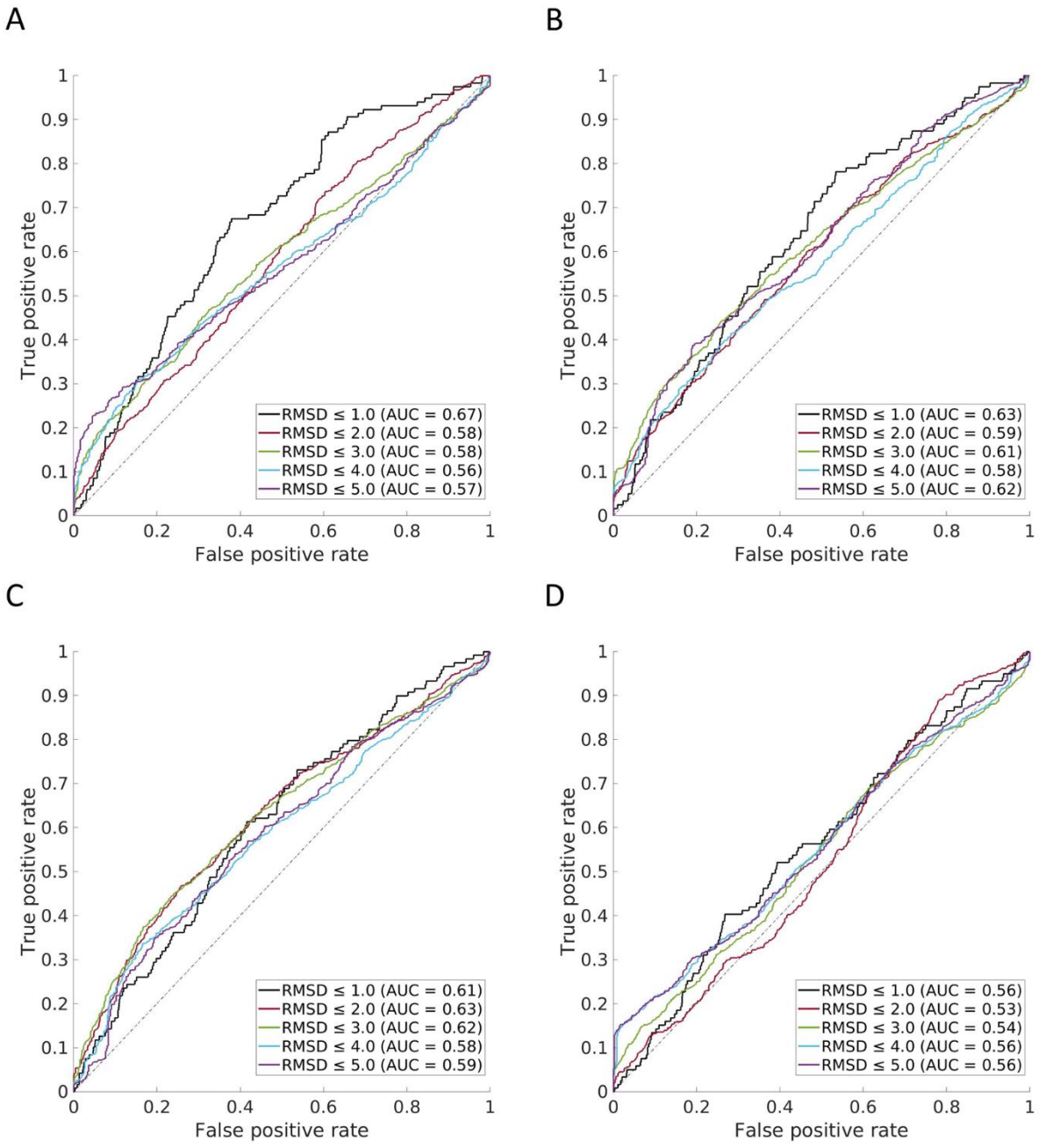


Fig. 5.13 Relationship between RMSD and statistical energies over a set of predicted CASP models and their corresponding native structures via ROC curves. **A-C**: SCEs vs RMSD. (A), (B), and (C) correspond to 1x1, 3x3, and 5x5 SCEs. **D**: CEs vs RMSD.

5.3 Discussion

The long history and widespread use of contact potentials suggests that extensions to the concept may also prove fruitful. Here, we demonstrated such an extension and showed that conditioning the contact potential's amino-acid pair statistics on the backbone conformation of the interacting fragment is a promising way to increase the relevance of the resultant statistical energies. Moreover, examining these structure-conditioned energies and the fragment geometries they are conditioned on has revealed a general relationship between structural similarity and energetic similarity.

The results summarized in Fig. 5.4 provide a good example of this increase in relevance, highlighting how the energies of the SCP contain more native sequence information than those of a traditional CP. Indeed, while the CP effectively encodes common patterns of amino-acid pair interactions—the most favorable energies after Cys-Cys being Glu-Lys, Asp-Lys, Arg-Asp, and Arg-Glu (all complementarily-charged amino-acid pairs which frequently interact)—it has no mechanism to recognize any detailed structural context in which these interactions occur. This can be seen as a limitation of CPs that is accepted in exchange for speed. Thus, a CP serves as a fast but very approximate measure of a structure's inter-residue interactions.

In contrast, the SCP described here performs much better. Importantly, one would not expect a contact potential of any kind, which inherently captures second-order sequence preferences only, to have a high sequence recovery. In fact, the most significant explanatory effect of sequence is expected to reside in the first-order contribution (e.g., preference for degree of burial, backbone dihedral angles, etc.). Further, when it comes to second-order effects, it is the sum of pair interactions involving a given residue that most matters for amino-acid choice at the position, while each pair contribution may play only a minor role. Despite this, the SCP stills exhibits considerable predictive performance in picking out native amino-acid pairs, with a rate of four times over that expected by chance. In terms of speed, the general relationship between structural similarity and energetic similarity shown in Fig. 5.8 implies that SCE matrices do not need to be computed for every fragment of interest. Instead, a database of energies can be pre-computed and the interaction motifs of interest can be evaluated on the fly by looking up the energies of the most similar motifs in the database. This would make the speed of

SCPs comparable to that of CPs, while providing a considerably more accurate sequence-structure linkage.

Additional evidence of the relevance of SCP-based energies is shown in Fig. 5.5, which demonstrates that SCE matrices are more closely related to thermodynamic coupling energies in coiled coils. While SCEs correlate more highly with experimental coupling values than CEs do using the "default" parameters, the highest correlations occur when the ensemble of motifs used to condition the amino-acid pair statistics were constrained to include only those fragments with very high structural similarity to the query motifs (i.e., the motifs centered on the a-a' and g-e' interaction pairs). This sensitivity to the chosen ensemble reinforces the point that structural context matters when estimating pairwise energies and suggests that using statistical energies as proxies for coupling energies requires knowledge of the native ensemble, not just knowledge of the crystallized conformation. While such knowledge would be rarely available, the SCP at least provides a means of tuning the statistics based on the predicted or assumed ensemble, which cannot be done with a CP. Moreover, in some cases such as sequence design, it may be useful to impose a desired ensemble, making this tuning capability of SCP convenient. Considering the results in Fig. 5.5 more broadly, it is interesting that the statistics in the PDB, when conditioned on an appropriate set of fragments, correspond even approximately with coupling energies, which are derived from measurements of thermodynamic equilibria in specific systems. This correspondence suggests that thermodynamic preferences contribute to the distribution of amino-acid pair statistics in the PDB.

Results in Figs. 5.10-13 corroborate those in Figs. 5.4 and 5.5 by directly evaluating how effective each energy function is at predicting the sequence-structure compatibility in structural models. The increase in AUCs (which quantify how well high-quality structures can be differentiated from low-quality ones) achieved by SCEs shows that incorporation of structural context helps in the evaluation of pairwise interactions in structural models. Importantly, the dataset involved in this analysis is expansive, comprising thousands of structural models submitted by CASP participants, and includes models derived from a variety of techniques tried over many years. The consistent increase in AUCs, whether using GDT_TS, TM-score, or RMSD, and regardless of the

threshold chosen to split high-quality and low-quality models, allows us to conclude that an SCP-based scoring function is reliably better than an equivalent CP-based one. This holds even when the SCEs are computed with the minimally-contextual 1x1 motifs, but the further increase when using 3x3 motifs confirms that the increase in performance is driven by incorporating additional structural context.

Perhaps the most striking observation about the SCP is the bidirectional relationship it reveals between contact geometry and pairwise sequence preferences, which is shown via clustering in Fig. 5.7 and more generally in Fig. 5.8. It may be expected that structurally similar pairs of interaction motifs induce similar SCE matrices. After all, these matrices are computed by collecting ensembles of structurally similar fragments and pooling their amino-acid pair statistics. However, it is interesting that the relationship holds in the other direction—that energetically similar pairs of motifs tend to be structurally similar as well. It could have been the case that most contact geometries would induce similar SCE matrices, which would have precluded using energetic similarity to predict structural similarity. This is not the case, however, and Fig. 5.7D makes it clear that structurally unrelated motifs are very unlikely to have similar SCE matrices. This coupling between structural and energetic similarity, the limits of which are quantified in Fig. 5.8, implies that not only does a particular contact geometry impose constraints on sequence preferences, but that particular sequence preferences impose constraints on the contact geometry. In effect, there appears to (usually) be at most only one way, in local structural space, to achieve a specific set of pair amino-acid preferences.

At a high level, our results show that local geometry-to-sequence mappings are inherently learnable, generalizable, and with a tight coupling between local structure and the pattern of amino-acid preferences. This fact may well be a part of the reason behind recent results having been able to achieve excellent generalization capabilities in going from multiple-sequence alignments to accurately predicted structures[17,121]. These results once again suggest that the amount of protein structural data amassed to date is sufficient to establish robust generalizations about sequence-structure relationships. Thus, continued exploitation of these data is likely to produce additional insights and

generalizations and may enable novel techniques for the design and modeling protein structure and properties.

5.4 Methods

5.4.1 Contact degree

Contacts in this study were defined using our previously described contact degree (CD) metric[7–9]. CD quantifies the extent to which a pair of positions is poised to host a contact by placing all possible rotamers (of all natural amino acids) at both positions and calculating the probability-weighted fraction of mutually exclusive rotamer pairs (i.e., those with clashing heavy atoms; see Holland *et al.*[8]).

5.4.2 Contact database creation

In order to sample a diverse distribution of native contacts, a high quality, non-redundant subset of the PDB was collected and around 200,000 contacts were sampled from it. In more detail, the PISCES server[46] was used to collect a non-redundant subset of the PDB. Only structures solved by X-ray crystallography were included, with the maximum resolution capped at 2.3 Å and the maximum R-value at 0.3. Structures were filtered by chain, keeping only those with between 40 and 10,000 residues, and with the maximum sequence identity of any pair restricted to 25%. The list of chains meeting these criteria was collected on January 1st, 2020 and resulted in 12,148 entries. Contacts were computed between every pair of residues in every chain. In order to sample inter-protein contacts, contacts were also computed between every pair of residues between each chain and every other chain in its PDB file. This resulted in 61,510,642 contacts in total. Contacts were then filtered to include only those for which both contacting residues had canonical amino acid names (with MSE being considered equivalent to MET), and enough sequence separation (at least 5 residues in between the two contacting ones) to ensure the largest considered interaction motifs (5x5-mers) were composed of two non-contiguous segments. Furthermore, a contact was included only if each residue involved in the 5x5-mer motif had all four backbone atoms. Because most contacts are weak, and we wanted to sample many strong contacts in addition to weak ones, the contact database

was created by sampling an equal number of contacts from 11 bins of contact degree: [0, 2^{-10}), [2^{-10}, 2^{-9}), [2^{-9}, 2^{-8}), ..., [2^{-1}, 1]. DB200K was created by sampling 18,182 contacts per bin, resulting in 200,002 contacts spanning 10,839 structures/complexes. The set of contacts comprising DB200K can be found in the "DB200K.tar.gz" file hosted on Zenodo[136].

5.4.3 Structure-conditioned potentials

An interaction motif is a structural fragment centered around a pair of contacting residues. The fragment may contain just the pair of interacting residues (1x1) or one or more flanking residues on each side of the pair (3x3, 5x5, etc.). Corresponding to the pair component of the dTERMen energy function, the structure-conditioned energy (SCE) matrix for an interaction motif is a 20x20 matrix containing log-transformed ratios of amino-acid pair observations over expectations, computed using the amino-acid statistics from an ensemble of fragments structurally similar to the interaction motif. Only the statistics of the two contacting residues are considered; the flanking residues control how much structural context is considered when collecting the ensemble of similar fragments, but their amino-acid identities do not contribute directly to the energies. The ensemble is collected by using the interaction motif to query into a structural database, finding all fragments (matches) of the same size as the query in the order of their backbone-atom root-mean-square deviation (RMSD) from the query motif. The search is limited by both the number of fragments returned (max count) and the maximal RMSD to the query (RMSD cutoff). Except for the SCEs used to compare with experimental coupling energies, which were recomputed using a wide variety of match counts (see Fig. 5.5 and the "Coupling Energies" section below), the max count was fixed at 50,000 for all SCEs computed here. The RMSD cutoff was set in a size-based manner according to our previously derived cutoff function (Eqs. 25 and 26 in the supplementary information of Zhou *et al.*[9]). This resulted in cutoffs of 1.0 Å, ~0.79 Å, and ~0.77 Å for 1x1, 3x3, and 5x5 motifs, respectively.

The equation below details how the 400 structure-conditioned energies (SCEs) of an SCP are computed for an interaction motif f centered around a pair of interacting positions (i, j):

$$\mathrm{SCE}(a_i, a_j) = -\log \frac{\mathrm{N}_{\mathrm{obs}}(a_i, a_j) + \varepsilon}{\mathrm{N}_{\mathrm{exp}}(a_i, a_j) + \varepsilon}$$

Eq. 5.3

Here, (a_i, a_j) is the amino-acid pair whose SCE is being computed at these positions (with a_i being the amino acid at position i and a_j the amino acid at position j), $\mathrm{N}_{\mathrm{obs}}(a_i, a_j)$ is the number of occurrences of pair (a_i, a_j) in f's ensemble of matching fragments, and $\mathrm{N}_{\mathrm{exp}}(a_i, a_j)$ is the number of occurrences of pair (a_i, a_j) that would be expected if there were no pair preferences. The pseudocount ε is set to $20 / \max(\mathrm{N}_{\mathrm{obs}}(a_i, a_j), \mathrm{N}_{\mathrm{exp}}(a_i, a_j), 1)$. Note the above equation is equivalent to Eq. 5.2 but with a replaced by a_i and b replaced by a_j for clarity about how i and j play a role. $\mathrm{N}_{\mathrm{exp}}(a_i, a_j)$ is defined as follows:

$$\mathrm{N}_{\mathrm{exp}}(a_i, a_j) = \sum_{m \in M} \frac{\exp(-\mathrm{E}_1(a_i \mid m_i) - \Delta_i(a_i, M))}{\sum_{a \in AA} \exp(-\mathrm{E}_1(a \mid m_i) - \Delta_j(a, M))} \times \frac{\exp(-\mathrm{E}_1(a_j \mid m_j) - \Delta_i(a_j, M))}{\sum_{a \in AA} \exp(-\mathrm{E}_1(a \mid m_j) - \Delta_j(a, M))}$$

Eq. 5.4

Here, m is a match in f's ensemble of matching fragments M, $\mathrm{E}_1(a_i \mid m_i)$ is the pseudo-energy associated with amino acid a_i at position i in match m, AA is the set of the 20 natural amino acids, and $\Delta_i(a_i, M)$ is a residual energy associated with amino acid a_i that is set (for each motif) to ensure that the expected marginal counts of each amino acid at each position coincide with observed counts. See the "Pair contributions" section of the supplementary information of Zhou *et al.*[9] for more details. The pseudo-energy in E_1 is a first-order energy model that takes into account the backbone dihedral angles φ, ψ, and ω and environment to estimate how favorable each amino acid is at a given position. See the "Pre-computed contributions" section of the supplementary information of Zhou *et al.*[9] for more information about how these pseudo-energies are computed. The first term under the outer sum is thus a ratio between how likely (according to E_1) amino acid a_i is to occur at position i and the sum of the likelihoods over all possible amino acids at this

position. The second term under the outer sum computes the same ratio but for a_j at position j. Multiplied together, these two terms capture the probability of observing pair (a_i, a_j) in match m under a model that knows about first-order amino-acid preferences but assumes no second-order dependencies. This means the sum over each match in the ensemble M is the expected number of observations in M of the amino acid pair (a_i, a_j) if there were no second-order (or other higher-order) dependencies. Thus, the ratio between $N_{obs}(a_i, a_j)$ and $N_{exp}(a_i, a_j)$ estimates to what extent true observations exhibit apparent correlations.

5.4.4 Contact potential

The contact potential was computed for every canonical amino-acid pair using Eq. 5.1, with DB200K used as the set of contacts. Since by construction, the energy is invariant to the order of the amino-acid pairs, with $E(a, b) = E(b, a)$, the number of observed pairs $N_{obs}(a, b)$ must take this into account by summing (a, b) and (b, a) contacts together:

$$N_{obs}(a, b) = N_{obs}(b, a) = N(a, b) + (1 - \mathbb{I}(a, b)) \cdot N(b, a)$$

Eq. 5.5

Here, $N(a, b)$ is the number of contacts in the database between a and b in the order (a, b), and $\mathbb{I}(a, b)$ is 1 if $a=b$ and 0 otherwise, ensuring homotypic contacts are not counted twice. To compute the number of expected pairs $N_{exp}(a, b)$, the expectation for heterotypic pairs must be doubled so that, in accordance with the chosen reference state, in the absence of amino-acid pair preferences between a and b, $N_{obs}(a, b) = N_{exp}(a, b)$, thereby making $E(a, b) = 0$:

$$N_{exp}(a, b) = N_{exp}(b, a) = N_{obs}(a) \cdot N_{obs}(b) \cdot H(a, b) / N$$

$$H(a, b) = 2 - \mathbb{I}(a, b)$$

Eq. 5.6

The empirically derived amino acid-dependent pseudocount ε was structured identically to the one used for the SCP (see the "Structure-conditioned potentials" section above and Zhou *et al.*[9] for details), although in the case of the contact potential, the large set of contacts used to compute the statistics ensured there was no data sparsity and thus the values were negligible.

5.4.4 AA pair identification

The modified Z-score used in Fig. 5.4B is computed by analogy to a traditional Z-score, but with the mean replaced by the median and the standard deviation replaced by the median absolute deviation (MAD). Below, E is the set of 400 energies in an SCP and E_i is the one of these energies whose modified Z-score is being computed:

$$Z_{\mathrm{mod}}(E_i) = \frac{E_i - \mathrm{median}(E)}{\mathrm{MAD}(E_i)} = \frac{E_i - \mathrm{median}(E)}{\mathrm{median}(\left|E_i - \mathrm{median}(E)\right|)}$$

Eq. 5.7

5.4.5 Coupling energies

Experimentally determined coupling energies for the a-a' and g-e' interhelical interactions were taken from Table 4 and Table III from Acharya *et al.*[127] and Krylov *et al.*[76], respectively. Since experimentally solved structures for the coiled coils were not available, they were modeled using CCFold[137] based on the sequences specified in the papers, trimming off N- and C-terminal residues distal to the interaction. In particular, for the system used for measuring the a-a' interactions, the sequences used were RAAFLEKENTALRTRLAELRKRVGRCRNIVSKYETRYG (chain A) and RAAFLEKENTALRTELAELEKEVGRCENIVSKYETRYG (chain B), with the residue pair (A16, B16) used to compute the energies (note that these residues, both labeled as X in the sequence in Acharya *et al.*[127], were replaced with leucine when modeled with CCFold). For the system used for measuring the g-e' interactions, the sequence used for both chains was KVFVPDEQKDEKYWTRRKKNNVAAKRSRDARRLKENQITIRAAFLEK

ENTALRTEVAELRKEVGRCKNIVSKYETRYGPL, with the residue pair (A41, B46) used to compute the energies.

5.4.6 Clustering

Each clustering was performed by randomly sampling without replacement a set *S* of 50,000 motifs from DB200K and running 100 rounds of an in-house greedy clustering method which accepts two parameters, a distance cutoff *d* and a sampling count *n*, and returns one cluster per round. In any round *i*, *n* motifs are randomly sampled without replacement from *S*. The distance between each pair of *n* motifs is computed and the motif with the largest number of distances of at most *d* to the other sampled motifs is chosen as the cluster representative *r*. The distance between r and each motif in S is computed and every motif with a distance of at most *d* is included in the returned cluster *C*. For round *i*+1, $S = S \setminus C$ (i.e., the elements assigned to the *i*th cluster are not available for rounds *i*+1, *i*+2, …). When clustering by structure, the distance metric used was best-fit RMSD and *d* was chosen to be 0.5. When clustering by SCE, the distance metric used was $r_E = 1 - r$, where *r* is the linear correlation coefficient between the pair of SCE matrices, and *d* was chosen to be 0.3. Note that the chosen values of d, 0.5 when clustering by structure and 0.3 when clustering by energy, result in the top 100 clusters including a similar number of motifs, making the comparisons of distributions in Fig. 5.7D and 5.7E fair. To make the sets of random clusters a similarly fair control, the clusters were chosen to have the same number of elements as the structure clusters (in Fig. 5.7D) and the energy clusters (in Fig. 5.7E) as well.

5.4.7 CASP model evaluation

All publicly available refinement targets from CASP9-14 were considered, with the solved structure and up to 20 models included per target. To sample a set of structures with a wide range of structure quality, the 20 models included for each target were selected by sorting the models best-to-worst (by GDT_TS) and alternately adding the next-best and next-worst models remaining until either 20 were added or no more models were available. The set of targets and their included models is listed in the "CASP-models.xlsx" file hosted on Zenodo[136]. Structures and GDT_TS scores were taken from

the CASP website (https://predictioncenter.org/). TM-scores and RMSDs were computed using the TMscore program[135] with the default settings. SCEs were computed over every contact with a CD of at least 0.1.

5.4 Acknowledgements

This chapter is adapted from a paper that has been submitted to Protein Science. This work was supported by NIH award R01-GM132117 and NSF award DMR1534246 (GG)

6 Conclusions

6.1 Residue-level statistical potentials

Much of my work has focused on residue-level statistical potentials, whether in the form of a simple, binary contact potential (Chapter 2), a more intricate, hierarchical statistical potential (Chapter 3), or a contact potential conditioned on structural context (Chapter 5). There are distinct advantages to sticking to residue-level potentials, such as simplicity and interpretability. Describing how pairs of amino acid types interact with each other is of fundamental interest to protein science and the conclusions are easy to understand. In fact, it is remarkable how much about protein structure can be understood at just the level of residues and residue pairs given how many diverse conformations most types of side-chain can occupy.

Discovering that a simple residue-level contact potential could boost the performance of contact prediction methods (Fig. 2.4) was surprising, and conditioning such a potential on structural context revealed interesting relationships between sequence and structure (Figs. 5.3-5.5). However, residue-level statistical potentials of all kinds have significant limitations. The particular conformations of a protein's side-chains clearly do matter, and so not all interactions can be understood in the framework of these potentials. Atomic-level potentials such as DFIRE[67] and GOAP[50] approach this problem by delving into the atomic-level details of protein structure and have had many successes doing so, although they lose some of the simplicity and interpretability in the process. Taking a very different tack, many recent methods instead focus on relating sequence to

structure through machine learning, in particular neural networks. It will be interesting to see whether neural networks can address the limitations of residue-level potentials by putting them into the right context. Can the predictions of residue-level potentials be made more accurate by considering the larger context? In other words, can the energies of particular interactions be refined by examining the interactions and structure around it, as neural networks are poised to do? It is not clear that atomic-level energies are necessary if sufficient context is incorporated into residue-level energies. The structure-conditioned potential discussed in Chapter 5 is an exciting foray into this idea, and it is clear that conditioning on structural context opens up new possibilities for residue-level potentials, but the concept can be taken further, especially in light of these recent developments in neural networks.

Taking a broader view, it is interesting to note how dependent the structure-conditioned potential is on recently developments. Without the extraordinary growth of the PDB and structural search tools, it would not have been feasible to investigate the effect of structural context on amino acid pair preferences. Even now, the time it takes to search for relevant fragments hinders the applicability of structural searches; imagine what could be done if, say, a neural network could be trained to accurately predict the pair energies of any given interaction motif in a matter of moments. It would then be trivial to construct much larger databases of statistical energies and perhaps draw deeper conclusions about the relationship between sequence and structure from them. While I do not know which yet-to-be-developed tools will most significantly impact our understanding of pairwise sequence-structure relationships, it is clear that the growth of the PDB and the development of tools to mine its information will continue to limit what kinds of statistical potentials are feasible to construct and study. While there are certainly many fruitful ways to better understand pairwise sequence-structure relationships, with the creation of more detailed and intricate statistical potentials among them, one of the most impactful may be to develop structural tools to aid those studying such relationships. The Protein Builder could be seen as one of these tools, and I expect users other than myself to find ways of using it that I have not anticipated.

6.2 Tertiary fragments

The other primary focus of my work has been the study of tertiary fragments and how they relate sequence to structure. The Protein Builder (Chapter 4) started as a simple demo of how tertiary fragments could be pieced together by exploiting empirically known overlaps. The richness of the overlap space encouraged further work on the project, and the final result showcases not only how many ways tertiary fragments can be pieced together into an assembly but how this assembly, through the fairly simple optimization process of fusion, can be turned into a novel backbone. While the creation of novel backbones is certainly not new, and there are many techniques to do so, the interactivity and intuition-based style of the tool is a distinct twist on the process and will hopefully encourage the creation of exotic (or even baroque) topologies and interaction networks. One of the major limitations of the original conception—the insistence that every fragment come from a representative database, which no matter how representative could never cover all of structure space—turned out to have a fairly simple solution, bridging. Enabling loop closure via bridging allow the database to focus on interaction motifs without needing an exorbitant number of linear fragments to close every possible loop geometry. This fits nicely with the overarching thesis of the project, which is to center backbone creation around interaction geometry, and acknowledges that loop closure is already a well-studied and successful subfield of backbone creation which does not need tertiary fragments to solve. Overall, the Protein Builder is an appealing demonstration of the discretized and geometric approach to structure I have been arguing for in this thesis.

While the Protein Builder utilized tertiary fragments in a primarily geometric way, slotting them together to create novel structures, the structure-conditioned potential (Chapter 5) conceptualized them as a means of probing sequence-structure relationships. The most fascinating finding of the study was the tight, bidirectional relationship between interaction geometry and sequence preferences (encoded as statistical energies). While the structure to sequence direction was expected (being the basis for, e.g., dTERMen), the sequence to structure connection was a surprise, in particular the results shown in Fig. 5.4G. This is an instance of where the simplicity of the approach—structure-conditioned potentials are fundamentally just counting occurrences of sequence pairs in the context of particular geometries—benefits it greatly. If the energies were derived in a more intricate,

elaborate way, then it would be plausible that the energies were directly encoding information about structure without encoding fundamental sequence-structure relationships (e.g., an overtrained neural network encoding the coordinates in a convoluted way that does not generalize to unseen interaction motifs), but the fact that these energies come from a fairly simple counting procedure effectively precludes this.

One of the more intriguing observations structure-conditioned potentials revealed was that coiled-coil coupling energies are best reflected in very low-RMSD ensembles (Fig. 5.3A,C), in contrast to the much larger ensembles used in the rest of the project, which more reliably recover native sequence statistics. This is another area plagued by technical limitations, which the continued growth of the PDB should alleviate. The low-RMSD ensembles do not appear to have rich enough statistics to use in general, with the very common alpha-helical pair motifs being an exception to this. If low-RMSD ensembles were more densely populated, it could be feasible to study more particular interaction geometries, perhaps elucidating how the exact backbone coordinates affect the permissible side-chain conformations and their interactions.

6.3 Outlook

While the physics of proteins is exceedingly complex, the ability to conceptualize them as geometric objects whose patterns can be understood without explicitly engaging with the underlying physics continues to be a promising approach to predicting their behavior. Discretizing this geometry into an alphabet of motifs is a more recent endeavor, but it also seems to be bearing fruit, with my work adding to a number of projects which seek to elucidate sequence-structure relationships using these motifs. While this approach has answered questions, it has also raised even more. What is the best way to discretize structure space, if even the fairly straightforward method of extracting motifs around contacts works as well as it does? To what extent are higher order motifs which encompass many interactions necessary to understand fundamental sequence-structure relationships? How much context is needed to reliably predict sequence from structure and structure from sequence? There is a combinatorial explosion of possibilities in this space which we have only begun to explore. Recent advances in machine learning make this space even more exciting, as the ability to synthesize many disparate pieces of

information (such as tertiary fragments and their statistics) into accurate predictions should extend the reach of these geometric objects and their residue-level statistics. It is an exciting time for protein science and I expect a discretized and geometric approach to structure to play an integral role.

References

1. Sippl, M. J. Calculation of conformational ensembles from potentials of mena force. *J. Mol. Biol.* **213**, 859–883 (1990).
2. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95–99 (1963).
3. MacArthur, M. W. & Thornton, J. M. Deviations from Planarity of the Peptide Bond in Peptides and Proteins. *J. Mol. Biol.* **264**, 1180–1195 (1996).
4. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci.* **108**, E1293–E1301 (2011).
5. Zheng, F., Zhang, J. & Grigoryan, G. Tertiary Structural Propensities Reveal Fundamental Sequence/Structure Relationships. *Structure* **23**, 961–971 (2015).
6. Zhou, J. & Grigoryan, G. Rapid search for tertiary fragments reveals protein sequence-structure relationships: Tertiary Motif Search gives Structure Rules. *Protein Sci.* **24**, 508–524 (2015).
7. Mackenzie, C. O., Zhou, J. & Grigoryan, G. Tertiary alphabet for the observable protein structural universe. *Proc. Natl. Acad. Sci.* **113**, E7438–E7447 (2016).
8. Holland, J., Pan, Q. & Grigoryan, G. Contact prediction is hardest for the most informative contacts, but improves with the incorporation of contact potentials. *PLOS ONE* **13**, e0199585 (2018).
9. Zhou, J., Panaitiu, A. E. & Grigoryan, G. A general-purpose protein design framework based on mining sequence–structure relationships in known protein structures. *Proc. Natl. Acad. Sci.* **117**, 1059–1068 (2020).

10. Monod, J., Wyman, J. & Changeux, J.-P. On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.* **12**, 88–118 (1965).

11. Lovell, S. C., Word, J. M., Richardson, J. S. & Richardson, D. C. The penultimate rotamer library. *Proteins Struct. Funct. Bioinforma.* **40**, 389–408 (2000).

12. Shapovalov, M. V. & Dunbrack, R. L. A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure* **19**, 844–858 (2011).

13. Chen, R., Li, L. & Weng, Z. ZDOCK: An initial-stage protein-docking algorithm. *Proteins Struct. Funct. Genet.* **52**, 80–87 (2003).

14. Orengo, C. *et al.* CATH – a hierarchic classification of protein domain structures. *Structure* **5**, 1093–1109 (1997).

15. Ahnert, S. E., Marsh, J. A., Hernandez, H., Robinson, C. V. & Teichmann, S. A. Principles of assembly reveal a periodic table of protein complexes. *Science* **350**, aaa2245–aaa2245 (2015).

16. Fernandez-Fuentes, N. & Fiser, A. A modular perspective of protein structures; application to fragment based loop modeling. *Methods Mol. Biol. Clifton NJ* **932**, 141–158 (2013).

17. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* (2021) doi:10.1038/s41586-021-03819-2.

18. Fitch, W. M. & Markowitz, E. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem. Genet.* **4**, 579–593 (1970).

19. Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I. & Langmead, C. J. Learning generative models for protein fold families. *Proteins Struct. Funct. Bioinforma.* **79**, 1061–1078 (2011).

20. Göbel, U., Sander, C., Schneider, R. & Valencia, A. Correlated mutations and residue contacts in proteins. *Proteins Struct. Funct. Bioinforma.* **18**, 309–317 (1994).

21. Shindyalov, I. N., Kolchanov, N. A. & Sander, C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng. Des. Sel.* **7**, 349–358 (1994).

22. Taylor, W. R. & Hatrick, K. Compensating changes in protein multiple sequence alignments. *Protein Eng. Des. Sel.* **7**, 341–348 (1994).

23. Martin, L. C., Gloor, G. B., Dunn, S. D. & Wahl, L. M. Using information theory to search for co-evolving residues in proteins. *Bioinformatics* **21**, 4116–4124 (2005).

24. Dunn, S. D., Wahl, L. M. & Gloor, G. B. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **24**, 333–340 (2008).

25. Jones, D. T., Buchan, D. W. A., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28**, 184–190 (2012).

26. Olmea, O. & Valencia, A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold. Des.* **2**, S25–S32 (1997).

27. Pollock, D. D. & Taylor, W. R. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Eng. Des. Sel.* **10**, 647–657 (1997).

28. Lapedes, A. S., Giraud, B., Liu, L. & Stormo, G. D. Correlated mutations in models of protein sequences: phylogenetic and structural effects. *Stat. Mol. Biol. Genet.* **33**, 236–257 (1999).

29. Fares, M. A. & Travers, S. A. A. A Novel Method for Detecting Intramolecular Coevolution: Adding a Further Dimension to Selective Constraints Analyses. *Genetics* **173**, 9–23 (2006).

30. Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci.* **110**, 15674–15679 (2013).

31. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *eLife* **3**, e02030 (2014).

32. Olmea, O., Rost, B. & Valencia, A. Effective use of sequence correlation and conservation in fold recognition 1 1Edited by J. M. Thornton. *J. Mol. Biol.* **293**, 1221–1239 (1999).

33. Gao, X., Bu, D., Xu, J. & Li, M. Improving consensus contact prediction via server correlation reduction. *BMC Struct. Biol.* **9**, 28 (2009).

34. Ovchinnikov, S. *et al.* Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta. *Proteins Struct. Funct. Bioinforma.* **84**, 67–75 (2016).

35. Jones, D. T., Singh, T., Kosciolek, T. & Tetchner, S. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* **31**, 999–1006 (2015).

36. Pokarowski, P. *et al.* Inferring ideal amino acid interaction forms from statistical protein contact potentials. *Proteins Struct. Funct. Bioinforma.* **59**, 49–57 (2005).

37. Vendruscolo, M. & Domany, E. Pairwise contact potentials are unsuitable for protein folding. *J. Chem. Phys.* **109**, 11101–11108 (1998).

38. Skolnick, J. In quest of an empirical potential for protein structure prediction. *Curr. Opin. Struct. Biol.* **16**, 166–171 (2006).

39. Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLOS Comput. Biol.* **13**, e1005324 (2017).

40. Liu, Y., Palmedo, P., Ye, Q., Berger, B. & Peng, J. Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks. *Cell Syst.* **6**, 65-74.e3 (2018).

41. Stahl, K., Schneider, M. & Brock, O. EPSILON-CP: using deep learning to combine information from multiple sources for protein contact prediction. *BMC Bioinformatics* **18**, 303 (2017).

42. Miyazawa, S. & Jernigan, R. L. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534–552 (1985).

43. Zhang, C. & Kim, S.-H. Environment-dependent residue contact energies for proteins. *Proc. Natl. Acad. Sci.* **97**, 2550–2555 (2000).

44. Sippl, M. J. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* **5**, 229–235 (1995).

45. Jernigan, R. L. & Bahar, I. Structure-derived potentials and protein simulations. *Curr. Opin. Struct. Biol.* **6**, 195–209 (1996).

46. Wang, G. & Dunbrack, R. L. PISCES: a protein sequence culling server. *Bioinformatics* **19**, 1589–1591 (2003).

47. Yang, J. *et al.* The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **12**, 7–8 (2015).

48. Simons, K. T. *et al.* Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* **34**, 82–95 (1999).

49. Lu, M., Dousis, A. D. & Ma, J. OPUS-PSP: An Orientation-dependent Statistical All-atom Potential Derived from Side-chain Packing. *J. Mol. Biol.* **376**, 288–301 (2008).

50. Zhou, H. & Skolnick, J. GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction. *Biophys. J.* **101**, 2043–2052 (2011).

51. Olson, M. A. & Lee, M. S. Structure refinement of protein model decoys requires accurate side-chain placement. *Proteins Struct. Funct. Bioinforma.* **81**, 469–478 (2013).

52. Mirzaie, M. & Sadeghi, M. Delaunay-based nonlocal interactions are sufficient and accurate in protein fold recognition. *Proteins Struct. Funct. Bioinforma.* **82**, 415–423 (2014).

53. Ruiz-Blanco, Y. B. *et al.* A physics-based scoring function for protein structural decoys: Dynamic testing on targets of CASP-ROLL. *Chem. Phys. Lett.* **610–611**, 135–140 (2014).

54. Zhou, J., Yan, W., Hu, G. & Shen, B. SVR_CAF: An integrated score function for detecting native protein structures among decoys. *Proteins Struct. Funct. Bioinforma.* **82**, 556–564 (2014).

55. Hoque, M. T., Yang, Y., Mishra, A. & Zhou, Y. sDFIRE: Sequence-specific statistical energy function for protein structure prediction by decoy selections. *J. Comput. Chem.* **37**, 1119–1124 (2016).

56. Zhang, J. & Zhang, Y. A Novel Side-Chain Orientation Dependent Potential Derived from Random-Walk Reference State for Protein Fold Selection and Structure Prediction. *PLOS ONE* **5**, e15386 (2010).

57. Sułkowska, J. I., Morcos, F., Weigt, M., Hwa, T. & Onuchic, J. N. Genomics-aided structure prediction. *Proc. Natl. Acad. Sci.* **109**, 10340–10345 (2012).

58. Morcos, F., Jana, B., Hwa, T. & Onuchic, J. N. Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc. Natl. Acad. Sci.* **110**, 20533–20538 (2013).

59. Morcos, F., Schafer, N. P., Cheng, R. R., Onuchic, J. N. & Wolynes, P. G. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proc. Natl. Acad. Sci.* **111**, 12408–12413 (2014).

60. dos Santos, R. N., Morcos, F., Jana, B., Andricopulo, A. D. & Onuchic, J. N. Dimeric interactions and complex formation using direct coevolutionary couplings. *Sci. Rep.* **5**, 13652 (2015).

61. Kaján, L., Hopf, T. A., Kalaš, M., Marks, D. S. & Rost, B. FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics* **15**, 85 (2014).

62. Seemayer, S., Gruber, M. & Söding, J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics* **30**, 3128–3130 (2014).

63. Buchan, D. W. A. & Jones, D. T. Improved protein contact predictions with the MetaPSICOV2 server in CASP12. *Proteins Struct. Funct. Bioinforma.* **86**, 78–83 (2018).

64. He, B., Mortuza, S. M., Wang, Y., Shen, H.-B. & Zhang, Y. NeBcon: protein contact map prediction using neural network training coupled with naïve Bayes classifiers. *Bioinformatics* **33**, 2296–2306 (2017).

65. Kim, D. E., DiMaio, F., Wang, R. Y.-R., Song, Y. & Baker, D. One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. *Proteins Struct. Funct. Bioinforma.* **82**, 208–218 (2014).

66. Baker, D. Rosetta Decoy Datasets. (2016) doi:10.5281/zenodo.48780.

67. Zhou, H. & Zhou, Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* **11**, 2714–2726 (2009).

68. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic Acids Res.* **33**, W382–W388 (2005).

69. Eramian, D. *et al.* A composite score for predicting errors in protein structure models. *Protein Sci.* **15**, 1653–1666 (2006).

70. Samudrala, R. & Levitt, M. Decoys 'R' Us: A database of incorrect conformations to improve protein structure prediction. *Protein Sci.* **9**, 1399–1401 (2000).

71. Slovic, A. M., Lear, J. D. & DeGrado, W. F. De novo design of a pentameric coiled-coil: decoding the motif for tetramer versus pentamer formation in water-soluble phospholamban*. *J. Pept. Res.* **65**, 312–321 (2005).

72. Strauch, E.-M., Fleishman, S. J. & Baker, D. Computational design of a pH-sensitive IgG binding protein. *Proc. Natl. Acad. Sci.* **111**, 675–680 (2014).

73. Huang, P.-S. *et al.* De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol.* **12**, 29–34 (2016).

74. Marcos, E. *et al.* De novo design of a non-local β-sheet protein with high stability and accuracy. *Nat. Struct. Mol. Biol.* **25**, 1028–1034 (2018).

75. Broom, A. *et al.* Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico. *Nat. Commun.* **11**, 4808 (2020).

76. Krylov, D., Mikhailenko, I. & Vinson, C. A thermodynamic scale for leucine zipper stability and dimerization specificity: e and g interhelical interactions. *EMBO J.* **13**, 2849–2861 (1994).

77. Jones, D. T. De novo protein design using pairwise potentials and a genetic algorithm: De novo protein design. *Protein Sci.* **3**, 567–574 (1994).

78. Canutescu, A. A. & Dunbrack, R. L. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci.* **12**, 963–972 (2003).

79. Kolodny, R., Guibas, L., Levitt, M. & Koehl, P. Inverse Kinematics in Biology: The Protein Loop Closure Problem. *Int. J. Robot. Res.* **24**, 151–163 (2005).

80. Yanover, C., Fromer, M. & Shifman, J. M. Dead-end elimination for multistate protein design. *J. Comput. Chem.* **28**, 2122–2129 (2007).

81. Lee, J., Lee, D., Park, H., Coutsias, E. A. & Seok, C. Protein loop modeling by using fragment assembly and analytical loop closure. *Proteins Struct. Funct. Bioinforma.* **78**, 3428–3436 (2010).

82. Mitra, P., Shultis, D. & Zhang, Y. EvoDesign: de novo protein design based on structural and evolutionary profiles. *Nucleic Acids Res.* **41**, W273–W280 (2013).

83. Jacobs, T. M. *et al.* Design of structurally distinct proteins using strategies inspired by evolution. *Science* **352**, 687–690 (2016).

84. Chevalier, A. *et al.* Massively parallel de novo protein design for targeted therapeutics. *Nature* **550**, 74–79 (2017).

85. Koepnick, B. *et al.* De novo protein design by citizen scientists. *Nature* **570**, 390–394 (2019).

86. O'Meara, M. J. *et al.* Combined Covalent-Electrostatic Model of Hydrogen Bonding Improves Structure Prediction with Rosetta. *J. Chem. Theory Comput.* **11**, 609–622 (2015).

87. Anishchenko, I., Chidyausiku, T. M., Ovchinnikov, S., Pellock, S. J. & Baker, D. De novo protein design by deep network hallucination. 2020.07.22.211482 (2020).

88. Singh, A. Bottom-up de novo protein design. *Nat. Methods* **18**, 233–233 (2021).

89. Kleffner, R. *et al.* Foldit Standalone: a video game-derived protein structure manipulation interface using Rosetta. *Bioinformatics* **33**, 2765–2767 (2017).

90. Gonzalez, G., Hannigan, B. & DeGrado, W. F. A Real-Time All-Atom Structural Search Engine for Proteins. *PLOS Comput. Biol.* **10**, e1003750 (2014).

91. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.

92. Flask. https://flask.palletsprojects.com/en/2.0.x/.

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

94. Bauer, A. & Beyer, A. An improved pair potential to recognize native protein folds. *Proteins Struct. Funct. Genet.* **18**, 254–261 (1994).

95. Thomas, P. D. & Dill, K. A. Statistical Potentials Extracted From Protein Structures: How Accurate Are They? *J. Mol. Biol.* **257**, 457–469 (1996).

96. Huang, E. S., Subbiah, S., Tsai, J. & Levitt, M. Using a Hydrophobic Contact Potential to Evaluate Native and Near-native Folds Generated by Molecular Dynamics Simulations. *J. Mol. Biol.* **257**, 716–725 (1996).

97. Hamelryck, T. *et al.* Potentials of Mean Force for Protein Structure Prediction Vindicated, Formalized and Generalized. *PLoS ONE* **5**, e13714 (2010).

98. Shen, M. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15**, 2507–2524 (2006).

99. Zhao, F. & Xu, J. A Position-Specific Distance-Dependent Statistical Potential for Protein Structure and Functional Study. *Structure* **20**, 1118–1126 (2012).

100. Peng, J. & Xu, J. Raptorx: Exploiting structure information for protein alignment by statistical inference. *Proteins Struct. Funct. Bioinforma.* **79**, 161–171 (2011).

101. Gilis, D. & Rooman, M. PoPMuSiC, an algorithm for predicting protein mutant stability changes. Application to prion proteins. *Protein Eng. Des. Sel.* **13**, 849–856 (2000).

102. Dehouck, Y. *et al.* Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics* **25**, 2537–2543 (2009).

103. Wu, S., Skolnick, J. & Zhang, Y. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol.* **5**, 17 (2007).

104. Sankar, K., Jia, K. & Jernigan, R. L. Knowledge-based entropies improve the identification of native protein structures. *Proc. Natl. Acad. Sci.* **114**, 2928–2933 (2017).

105. Chen, J. R., Chang, B. H., Allen, J. E., Stiffler, M. A. & MacBeath, G. Predicting PDZ domain–peptide interactions from primary sequences. *Nat. Biotechnol.* **26**, 1041–1045 (2008).

106. López-Blanco, J. R. & Chacón, P. KORP: knowledge-based 6D potential for fast protein and loop modeling. *Bioinformatics* **35**, 3013–3019 (2019).

107. Zhang, C. An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. *Protein Sci.* **13**, 400–411 (2004).

108. Eyrich, V. A., Standley, D. M., Felts, A. K. & Friesner, R. A. Protein tertiary structure prediction using a branch and bound algorithm. *Proteins Struct. Funct. Bioinforma.* **35**, 41–57 (1999).

109. Rykunov, D. & Fiser, A. New statistical potential for quality assessment of protein models and a survey of energy functions. *BMC Bioinformatics* **11**, 128 (2010).

110. Makino, Y. & Itoh, N. A knowledge-based structure-discriminating function that requires only main-chain atom coordinates. *BMC Struct. Biol.* **8**, 46 (2008).

111. Mukherjee, A., Bhimalapuram, P. & Bagchi, B. Orientation-dependent potential of mean force for protein folding. *J. Chem. Phys.* **123**, 014901 (2005).

112. Buchete, N.-V., Straub, J. E. & Thirumalai, D. Anisotropic coarse-grained statistical potentials improve the ability to identify nativelike protein structures. *J. Chem. Phys.* **118**, 7658 (2003).

113. Arab, S., Sadeghi, M., Eslahchi, C., Pezeshk, H. & Sheari, A. A pairwise residue contact area-based mean force potential for discrimination of native protein structure. *BMC Bioinformatics* **11**, 16 (2010).

114. Vijayakumar, M. & Zhou, H.-X. Prediction of Residue−Residue Pair Frequencies in Proteins. *J. Phys. Chem. B* **104**, 9755–9764 (2000).

115. Melo, F., Sánchez, R. & Sali, A. Statistical potentials for fold assessment. *Protein Sci.* **11**, 430–448 (2009).

116. Nanias, M., Chinchio, M., Pillardy, J., Ripoll, D. R. & Scheraga, H. A. Packing helices in proteins by global optimization of a potential energy function. *Proc. Natl. Acad. Sci.* **100**, 1706–1710 (2003).

117. Zimmer, R., Wohler, M. & Thiele, R. New scoring schemes for protein fold recognition based on Voronoi contacts. *Bioinformatics* **14**, 295–308 (1998).

118. Yuan, C., Chen, H. & Kihara, D. Effective inter-residue contact definitions for accurate protein fold recognition. *BMC Bioinformatics* **13**, 292 (2012).

119. Chhajer, M. & Crippen, G. M. A protein folding potential that places the native states of a large number of proteins near a local minimum. *BMC Struct. Biol.* **2**, 4 (2002).

120. Jones, D. T. & Kandathil, S. M. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics* **34**, 3308–3315 (2018).

121. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

122. Egbert, M. *et al.* Assessing the binding properties of CASP14 targets and models. *Proteins Struct. Funct. Bioinforma.* prot.26209 (2021) doi:10.1002/prot.26209.

123. Zhou, P., Tian, F., Lv, F. & Shang, Z. Geometric characteristics of hydrogen bonds involving sulfur atoms in proteins: Hydrogen Bonds Involving Sulfur Atoms in Proteins. *Proteins Struct. Funct. Bioinforma.* **76**, 151–163 (2009).

124. Nisthal, A., Wang, C. Y., Ary, M. L. & Mayo, S. L. Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. *Proc. Natl. Acad. Sci.* **116**, 16367–16377 (2019).

125. Nikam, R., Kulandaisamy, A., Harini, K., Sharma, D. & Gromiha, M. M. ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. *Nucleic Acids Res.* **49**, D420–D424 (2021).

126. Horovitz, A. & Fersht, A. R. Strategy for analysing the co-operativity of intramolecular interactions in peptides and proteins. *J. Mol. Biol.* **214**, 613–617 (1990).

127. Acharya, A., Rishi, V. & Vinson, C. Stability of 100 Homo and Heterotypic Coiled−Coil **a** − **a** ‘ Pairs for Ten Amino Acids (A, L, I, V, N, K, S, T, E, and R). *Biochemistry* **45**, 11324–11332 (2006).

128. MacCallum, J. L. *et al.* Assessment of protein structure refinement in CASP9: Assessment of Protein Structure Refinement in CASP9. *Proteins Struct. Funct. Bioinforma.* **79**, 74–90 (2011).

129. Nugent, T., Cozzetto, D. & Jones, D. T. Evaluation of predictions in the CASP10 model refinement category: Assessment of Model Refinement Predictions. *Proteins Struct. Funct. Bioinforma.* **82**, 98–111 (2014).

130. Modi, V. & Dunbrack, R. L. Assessment of refinement of template-based models in CASP11: Template-Based Models in CASP11. *Proteins Struct. Funct. Bioinforma.* **84**, 260–281 (2016).

131. Hovan, L. *et al.* Assessment of the model refinement category in CASP12. *Proteins Struct. Funct. Bioinforma.* **86**, 152–167 (2018).

132. Read, R. J., Sammito, M. D., Kryshtafovych, A. & Croll, T. I. Evaluation of model refinement in CASP13. *Proteins Struct. Funct. Bioinforma.* **87**, 1249–1262 (2019).

133. Simpkin, A. J., Sánchez Rodríguez, F., Mesdaghi, S., Kryshtafovych, A. & Rigden, D. J. Evaluation of model refinement in CASP14. *Proteins Struct. Funct. Bioinforma.* prot.26185 (2021) doi:10.1002/prot.26185.

134. Zemla, A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.* **31**, 3370–3374 (2003).

135. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins Struct. Funct. Bioinforma.* **57**, 702–710 (2004).

136. Holland, Jack & Grigoryan, Gevorg. Supplementary data for structure-conditioned amino-acid couplings. (2021) doi:10.5281/ZENODO.5643829.

137. Guzenko, D. & Strelkov, S. V. CCFold: rapid and accurate prediction of coiled-coil structures and application to modelling intermediate filaments. *Bioinformatics* **34**, 215–222 (2018).