Evolutionary couplings detect side-chain interactions

Adam J. Hockenberry[a,1] **and Claus O. Wilke**[a]

[a]Department of Integrative Biology, The University of Texas at Austin

This manuscript was compiled on October 18, 2018

Patterns of amino acid covariation in large protein sequence alignments can inform the prediction of *de novo* protein structures, binding interfaces, and mutational effects. While algorithms that detect these so-called evolutionary couplings between residues have proven useful for practical applications, less is known about how and why these methods perform so well, and what insights into biological processes can be gained from their application. Researchers frequently benchmark the performance of evolutionary coupling algorithms by comparing results with true structural contacts that are derived from solved protein structures. However, the method used to determine true structural contacts is not standardized and different definitions of structural contacts may have important consequences for comparing methods and understanding their overall utility. Here, we show that structural contacts between side-chain atoms are significantly more likely to be identified by evolutionary coupling analyses compared with backbone-based interactions. We use both simulation and empirical analyses to highlight that backbone-based definitions of true residue-residue contacts may underestimate the accuracy of evolutionary coupling algorithms by as much as 40%. These findings suggest that more advanced machine learning and neural network models developed to predict residue-residue contacts may be hindered by the use of mislabeled true positive training data.

Protein evolution | Structural constraints | Contact prediction

A long-standing problem in biology is to predict the structure of a protein based solely on its primary amino acid sequence (1–3). Despite advances in x-ray crystallography, NMR spectroscopy, and cryo-electron microscopy, the pace at which researchers are accumulating new genomes and gene sequences far out strips the ability of traditional biophysical methods to describe these genomes at the level of 3D-structure (4–7). A variety of computational methods–such as homology modeling (8, 9)–have been developed to support traditional biophysical methods, but *de novo* structural determination from primary sequence information alone remains elusive for all but the smallest proteins.

In recent years, however, computational researchers have made substantial improvements to *de novo* structural determination by leveraging co-evolutionary information contained within large sequence databases (10–13). Residues that co-evolve with one another across time may do so as a result of their spatial proximity with protein structures–i.e. mutations to an individual residue may be compensated for by subsequent mutations to other directly interacting residues(14–16). By determining an 'evolutionary coupling' score for all pairs of amino acid residues within a structure–and assuming that the highest-scoring residue-residue pairs are in close spatial proximity–researchers can constrain the search space of protein folding methods and accurately predict 3D-structures (10, 17, 18). Other applications have used evolutionary coupling scores to predict protein binding partners and interfaces (19, 20), as well as to predict the effect of mutations on protein stability and function (21). Many of these applications have been further been improved through the use of machine learning (22–24) and deep neural networks that leverage evolutionary couplings along-side numerous other protein features (25–36).

Despite the progress that has been made in this field–spurred by the development of so-called Direct Coupling Analyses and related methods–there are a number of known limitations to current methods for computing evolutionary couplings (37–40). Perhaps most importantly is a requirement for vast numbers of sequence homologs (18). Additionally, the evolutionary relatedness of sequences and the heterogeneity of substitution rates across sites may impose further constraints on the overall identifiability of evolutionary couplings. Finally, the more distantly related a given homolog is to the target structure, the more likely it is that there will be actual structural differences between molecules making the designation of a protein family based solely on sequence homology potentially problematic.

As researchers develop and refine algorithms to better predict evolutionary couplings from large multiple sequence alignments, a common work-flow is to benchmark methods against known protein structures to determine thee accuracy of residue-residue contact predictions (41, 42). The large number of protein structures that have been solved at atomic resolution provides a training data set where intra-molecular contacts are *known* (7). However, even the most high-resolution crystal structures of proteins require researchers to extrapolate

Significance Statement

Evolutionary couplings between residues within a protein can provide valuable information about protein structures, protein-protein interactions, and the mutability of individual residues. However, the mechanistic factors that determine whether two residues will co-evolve remains unknown. We show that structural proximity by itself is not sufficient for co-evolution to occur between residues. Rather, evolutionary couplings between residues are specifically governed by interactions between side-chain atoms. By contrast, intramolecular contacts between atoms in the protein backbone display only a weak signature of evolutionary coupling. These findings highlight that different types of stabilizing contacts exist within protein structures and that these types have a differential impact on the evolution of protein structures.

Please provide details of author contributions here.

Please declare any conflict of interest here.

[1]To whom correspondence should be addressed. E-mail: adam.hockenberry@utexas.edu

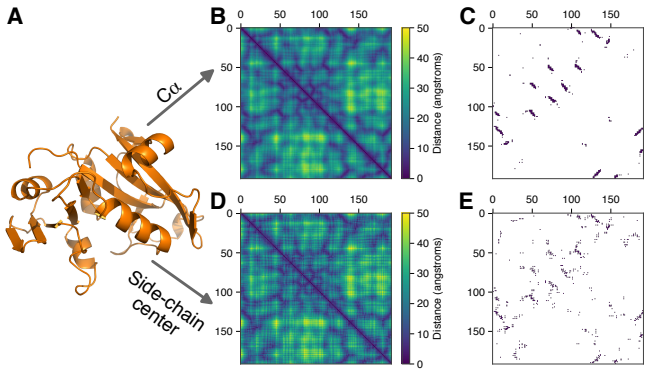

Fig. 1. Constructing contact maps from protein structures. (*A*) An example structure (PDB:1AOE). (*B*) A symmetrical distance matrix between all pairs of amino acid residues measured from each residues Cα atom. (*C*) Medium- to long-range contacts ($>$ 12 residues apart) are identified using an 8Å cutoff (dark blue). (*D*) and (*E*) Same methodology as depicted in (B) and (C), using the geometric center of each residues side-chain as a reference point for measuring distances.

from the location of particular atoms and residues to classify residue-residue 'bonds' or 'contacts' (43–46). A commonly used hueristic is to determine that any amino acid residue that lies within some pre-defined physical distance–frequently 8Å–of another amino acid residue is said to be in structural 'contact' (10).

Some current applications of evolutionary coupling analyses use Cα atoms as a reference point for determining residue-residue distances while others use Cβ or other complex methods such as the minimum distance between heavy atoms between two residues (18). Prior research has shown that the number of residue-residue contacts identified via side-chain centers is more closely related to evolutionary rates than similar metrics derived from Cα atoms (47–49). However, the consequences of choosing different reference points to determine the accuracy of modern evolutionary coupling approaches is unknown.

Currently, there are no accepted standards in the field for how to determine a network of residue-residue contacts for a given protein structure. Further, there has yet to be a systematic analysis of whether co-evolutionary signatures are more- or less-closely related to different types of intra-molecular contacts that may exist within a protein structure. Here, we systematically test the the accuracy of several evolutionary coupling algorithms against true positive contacts defined via Cα, Cβ, or side-chain geometric centers. We find that residue-residue contacts defined according to the distances between side-chain centers are much more accurately predicted by evolutionary couplings. These results imply that the dominant epistatic effects resulting in co-evolutionary signatures arise from side-chain::side-chain interactions. Our findings highlight the importance of the choice of contact-definition and provide insight into the constraints governing the evolution of protein structures.

Results

Structural contact definitions. Putatively true interactions between amino acid residues within a given protein are frequently derived from the distance between residues in known protein structures. Figure 1 depicts an example protein structure (PDB:1AOE) as well as a symmetric matrix depicting all residue-residue contact distances (in angstroms, Å) defined according to the distance between the Cα atoms of individual residues. By convention, we define true contacts as residue-residue pairs that are less than 8Å apart. We further note that, for most applications, the most structurally *interesting* contacts are mid- to long-range, which we define here as amino acid pairs separated by a primary chain distance of at least 12 residues (Fig. 1B,C). We only consider this subset of possible contacts for the remainder of this manuscript.

Many researchers have noted that the distance between amino acid residues need not be defined by Cα atom-based distances, and many applications rely on Cβ atoms (43–46). A logical question is whether using different reference points to define contacts matters in practice. To compare the consequences of choosing different reference points, we define all residue-residue contacts according to the 8Å, Cα atom-based distance threshold for a given protein. Next, we use the same absolute number of contacts to determine a comparable distance threshold (specific to each protein) to use for both Cβ atom and side-chain center based distances such that an equal number of putatively true contacts are identified regardless of the distance metric employed (SI Fig. S1). Although the distance matrices look similar for an example protein when calculated via Cα atoms or side-chain centers (Fig. 1B compared to D), the resulting maps of residue-residue contacts show considerable heterogeneity (Fig. 1C compared to E). More quantitatively, the set of all residue-residue distances measured by either Cα atoms, Cβ atoms, or side-chain centers are highly correlated with one another (Fig. 2A (left), SI Fig. S2). However, this strong *overall* correlation obscures important differences in *contact* definitions which we observe when focusing within the narrow region where direct amino acid residue contacts are defined (Fig. 2A (right), SI Fig. S2). For 1AOE,

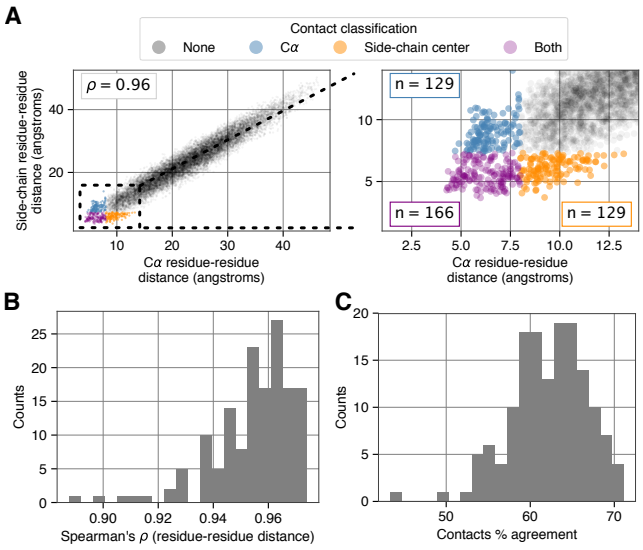

Fig. 2. Relationship between different contact identification methods. (*A*) Correlation between residue-residue distances in PDB:1AOE measured according to Cα atoms and side-chain centers (left). A zoomed in view (right) highlights variably defined residue-residue contacts indicated by the various colors. (*B*) Distribution of Spearman's correlation coefficient values (ρ) between residue-residue distances for 150 different proteins. (*C*) Distribution of the percent agreement for contact definitions for the same set of proteins. (SI Fig.S2 shows a comparable comparison between Cβ and side-chain center-based distances.)

we identified a total of 295 contacts according to the 8Å Cα atom-based distance threshold. Of the shortest 295 contact distances identified via side-chain centers–corresponding to a distance threshold of 7.33Å for this protein–the percent of residue-residue pairs that appear in both definitions is only 56% (78% for Cα compared to Cβ and 76% for Cβ compared to side-chain centers).

To assess the generality of these findings, we applied this analysis to a commonly used benchmark set of 150 proteins (23, 34, 38). Across all of these proteins, we observed a median correlation of 0.97 between residue-residue distances calculated via Cα atoms and side-chain centers (Fig. 2B) and a median overlap of 63% between contacts defined via Cα and side-chain centers (Fig. 2C). Cα- and Cβ-based definitions, as well as Cβ- and side-chain center-based definitions, both had median overlaps of 78% (SI Fig. S2). Together, these results highlight that true contacts vary substantially according to the reference point used to measure residue-residue distances.

Simulation analyses. While the previous analysis of empirical structures shows that the choice of reference point has important consequences for true contact identification, it is not clear which of the different methods is more biologically "correct" or practically meaningful. We thus turned our attention to a simplified biophysical system to test the ability of evolutionary coupling analyses to recover intramolecular contacts. We used the ROSETTA modeling software (50–52) to perform all-atom evolutionary simulations of the evolutionary process (53, 54) while selecting for the maintenance of protein stability (expressed as a fraction of the initial PDB model stability). We simulated thousands of independent evolutionary trajectories, and used the resulting amino acid sequences from these simulations to calculate evolutionary couplings. We used 3 separate algorithms to calculate evolutionary couplings, but the main text results depict predictions using CCMpred. Within this defined system, we are able to remove the constraints of phylogenetic biases, limited data availability, homopolymerization, and changes in evolutionary pressures over time between species–all of which partially limit the power of algorithms to detect true evolutionary couplings in real data (55).

We continued to use 1AOE as an example protein and varied several parameters of our simulation to ensure robust results. We defined true positive residue-residue contacts according to the original PDB structure using residue-residue distances calculated between different quantities for comparison (Cα, Cβ, and side-chain center). To assess the accuracy of evolutionary couplings, we determined the positive predictive value (PPV) of the top $L/2$ couplings—where L is the primary chain length of the protein under investigation. For Cα-based contact definitions, we found that the PPV increases rapidly according to the number of independent sequences that we simulated and consequently used as input for evolutionary coupling analyses (Fig. 3A). In each case, we ran these simulations until we accepted mutations totaling 10x the length of the protein sequence. However, regardless of the selection strength that we imposed on the sequence evolution (colored lines in Fig. 3A), PPV values plateaued at a value $<< 1$ indicating that evolutionary couplings were failing to accurately capture protein contacts. By contrast, when we analyzed the same evolutionary coupling values but used side-chain center-based distances to define contacts we observed that PPV values approached 1 (representing perfect prediction accuracy for this subset of couplings) in almost all cases.

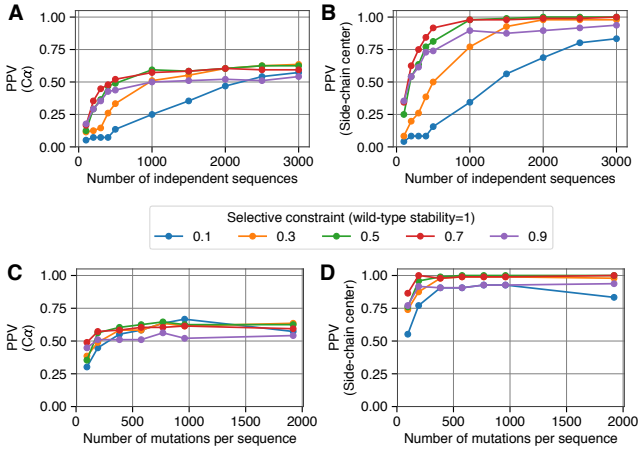

Fig. 3. Comparing simulation-derived evolutionary couplings to different contact definitions. (*A*) For each of 5 separate selection strengths (colored lines), we ran simulations until a number of mutations totaling 10 times the length of the protein were accumulated per replicate. We varied the number of independent replicate sequences (x-axis) that were used as input for evolutionary coupling analysis, and found that couplings fail to fully recover Cα defined contacts for PDB:1AOE. (*B*) By contrast, contacts defined via side-chain centers are near-perfectly recovered across a range of simulation parameters. (*C*) and (*D*) Similar to parts (A) and (B), but along the x-axis we now show results from simulations where a different number of accepted mutations were accumulated per sequence. We fixed the number of replicate sequences that were simulated–and used for evolutionary coupling analysis–at 3,000 for each of these data points. (Results comparing Cβ and side-chain center-based contact definitions, can be found in SI Fig. S3.)

We additionally explored how the number of mutations accumulated per sequence affected the ability of evolutionary coupling algorithms to recover intramolecular contacts. We fixed the number of replicate sequences at 3000, and observed that PPV values showed minimal variation according to the number of accepted mutations per sequence (Fig. 3C). As before, however, prediction accuracies were substantially higher when we defined true contacts according to side-chain center distances (Fig. 3D).

These simulation results highlight that–across numerous parameter combinations–the top $L/2$ evolutionary couplings were all true positive intramolecular contacts as long as true positives were defined according to side-chain centers and not Cα carbons. Additionally, we depict Cα and side-chain center based methods here because they represent extreme ends of the spectrum from backbone to side-chains. Cβ-based contact definitions had intermediate accuracy, plateauing at higher values than Cα but lower than side-chain center definitions (SI Fig. S3).

Empirical analyses. To see how evolutionary couplings compare to different definitions of true residue-residue contacts in empirical data, we used PHMMER to identify sequence homologs for each of the 150 proteins (see Materials and Methods for details). We assessed the relationship between evolutionary couplings and structural contacts for all proteins by calculating the positive predictive value (PPV) of the highest $L/2$ couplings.

As expected, the PPV between the top $L/2$ evolutionary couplings and Cα-based contacts varied substantially across the 150 structures. This variation may result from a number of different effects, and we observed a clear and expected

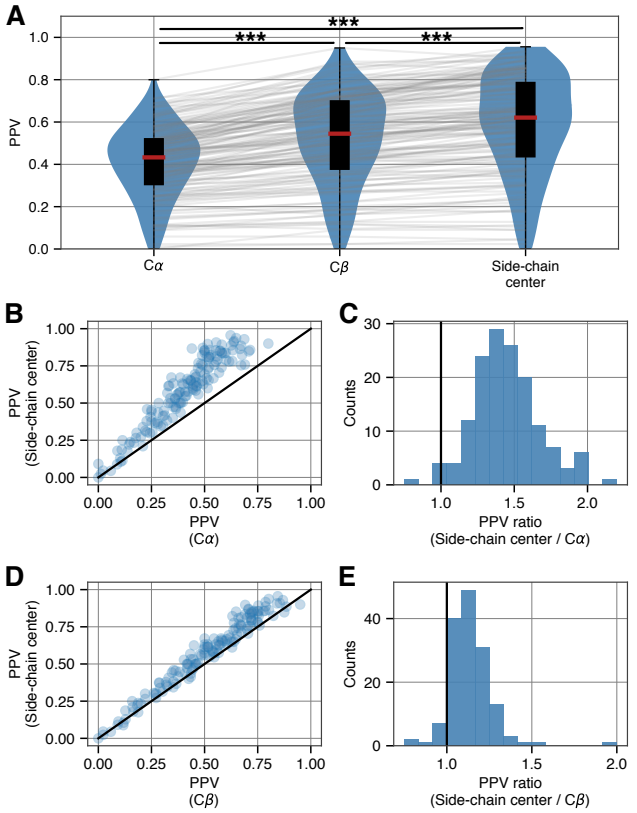

Fig. 4. Accuracy of evolutionary couplings derived from empirical alignments. (*A*) For a diverse set of 150 proteins, the PPV of the top $L/2$ evolutionary coupling scores–derived from empirical sequence alignments–is progressively higher when intramolecular contacts are defined according to Cα atoms, Cβ atoms, and side-chain centers. (*** indicates $p < 10^{-20}$, Wilcoxon signed-rank test) (*B*) Scatter plot of PPVs for each protein according to Cα and side-chain center-based contact identification methods. (*C*) Histogram of the ratios from the data in (B) indicate that the median percent increase in accuracy is 43%. (*D*) and (*E*) As in (B) and (C), comparing Cβ and side-chain center-based contact identification methods. Results show a median percent increase in contact identification accuracy of 13%. (Results for other evolutionary coupling algorithm implementations can be found in SI Figs. S5 & S6.)

correlation between PPV values and the number of available homologous sequences used to determine evolutionary couplings (SI Fig. S4). Despite the variability in prediction accuracy between proteins, we observed systematic variability in the PPV according to which metric was used to identify true positive contacts (Fig. 4A). When compared with Cα-based distances, residue-residue distances measured according to Cβ atoms resulted in significantly higher PPVs, and side-chain-based contact distances resulted in even further improvements. Further, the magnitude was substantial: across all 150 proteins the median percent increase in PPV between Cα- and side-chain center-based contact identification methods was 43% (Fig. 4B,C). Even between the more similar Cβ- and side-chain center-based methods, the median percent increase in accuracy was 13% (Fig. 4D,E). Both comparisons were highly significant and persisted across the entire range of PPVs represented within our dataset (Fig. 4B,D). Additionally, these results were highly consistent across different evolutionary coupling algorithms (SI Figs. S5 & S6).

Side-chain orientation and evolutionary couplings. Using the exact same evolutionary couplings, the previous analyses have shown that PPVs are substantially higher when using side-chain-based distances to identify true positive intramolecular contacts compared with either Cα or Cβ-based distances. To look more specifically at why these differences were so pronounced, we decided to investigate the orientation of residue-residue pairs identified by the various criteria. At a simple level, any two residues can be in structural contact across a number of orientations of their respective side-chains: i) both residue's side-chains may point towards one another with the energetic interactions occurring through side-chain atoms, ii) one residue's side-chain may point towards the other residue while that residue's side-chain points away, or iii) both residue's side-chains may point away from one another with energetic interactions occurring between the respective amino acid backbones (Fig. 5A). As expected, when we look only at residue-residue pairs that are defined as contacts via different reference points, we see that side-chain based contact definitions strongly enrich for cases where both side-chains point towards one another in an example protein (Fig. 5B).

Across all 150 proteins in our dataset, we calculated the fraction of *all* residue-residue pairs (regardless of whether they are putative contacts, but subject to the same primary chain distance constraints applied throughout this manuscript) where both side-chains point towards one another and found it to be relatively small (Fig. 5C, "All pairs"). However, this fraction increases progressively when we limit our analysis to the subset of residue-residue pairs identified as true contacts for each protein according to Cα, Cβ and side-chain centers–illustrating that the trend observed in (Fig. 5B) applies broadly. If instead we only look at the top-ranked evolutionary couplings (ignoring whether or not the residue-residue pairs are putatively true structural contacts), we observe that a large fraction of the strongest identified evolutionary couplings are between residues that point towards one another in the reference protein structure. Additionally, this fraction is highest for the most highly ranked evolutionary couplings and is substantially higher than the proportion identified by Cα-based distances.

To further illustrate this point, we turned to an alternative method for determining intramolecular contacts that we have not yet systematically explored: determining structural contacts based on the minimum distance between any two heavy atoms for each residue-residue pair. We implemented two versions of this algorithm, determining the minimum distance between: i) all heavy atoms within residues and ii) side-chain heavy atoms only. In each case, and to facilitate comparison between methods, we again selected the shortest X distances as contacts where X is the number of contacts identified for each protein via the 8Å distance threshold using Cα. For the set of 150 proteins, the resulting PPVs were significantly higher when contacts were defined *only* according to side-chain atoms as opposed to the complete set of backbone and side-chain atoms (SI Fig. S7). Furthermore, PPVs calculated via side-chain center distances were statistically indistinguishable from PPVs derived from the minimum distance between all heavy atoms within side-chains.

Taken together, our analysis of side-chain orientations and our analysis of contacts identified via minimum atomic distances both highlight that evolutionary couplings frequently occur between residues whose side-chains point towards one another. Cα-(and to a lesser extent Cβ-) based contact defini-

tions classify a smaller number of contacts in this orientation, and including backbone atoms in minimum-distance based contact identification methods actually decreases the accuracy of contact predictions based on evolutionary couplings.

Discussion

The co-evolutionary patterns of amino acid substitutions can provide important information about protein structures. There are a number of competing methods currently employed by different researchers to detect evolutionary couplings between residues, and the ability to recover true residue-residue contacts has been the primary metric used to assess performance of various methods. However, true structural contacts are ill-defined and variability in contact definitions can prohibit comparison between the efficacy of different methods, as well as obscure the biological interpretation of evolutionary constraints. We show here that evolutionary couplings are significantly more accurate at detecting true residue-residue contacts based off of side-chain center distances. Critically, these findings provide important biological insight protein evolution and epistatic interactions between residues. Our model posits that although different types of interactions between amino acid residues may stabilize protein structures, evolutionary couplings predominantly consist of residues whose contact occurs via interactions between the side-chain atoms of both residues.

Evolutionary couplings are themselves important, but more recently these values have been used as input into more advanced machine learning and neural network-based algorithms that supplement evolutionary couplings with a variety of other information to predict intramolecular contacts. However, our analysis suggests that there may be biases in training data–essential to supervised learning techniques–based on how intramolecular contacts are defined; the definition of putatively "true" structural contacts relies on the method used to calculate residue-residue distances. We show that evolutionary couplings more accurately predict side-chain center-based contacts, and the strongest evolutionary couplings are consistently enriched for residue-residue pairs where the side-chains are oriented towards one-another. We speculate that accuracy of supervised algorithms may be improved with more properly labeled training data that corresponds with these known biophysical constraints. Alternatively, supervised learning algorithms may be able to achieve even greater improvements in accuracy by separating different *types* of residue-residue contacts according to their atomic interactions, training separate models to detect each type, and integrating the results.

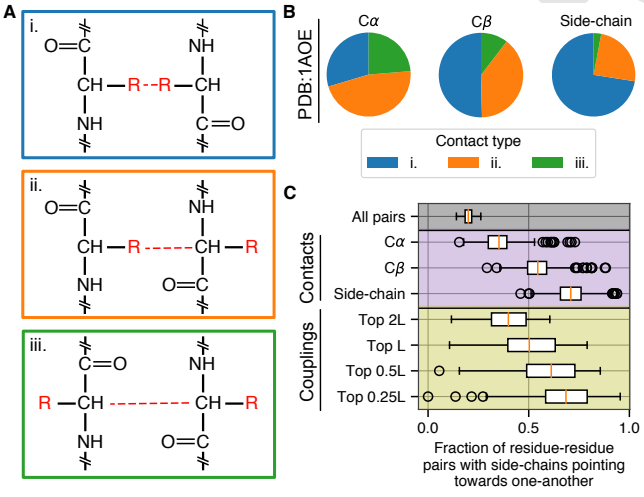

Fig. 5. Different types of residue-residue interactions are possible. (*A*) Two interacting residues may interact via: each residue's side chain atoms (type i), the side-chain of one residue and the backbone of the other (type ii), or the backbone atoms of each residue (type iii). (*B*) For intramolecular contacts identified in PDB:1AOE, the relative proportion of different interaction types varies according to contact identification method. Residue-residue contacts defined via side-chain centers are enriched in type i interactions (blue). (*C*) For 150 proteins, the fraction of residue-residue pairs where the side-chains point towards one-another is highest in contacts defined via side-chain centers (purple). Further, the top ranked evolutionary couplings (regardless of whether they are defined as contacts) are progressively enriched in residue-residue pairs where the side-chains point towards one another (yellow).

A number of issues constrain the maximal accuracy that can be expected from using evolutionary couplings alone to predict contacts. Anischenko *et al.* (2017) illustrated that many so-called false positive signals resulting from evolutionary coupling analyses arise from repeat proteins, homo-oligomerization, and structural variation within protein families (55). Here, we show that another source of false positive signals may simply be ill-defined true positive contacts. Without changing anything about the way evolutionary couplings are calculated, we show their accuracy at predicting intramolecular contacts is progressively higher for different contact-definitions. Further, the magnitude of this difference is not trivial: across a diverse set of proteins we show that side-chain center based contacts are predicted with a median of 43% and 13% higher accuracy than comparable Cα and Cβ based contacts. Thus a substantial number of *false positive* predictions made by evolutionary coupling analyses may simply due to the false classification of true positives.

While improving contact identification methods is an important practical result, our findings improve our understanding of protein evolution by showing that side-chain interactions are more important for governing epistasis between amino acid residues within individual protein structures. Although the overall structural geometry of a protein is dictated by the shape of the protein backbone, consideration of side-chains is critical for maintaining this geometry and determining the co-evolutionary dynamics of substitutions. Our findings do not suggest that intra-molecular contacts between the backbone atoms of residues are not important for folding or stabilizing protein structures. Rather, our results suggest that contacts between backbone atoms are not likely to be detected by evolutionary coupling analyses and imply that epistatis between residues is largely governed by whether side-chain atoms are in direct contact.

Direct coupling analyses and related methods have significantly improved our ability to leverage the experiment of natural sequence evolution for the purpose of predicting important properties of proteins. While these methods continue to find novel applications, they are beginning to provide mechanistic insight into the evolutionary process (56). Ultimately, it may even be possible to incorporate more realistic pair-wise interactions into models of sequence evolution and inference, which are almost exclusively site-independent. Further technical improvements, such as explicitly accounting for the phylogenetic relatedness of sequences, may allow for even more accurate inference of evolutionary couplings and consequently insight into biological mechanisms.

Materials and Methods

Dataset compilation and processing. We downloaded the so-called PSICOV dataset of 150 proteins that have been extensively studied (23, 34, 38). We processed each starting ".PDB" file to select a single chain, ensure a consistent numbering of residues (1...n), test for unknown or non-standard residues, select the most likely state for all disordered sequence atoms, and remove all extraneous information (including "HETATM" lines). Next, to ensure that all residues were represented in full and repair those that were not, we used PYROSETTA to read in the ".PDB" files using the 'pose_from_pdb' function and wrote the output as our final clean structure.

Determining structural contacts and contact-types. From each cleaned ".PDB" file, we calculated residue-residue distance matrices using custom python scripts (the euclidean distance from 3-dimensional atomic coordinates). All residues contain a Cα atom so this calculation was straightforward. For Cβ calculations, we used the Cβ atom of all residues except glycine, where we continued to use the Cα atom. For side-chain center calculations, we calculated the geometric center of each residue based on the coordinates of all non-backbone heavy atoms. This calculation included Cβ atoms but excluded Cα atoms for all amino acids except glycine, where we continued to use Cα as the reference point.

To calculate minimum atomic distances between two residues, we calculated all pairwise euclidean distances between heavy atoms and selected the minimum distance. In extending this analysis to only consider side-chain atoms, we continued to consider Cβ atoms as part of the side-chain but not Cα. Again, we relaxed this restriction for glycine and included Cα as a side-chain atom to permit calculations.

For all methods, contacts were assessed by first removing all residue-residue pairs where the two residues were shorter than 12 amino acids apart in primary chain distance. Contacts were determined throughout this manuscript for each structure according to an 8Å cutoff between Cα atoms. Since accuracy values are partially dependent on the number of true positives that are called, we maintained a constant number of true positive contact classifications throughout to facilitate comparison between methods. For each contact definition (Cβ, side-chain center, minimum atomic distances), we selected n residue-residue pairs with the shortest distances where n is the number of contacts defined according to the aforementioned Cα-based method.

To classify residue-residue pairs (a and b) via their side-chain orientations, we chose a residue (a) and drew two vectors: i) from the Cα atom coordinate to the side-chain center for that residue and ii) from the Cα atom coordinate to the Cα atom coordinate for the other residue in question (b). If the angle between these two vectors was less than $\pi/2$ radians, the side-chain of residue a was said to point towards residue b. To determine the residues classification as in Fig.5A, we next repeated the calculation using residue b as the reference and classified the residue-residue pair accordingly.

Evolutionary coupling algorithms. For each of the 150 proteins in our dataset, we followed a principled method to retrieve homologous sequences. We first extracted the primary amino acid sequence from the ".PDB" file. We next used PHMMER to search through progressively larger databases in order to retrieve up to 10,000 homologous sequences. To do so, we downloaded local versions of the rp15, rp35, rp55, and rp75 databases. We first searched the smallest, least redundant, database for each sequence using an E-value threshold of 0.0001. For any sequence with greater than 10,000 hits we stopped and selected the top scoring 10,000 hits for further analysis. For sequences with fewer than 10,000 hits we moved to the next largest database and repeated the process. Finally for the small number of sequences for which we did not accumulate at least 1,000 sequences in the largest database (rp75), we used the online version of PHMMER to search the UniprotKB database and downloaded the maximum results.

For each protein, we next aligned the hits along-side the reference sequence using MAFFT (L-INS-i method with default parameters). Next, we cleaned these results to remove all columns that were gapped in the reference (".PDB") sequence. All other columns and sequences in the sequence alignments were retained regardless of gap coverage.

Using these alignments, we next calculated evolutionary couplings between residue-residue pairs. All results in the main manuscript are displayed using CCMpred with default parameters (0.8 local sequence re-weighting threshold, 0.2 pairwise regularization coefficients, average product correction). We additionally used the 'plmc' method from the EVcouplings framework with default parameters (no average product correction) and PSICOV (default parameters excepting: "-z 50 -r 0.001") to ensure the robustness of our findings.

Except where otherwise noted (Fig. 5), main text results (Fig. 3, Fig. 4) were calculated using the top $L/2$ couplings for each protein where L is the length of the reference amino acid sequence. Positive Predictive Value (PPV) is calculated as the number of classified contacts among these top couplings divided by the total number of top couplings considered.

Evolutionary simulations. For the example protein used throughout the text (PDB:1AOE) we performed mutation accumulation simulations using PYROSETTA. We first read in the ".PDB" structure (with di-sulfide bonds turned off), and minimized it so as to optimize thermodynamic stability by rotamer selection and backbone movements. We next fixed the backbone, and implemented an expedited evolutionary algorithm to select amino acid point mutations (no insertions or deletions were allowed) according to their predicted impact on structural stability. At each step, we selected a random amino acid position, and attempted to mutate it randomly to one of the remaining 19 amino acids. We re-packed the structure within a 12Å radius of the mutation and determined whether or not to accept it based off of the resulting change in structural stability. Mutations which either did not alter or which increased stability (i.e. resulted in a decreased ΔG) were accepted. Mutations that decreased stability were accepted with a probability proportional to their $\Delta\Delta G$ as in Teufel and Wilke (2017). At the end of the evolutionary process, the resulting amino acid sequence was stored for future analysis.

We performed thousands of independent replicates of this expedited evolutionary process where we altered the number of accepted mutations that we accumulated, the number of replicate evolutionary experiments that we performed, and the fraction of the initial wild-type stability value that we used for our selection criteria. Collections of the resulting sequences were analyzed via evolutionary coupling algorithms in the same manner as empirical sequences, with no need for sequence alignment.

Data access. All code and data are currently being compiled and edited. They will be made freely available at the following link: https://github.com/adamhockenberry/side-chain-couplings which will itself evolve throughout the submission process to reflect final manuscript analyses.

ACKNOWLEDGMENTS. The authors acknowledge valuable feedback and support from members of the Wilke lab.

1. Anfinsen CB (1973) Principles that Govern the Folding of Protein Chains. *Science* 181(4096):223–230.
2. Sadowski MI, Jones DT (2009) The sequence-structure relationship and protein function prediction. *Current Opinion in Structural Biology* 19(3):357–362.
3. Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. *Nature Biotechnology* 30(11):1072–1080.
4. Liao M, Cao E, Julius D, Cheng Y (2013) Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature* 504(7478):107–112.
5. Amunts A, et al. (2014) Structure of the Yeast Mitochondrial Large Ribosomal Subunit. *Science* 343:1485–1489.
6. Punjani A, Rubinstein JL, Fleet DJ, Brubaker MA (2017) CryoSPARC: Algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods* 14(3):290–296.
7. Rose PW, et al. (2017) The RCSB protein data bank: Integrative view of protein, gene and 3D structural information. *Nucleic Acids Research* 45:D271–D281.
8. Söding J (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21(7):951–960.
9. Biasini M, et al. (2014) SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research* 42:252–258.
10. Marks DS, et al. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* 6(12).
11. Morcos F, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences* 108(49):E1293–E1301.

12. Sulkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN (2012) Genomics-aided structure prediction. *Proceedings of the National Academy of Sciences* 109(26):10340–10345.
13. Kamisetty H, Ovchinnikov S, Baker D (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences* 110(39):15674–15679.
14. Gobel U, Sander C, Schneider R, Valencia A (1994) Correlated Mutations and Residue Contacts in Proteins. *Proteins* 18:309–317.
15. Shindyalov IN, Kolchanov NA, Sander C (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Engineering, Design and Selection* 7(3):349–358.
16. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences* 106(1):67–72.
17. Hopf TA, et al. (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 149(7):1607–1621.
18. Ovchinnikov S, et al. (2017) Protein structure determination using metagenome sequence data. *Science* 355(6322):294–298.
19. Hopf TA, et al. (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* 3:1–45.
20. Ovchinnikov S, Kamisetty H, Baker D (2014) Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife* pp. 1–21.
21. Hopf TA, et al. (2017) Mutation effects predicted from sequence co-variation. *Nature Biotechnology* 35(2):128–135.
22. Cheng J, Baldi P (2007) Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics* 8(1).
23. Jones DT, Singh T, Kosciolek T, Tetchner S (2015) MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* 31(7):999–1006.
24. Michel M, Skwark MJ, Hurtado DM, Ekeberg M, Elofsson A (2017) Predicting accurate contacts in thousands of Pfam domain families using PconsC3. *Bioinformatics* 33(18):2859–2866.
25. Tegge AN, Wang Z, Eickholt J, Cheng J (2009) NNcon: Improved protein contact map prediction using 2D-recursive neural networks. *Nucleic Acids Research* 37:515–518.
26. Lena PD, Nagata K, Baldi P (2012) Deep architectures for protein contact map prediction. *Bioinformatics* 28(19):2449–2457.
27. Xiong D, Zeng J, Gong H (2017) A deep learning framework for improving long-range residue-residue contact prediction using a hierarchical strategy. *Bioinformatics* 33(17):2675–2683.
28. Stahl K, Schneider M, Brock O (2017) EPSILON-CP: Using deep learning to combine information from multiple sources for protein contact prediction. *BMC Bioinformatics* 18(1):1–11.
29. He B, Mortuza SM, Wang Y, Shen HB, Zhang Y (2017) NeBcon: Protein contact map prediction using neural network training coupled with naïve Bayes classifiers. *Bioinformatics* 33(15):2296–2306.
30. Wang S, Sun S, Li Z, Zhang R, Xu J (2017) *Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model*. Vol. 13.
31. Riesselman AJ, Ingraham JB, Marks DS (2018) Deep generative models of genetic variation capture mutation effects. *Nature Methods*.
32. Liu Y, Palmedo P, Ye Q, Berger B, Peng J (2018) Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks. *Cell Systems* 6(1):65–74.e3.
33. Wozniak PP, Pelc J, Skrzypecki M, Vriend G, Kotulska M (2018) Bio-knowledge based filters improve residue- residue contact prediction accuracy. *Bioinformatics* (May):1–9.
34. Jones DT, Kandathil SM (2018) High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics* (April):1–8.
35. Adhikari B, Hou J, Cheng J (2018) DNCON2: Improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics* 34(9):1466–1472.
36. Hanson J, Paliwal K, Litfin T, Yang Y, Zhou Y (2018) Accurate Prediction of Protein Contact Maps by Coupling Residual Two-Dimensional Bidirectional Long Short-Term Memory with Convolutional Neural Networks. *Bioinformatics* (June):bty481–bty481.
37. Burger L, Van Nimwegen E (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Computational Biology* 6(1).
38. Jones DT, Buchan DW, Cozzetto D, Pontil M (2012) PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28(2):184–190.
39. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E (2013) Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E* 87(1):1–16.
40. Seemayer S, Gruber M, Söding J (2014) CCMpred - Fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics* 30(21):3128–3130.
41. Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, Bonvin AM (2018) Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. *Proteins: Structure, Function and Bioinformatics* 86(October 2017):51–66.
42. Wang S, Sun S, Xu J (2018) Analysis of deep learning methods for blind protein contact prediction in CASP12. *Proteins: Structure, Function and Bioinformatics* 86(August 2017):67–77.
43. Seeliger D, de Groot BL (2007) Atomic contacts in protein structures. A detailed analysis of atomic radii, packing, and overlaps. *Proteins* 68:595–601.
44. Sathyapriya R, Duarte JM, Stehr H, Filippis I, Lappe M (2009) Defining an essence of structure determining residue contacts in proteins. *PLoS Computational Biology* 5(12).
45. Duarte JM, Sathyapriya R, Stehr H, Filippis I, Lappe M (2010) Optimal contact definition for reconstruction of Contact Maps. *BMC Bioinformatics* 11.
46. Yuan C, Chen H, Kihara D (2012) Effective inter-residue contact definitions for accurate protein fold recognition. *BMC Bioinformatics* 13(1).
47. Lin CP, et al. (2008) Deriving protein dynamical properties from weighted protein contact number. *Proteins* 72(3):929–935.
48. Marcos ML, Echave J (2015) Too packed to change: side-chain packing and site-specific substitution rates in protein evolution. *PeerJ* 3:e911.
49. Shahmoradi A, Wilke CO (2016) Dissecting the roles of local packing density and longer-range effects in protein sequence evolution. *Proteins* 84(6):841–854.
50. Chaudhury S, Lyskov S, Gray JJ (2010) PyRosetta: A script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* 26(5):689–691.
51. Leaver-Fay A, et al. (2011) Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology* 487(Chapter 19):545–574.
52. Kellogg EH, Leaver-Fay A, Baker D (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* 79(3):830–838.
53. Teufel AI, Wilke CO (2017) Accelerated simulation of evolutionary trajectories in origin-fixation models. *Journal of the Royal Society Interface* 14(127).
54. Jiang Q, Teufel AI, Jackson EL, Wilke CO (2018) Beyond Thermodynamic Constraints: Evolutionary Sampling Generates Realistic Protein Sequence Variation. *Genetics* 208:1387–1395.
55. Anishchenko I, Ovchinnikov S, Kamisetty H, Baker D (2017) Origins of coevolution between residues distant in protein 3D structures. *Proceedings of the National Academy of Sciences* p. 201702664.
56. Figliuzzi M, Barrat-Charlaix P, Weigt M (2018) How pairwise coevolutionary models capture the collective residue variability in proteins? *Molecular Biology and Evolution* 35(4):1018–1027.