





CovET: A covariation-evolutionary trace method that identifies protein structure–function modules

Received for publication, January 25, 2023, and in revised form, June 1, 2023. Published, Papers in Press, June 7, 2023.
<https://doi.org/10.1016/j.jbc.2023.104896>

Daniel M. Konecki^{1,‡}, Spencer Hamrick^{2,‡}, Chen Wang (王忱)^{3,‡} , Melina A. Agosto⁴, Theodore G. Wensel^{1,3,4,5} , and Olivier Lichtarge^{1,3,4,5,6,*}

From the ¹Quantitative and Computational Biosciences Graduate Program, ²Chemical, Physical, and Structural Biology Graduate Program, ³Department of Molecular and Human Genetics, ⁴Verna and Marrs McLean Department of Biochemistry and Molecular Biology, ⁵Cancer and Cell Biology Graduate Program, and ⁶Computational and Integrative Biomedical Research Center, Baylor College of Medicine, Houston, Texas, USA

Reviewed by members of the JBC Editorial Board. Edited by Henrik Dohlman

Measuring the relative effect that any two sequence positions have on each other may improve protein design or help better interpret coding variants. Current approaches use statistics and machine learning but rarely consider phylogenetic divergences which, as shown by Evolutionary Trace studies, provide insight into the functional impact of sequence perturbations. Here, we reframe covariation analyses in the Evolutionary Trace framework to measure the relative tolerance to perturbation of each residue pair during evolution. This approach (CovET) systematically accounts for phylogenetic divergences: at each divergence event, we penalize covariation patterns that belie evolutionary coupling. We find that while CovET approximates the performance of existing methods to predict individual structural contacts, it performs significantly better at finding structural clusters of coupled residues and ligand binding sites. For example, CovET found more functionally critical residues when we examined the RNA recognition motif and WW domains. It correlates better with large-scale epistasis screen data. In the dopamine D2 receptor, top CovET residue pairs recovered accurately the allosteric activation pathway characterized for Class A G protein-coupled receptors. These data suggest that CovET ranks highest the sequence position pairs that play critical functional roles through epistatic and allosteric interactions in evolutionarily relevant structure–function motifs. CovET complements current methods and may shed light on fundamental molecular mechanisms of protein structure and function.

Covariation analysis probes the multiple sequence alignment of a protein family to search for sequence positions whose mutational pattern of variations are not independent. Since most compensatory couplings are thought to arise from structurally neighboring residues, covariant residue positions

have been sought to predict structural contacts and thus constrain the search space of possible protein folds toward *de novo* computational structure prediction (1–6). AlphaFold (7, 8) and RoseTTAFold (9) both use covariation among many other training features in the context of machine learning.

The assumptions of structural proximity and compensatory mutations in covariation analysis bear closer scrutiny, however. First, concerted motions within a protein may lead to allosteric interactions. As a result, compensatory mutations may occur between structurally distant protein residues (10–12). Second, coupled residues must not necessarily undergo compensatory mutations during evolution. In fact, phylogenetic divergence involves changes in functional aptitude, with larger changes between increasingly distant species. Accordingly, mutational perturbations are increasingly less likely to be compensated between evolutionary distant orthologs—since mutation is precisely the instrument of evolution. These allosteric and evolutionary caveats are likely to impact the analysis of large protein sequence alignments but are not accounted for by current covariation algorithms in their search for coupled pairs of residues in structural contact (13, 14).

Therefore, we used the Evolutionary Trace (ET) framework to better account for phylogenetics in covariation analysis of structural contacts, allosteric interactions, and epistasis following a prior but different effort (15, 16). ET ranks every sequence position by its relative tolerance or sensitivity to mutation by correlating sequence variations with phylogenetic divergences between species. This approach has helped predict protein binding sites (17, 18), design separation of function mutations (19–22), annotate and alter protein function (23–36), design inhibitors (37, 38), and identify residues involved in diseases (39, 40). ET algorithms iteratively divide a multiple sequence alignment into progressively smaller sequence groupings following the branching pattern of the phylogenetic tree, penalizing each grouping by the total amount of entropy in each group. Accordingly, positions with patterns of variations that track closely with the evolutionary tree are deemed to be most functionally significant and penalized least. Positions that vary irrespective of the evolutionary tree and are deemed least important and penalized severely (17, 41). Here,

[‡] The first three authors are joint First Authors.

* For correspondence: Olivier Lichtarge, lichtarge@bcm.edu.

Present address for Melina A. Agosto: Retina and Optic Nerve Research Laboratory, Department of Physiology and Biophysics, and Department of Ophthalmology and Visual Sciences, Dalhousie University, Halifax, Nova Scotia, Canada.

CovET identifies protein structure–function modules

we reasoned that the same ET penalty scheme could be extended to pairs of residue positions, as to measure how well their variations are consistent with phylogenetic divergences. And thus developed our new covariation algorithm, CovET, which ranks the relative tolerance of residue pairs to mutations using the ET framework.

We tested CovET on both a large scale and in specific structures for its ability to predict structural features, functional sites, allosteric interactions, and experimental data on epistasis. We find that CovET not only identifies covarying residues that correspond to individual structural contacts of interest but outperforms other methods in the ability to identify mutually clustered pairs of covarying residues that reveal functional sites and pathways linked to allostery and to epistasis. The code for CovET is open and freely available on GitHub at: <https://github.com/LichtargeLab/Covariation-ET>.

Results

CovET method development

CovET uses the ET framework to identify phylogenetically supported residue couplings. ET has the general form: $ET_i = \sum_{n=1}^N \frac{1}{n} \sum_{g=1}^n x_i$, where i is a position or pair of positions, N is the depth of the phylogenetic tree, n is a specific level in the tree, g is a group of evolutionarily related sequences (represented by a branch in the tree), and x_i is a scoring function for residues at position i in group g in the phylogenetic tree. This general algorithm captures the process of traversing the

phylogenetic tree from root to leaves ($n = 1$ to N , represented by vertical-colored lines in Figure 1B) and of characterizing each group (g , represented by colored boxes in Fig. 1B) according to a penalty function x_i . CovET penalizes non-concerted variations ($AB > AC$ or $AB > CB$) between pairs of residues according to the exponential of the Shannon entropy, *i.e.*, the diversity metric or perplexity, resulting in the final formula in Equation 2 and Figure 1. v in the entropy calculation is any of the nonconcerted variations which can be observed when comparing a pair of positions between one sequence and another, a process highlighted in Figure 1C. CovET ignores concerted variation ($AB > CD$) and conservation ($AB > AB$) as neither is inconsistent with a coupling. The score resulting from a simple characterized pair of positions ij is demonstrated in Figure 1C. We hypothesize that this phylogenetic formulation will allow us to identify functionally relevant residue couplings.

CovET preferentially identifies higher order protein structures over local contacts

To test whether CovET identifies local structural contacts (residues where C β are within 8 Å of each other, C α for glycine), we studied its rankings of residue pairs in 943 protein families. ET and other alignment-based methods are inherently sensitive to the alignment itself, so we used the Pfam dataset to obtain a large, diverse protein set (42) (Table S1, see Experimental Procedures for detail). We compared CovET against standard methods: EVCouplings (13, 43) and DCA (44)

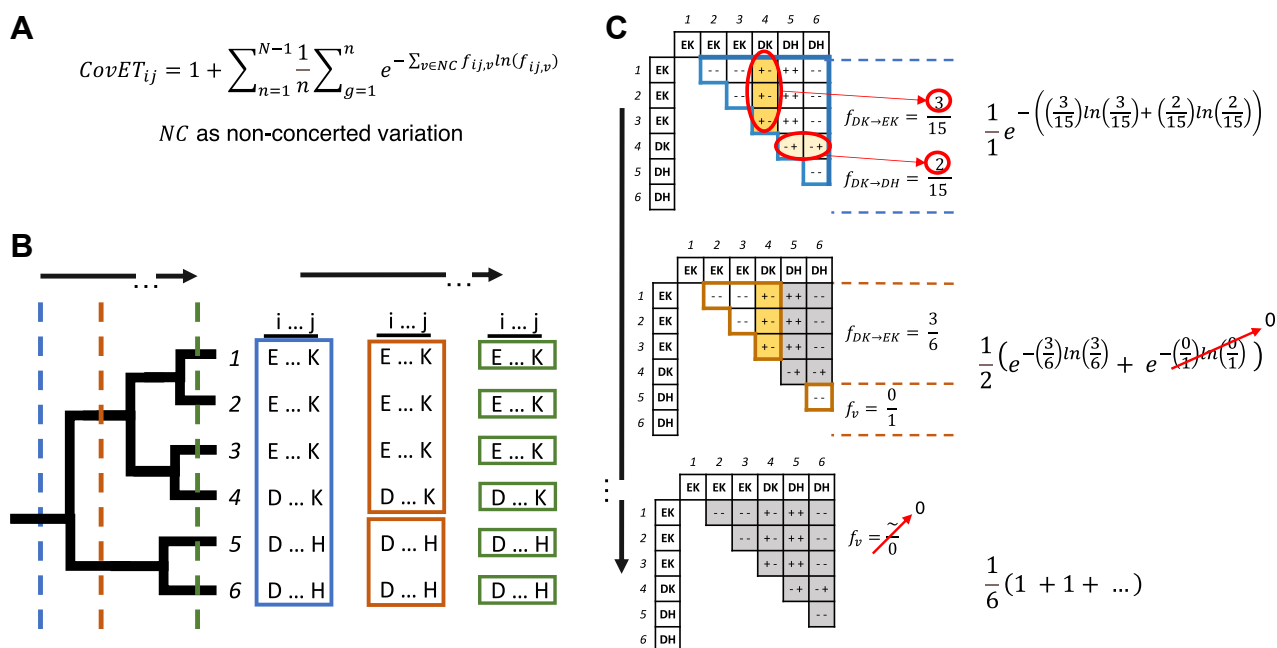


Figure 1. CovET Pipeline. A, CovET Equation, ij is a pair of residues, N is the number of sequences in the alignment, n is the division in the tree, g is the group of the tree at division n , v is one of the 17,640 possible nonconcerted variations in a 21 character alphabet, $f_{ij,v}$ is the frequency of nonconcerted variation v in group g . B, vertical dashed lines show division of the tree into groups (blue $n = 1$, orange $n = 2$, green $n = 6$), boxed sequences are resulting groups (g) from each division. C, the substitution matrix for each division in the tree. Instances of nonconcerted variation are shown in different shades of yellow and are used in the penalty function (right). The boxed regions of the matrix represent the pairwise comparison, defined on a per group basis (second substitution matrix has two boxed regions, corresponding to the two groups in the tree at that division). In each box the comparison of the residues in each pair is indicated by the + and - symbols, with + meaning variation and - meaning conservation (*i.e.*, ++ concerted variation, - conservation, and +- or -+ non-concerted variation).

as well as our previous covariation algorithm, ET-Mip which used the same ET framework but used mutual information to identify covariant residues (16). Based on the area under the receiver operating curve (AUROC, Fig. 2A), CovET improved over DCA and ET-Mip in predicting structural contacts, especially in the medium and long-range categories (Fig. S1) but was outperformed by EVCouplings. Based on the area under the precision recall curve (AUPRC, Fig. 2B), which is less sensitive to the class imbalance between positive and negative cases, CovET only outperforms ET-Mip. Importantly, all methods performed better than what would be expected from a random predictor.

A hallmark of phylogenetic strategies to predict the functional importance of individual residues is that the top predictions cluster structurally in 3D space (45–48), which would be useful in higher order protein structure and function determination. Therefore, we asked next whether covarying residue pairs identified by CovET and other methods were structurally random or clustered into subregions of structural and functional importance. For this, we converted the pairwise covariation scores for each of the 943 proteins in our Pfam dataset to single residue scores (based on the order of first appearance in a pair) and applied two complementary measures of structural clustering: the biased and unbiased selection cluster weighting (SCW) z-scores (41, 45, 46). Both scores

measure how clustered a group of residues are on a given protein structure, but the biased SCW z-score gives greater weight to clusters of residues more distant in sequence. SCW z-scores were measured for the top 30% of predicted residues in accordance with previous studies (45, 46). While all methods achieved significant clustering (unbiased SCW z-score >2) in most proteins (Fig. 2C), top-ranked CovET pairs are more significantly clustered than those of EVCouplings, DCA, and ET-Mip (CovET mean unbiased Z-Score 9.9, DCA 7.1, EVCouplings 7.0, ET-Mip 8.9). This advantage grew when measured by the biased SCW z-score (CovET mean biased Z-Score 3.9, DCA 0.9, EVCouplings 1.0, ET-Mip 0.1, Fig. 2D), with CovET strongly outperforming all other methods. These results showed that CovET is better at identifying amino acid structural clusters than individual local contacts, which is fundamentally different from the other three covariation methods.

Finally, given CovET's performance in the biased SCW z-score, we compared the average sequence separation (normalized by protein length) of top predicted pairs for all methods (Fig. 2E). DCA, EVCouplings, and ET-Mip disproportionately rank pairs of residues that are close together in the sequence among their top results. In sharp contrast, pairs predicted by CovET are farther apart across the sequence. These data show that CovET pairs are qualitatively and

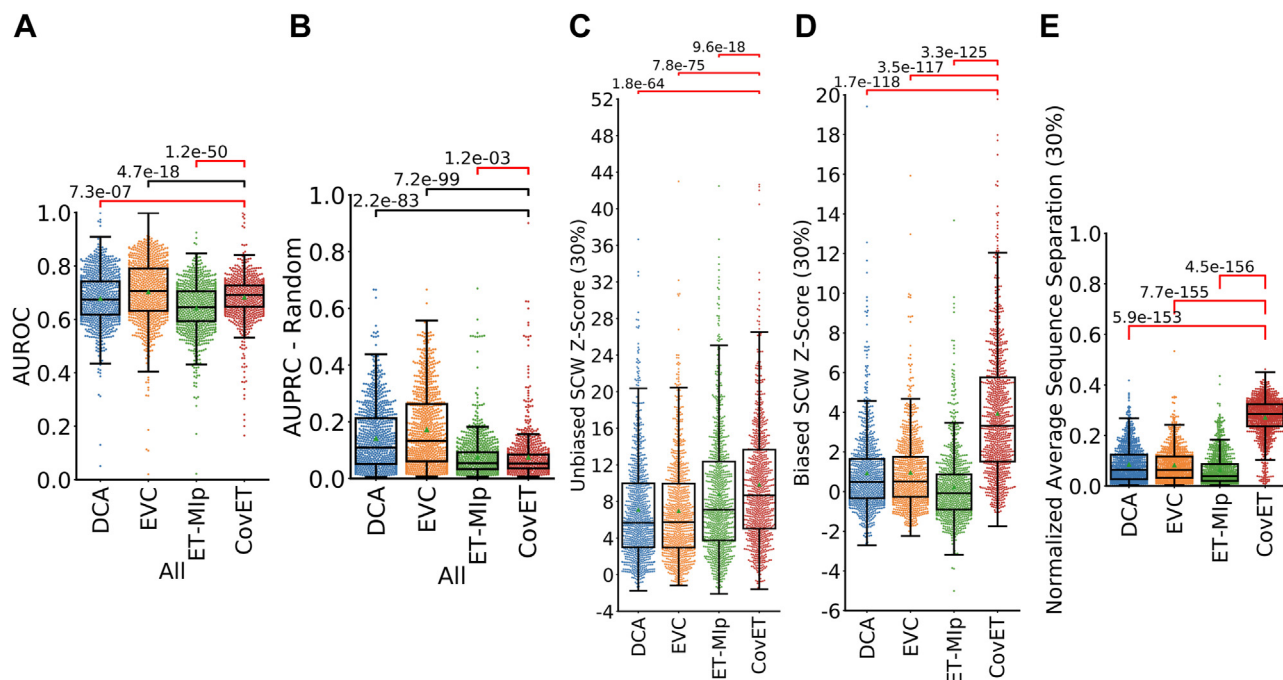


Figure 2. CovET recapitulates the performance of existing methods in predicting structural contacts, but its predictions are more significantly clustered structurally. A, AUROC measured for all contacts (residues where C β are within 8 Å of each other, C α for glycine) at least six residues apart. Pane (B) shows the same contacts measured at the same sequence separation categories evaluated by AUPRC. This evaluation which is less influenced by the class imbalance between contacts and noncontacts. The expected AUPRC value from a random predictor is the positive rate in the standard, which is the number of contacts/total residue pairs. AUPRC—Random evaluates whether the predictor performs better than random. C and D, unbiased (C) and biased (D) Selection cluster weighting (SCW) z-scores measured for each protein and each of the tested methods. While all methods show significant (z-score >2) clustering for most proteins, CovET shows significantly more clustering than DCA, EVCouplings (EVC), and ET-Mip in unbiased SCW z-scores and significantly better clustering than the three other methods in biased SCW z-scores. E, the average distance between pairs at 30% coverage is much higher for CovET than for any of the other methods assessed, confirming that the increase in biased SCW Z-Scores is due to clustering of residues which are significantly further apart in primary sequences. Significance was measured using the paired, two-sided Wilcoxon rank sum test. When CovET significantly outperforms other methods, the comparison bars are colored red. AUPRC, area under the precision recall curve.

CovET identifies protein structure–function modules

structurally different. They involve more distant positions in the sequence (Fig. 2E), yet these cluster together in the protein structure much better than the pairs predicted by other methods. Together, these results show that CovET captures coevolutionary relationships involving sequence neighbors, like other methods, but also positions far apart in the primary sequence, a unique property of this method.

CovET identifies important functional residues

Having confirmed the structural clustering of top CovET predictions, we assessed CovET's ability to recover functionally important residues, like other phylogenetic methods which cluster well. It is impractical to manually gather the functionally important residues for every protein in the Pfam dataset. Instead, we turned to the BioLiP database to gather information on biologically relevant ligands in our Pfam queries (49). Defining residues within 4 Å of biologically relevant ligands as being functionally important, we could characterize them in 329 of our Pfam queries from the BioLiP database. As shown in Figure 3A, CovET was the best at recovering these functionally important residues with an average AUROC of 0.746, while the other three methods have average AUROCs below random (0.5). As measured by AUPRC, CovET still significantly outperforms the

other methods (Fig. 3B). These results suggested that unlike other methods, CovET is uniquely able to identify functionally relevant residue pairs, as opposed to solely structural neighbors.

To assess CovET's ability to identify functional sites in greater detail, we gathered key functional site information from the literature for two protein domains (the RNA recognition motif (RRM) domain and the WW domain) and for the single-domain dopamine D2 receptor (D2R) and tested how well CovET and the other covariation methods could recover their functional sites. Multiple sequence alignments of each were obtained (see Experimental Procedures, Fig. S2 and Table S2) and covariation analyses performed. The RRM domain is found across a wide variety of organisms including eukaryotes, prokaryotes, and viruses (50–53). This domain is ~90 amino acids long and plays a role in almost all stages of RNA processing (51, 53–57). The RRM domain is built mainly from a β -sheet that contains two highly conserved motifs, RNP1 and RNP2, that are essential to the domain's main function, binding poly(A) RNA (50, 51, 53, 55, 58). CovET picked out many of the key residues from the β -sheets that bind RNA and others in the C-terminal region that stabilizes the RNA binding network. CovET significantly recovers the RNP1 motif at low coverage, with p -values of 0.042 at 20%

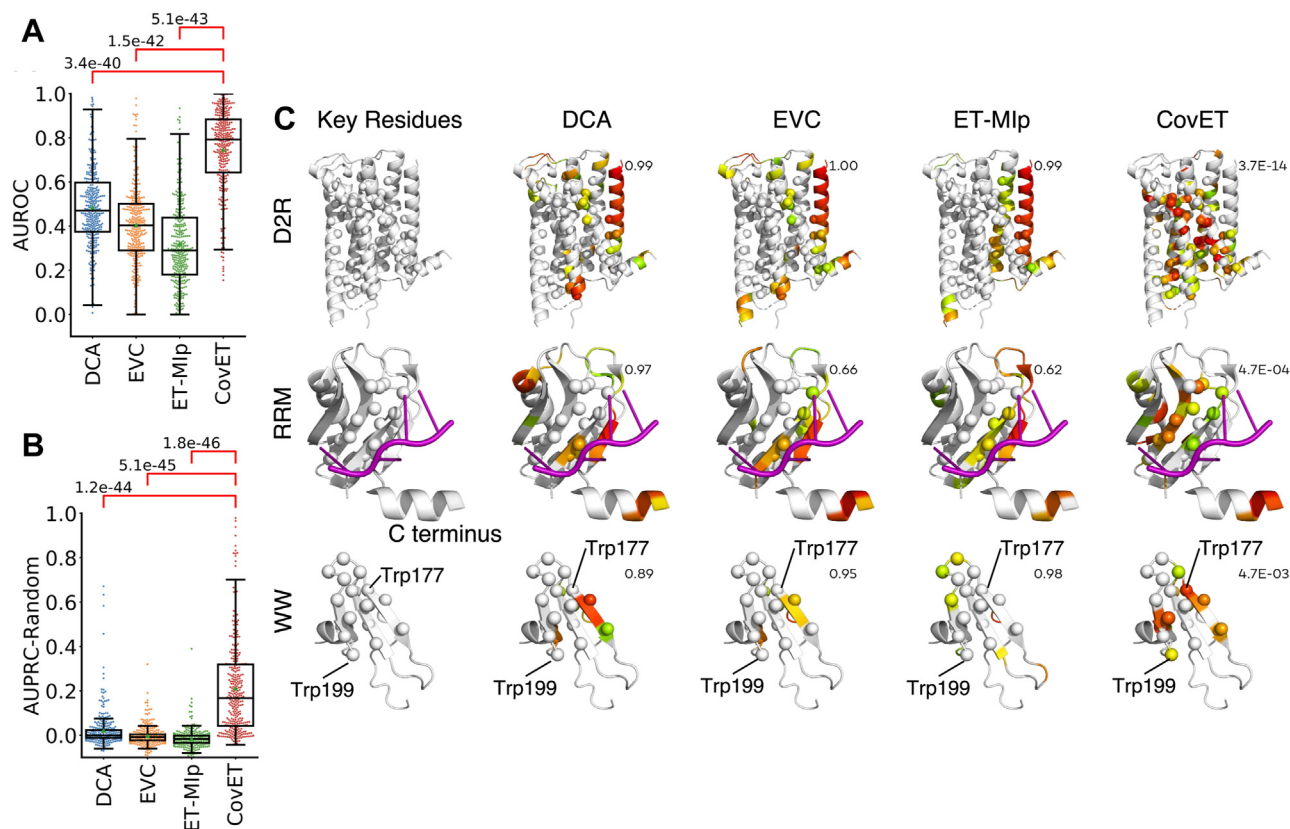


Figure 3. Top CovET residues significantly overlap with known functional sites. A and B, CovET performs the best in recovering ligand binding sites in the pfam dataset judging by both AUROC (A) and AUPRC (B). Similar to Figure 2B, the AUPRC values are adjusted with the AUPRC of random predictors for each protein. Significance was measured using the paired, two-sided Wilcoxon rank sum test. When CovET significantly outperforms other methods, the comparison bars are colored red. C, top CovET predictions significantly overlap with known functional residues in the dopamine 2 receptor, RRM, and WW domain. Known functional residues visualized as spheres. Top 30% of residues shown on a red to green color scale, with red indicating the most highly ranked residues, and green those just at the 30% threshold. False positives shown as color scaled ribbons, false negatives shown as white spheres. Significance was measured using the one-sided hypergeometric test. The overlap between top predictions and known functional residues are evaluated with hypergeometric test. p -values are shown to the upper right of each structure. AUPRC, area under the precision recall curve; AUROC, area under the receiver operating curve; D2R, dopamine D2 receptor; RRM, RNA recognition motif.

coverage and 0.008 at 30%, while the other methods failed to identify this site by 30% coverage (Figs. 3C and S3; Table S3). At 30% coverage, DCA, EVCouplings, and ET-MIP identified some residues in the C-terminal region, but these were disconnected from the few residues these methods recovered in the β -sheets (Fig. S4). In sharp contrast, the top pairs predicted by CovET connected the RNA binding β -sheets and the C terminus. These data show that CovET preferentially identifies key sequence positions in the RRM domain. Additionally, top CovET predictions formed a dense and interconnected network in the RRM domain: many of the top pairs had overlapping residues. In contrast, top pairs from other methods form a sparser network and reach a given residue coverage with fewer pairs than CovET. For the RRM and WW domains as well as the D2R, CovET has more residues pairs and a denser network for a given coverage cutoff (Figs. S4–S6). This is because coverage cutoffs are defined as all the pairs it takes to recover a certain percent of residues in the protein. For example, a 10% cutoff is defined as all the residue pairs it takes until 10% of residues are recovered by at least one pair. Methods whose top pairs repeatedly select the same residues will have more pairs for a given coverage. CovET consistently takes more pairs to reach a certain coverage, which highlights the interconnectedness of top CovET pairs.

We next assessed the recovery of key residues in the WW domain. WW domains are 30 to 50 amino acids long and have one of the smallest spontaneously folding β -sheet structures, consisting of three anti-parallel β -strands (59–66). This domain enables protein–protein interactions by binding short peptide sequences (e.g., PPxY in hYAP65) (59–66). The most important sites include the two tryptophans (W) for which the WW domain is named (59–61, 65, 66) and two hydrophobic patches, one involved in binding and the other in stability (61, 63–65). At 10% coverage, CovET recovers one titular tryptophan, while the second one is recovered at 20% coverage. Neither is recovered by DCA, EVCouplings, or ET-MIP by 30% coverage (Fig. 3C and Table S4). At 30% coverage, CovET is the only method to significantly recover the combined set of conserved residues and functional sites (Figs. 3C and S5). None of the other covariation methods reaches significance for the highlighted sites or the combined set of conserved residues and functional sites. In addition, the WW domain contains residues that can be mutated without causing significant perturbation to its structure or function, dubbed insensitive residues here, mostly in the N and C termini and the turns between β -strands (61, 65). Here, the pattern is reversed, and CovET does not pick up any insensitive residues by 30% coverage, while EVCouplings and ET-MIP begins to pick these residues up at 10% coverage and DCA by 20% coverage (Table S5). These data show that CovET preferentially identifies key conserved residues and functional sites in the WW domain, while DCA, EVCouplings, and ET-MIP recover known variable positions.

CovET highlights functional networks in the D2R

If top-ranked CovET residue pairs have important structural and functional interactions, we would expect that they would

form highly interconnected networks in proteins that mediate long-range information transfer across a protein *via* allosteric pathways. To test this hypothesis, we focused on the D2R, a member of the Class A G protein-coupled receptor (GPCR) family, the single largest family of eukaryotic proteins (67, 68). Class A GPCR's have arguably the most-studied allosteric signaling mechanism that depends on the concerted motions of its seven transmembrane helices, in which various functional modules have been described in detail (36, 67–77).

To predict covarying residues, we aligned the whole sequence of 2568 class A GPCRs probing a wide cross-section of bioamine receptors and species (Fig. S2A). The top-ranked pairs predicted by CovET created a modular network that significantly incorporates experimentally described components of the allosteric pathway extending from the ligand binding site to the G protein coupling and β -arrestin site (p -value $3.7E-17$ at 30% coverage) (Figs. 4 and S6; Table S6). Moreover, as the network is built-up over a span of CovET ranks, it reveals an underlying modular evolutionary hierarchy that implies various degrees of functional importance. This is because CovET estimates the covariation pattern of residue pairs in the context of the evolutionary tree. The best-ranked CovET pairs have residues that do not vary or if they do, always in concert with the main evolutionary tree branches. As coverage increases, concerted variation tracks with lesser tree branches until, eventually, additional pairs follow no discernible correlation with the phylogenetic tree branches.

Specifically, at only 2.5% coverage, CovET picks up two sets of known functional residues. One of the discrete clusters (Fig. 4A, lime, Table S7) overlaps entirely with the NPxxY (67, 68, 70, 72, 74, 78) motif (p -value $3.34E-6$), while both clusters partially recover the water channel (72) (p -value $3.83E-7$), the Na⁺ binding site (67, 68, 72, 74, 78) (p -value $1E-4$), and known state determinants (70, 78) (Fig. S7). At 5% coverage, CovET further adds three discrete clusters, one of which (Fig. 4B, red) overlaps with the CWxP (68, 70, 78) motif (p -value 0.006), and the canonical toggle switch (67) (p -value 0.012). At 7.5% coverage, CovET recovers the HHM (67, 68, 72) (p -value 0.014) and ionic lock (68, 70) (p -value 0.027) as well as switches (68, 70) involved in the allosteric activation pathway (p -value 0.001). Two of the prior clusters (Fig. 4B, blue and light blue) now merge and expand. As a result, nearly the entire structurally central and functionally pivotal transmembrane helix 3 (TM3) is recovered (Figs. 4C, blue and S7). At 10% coverage, CovET predictions significantly recover the DR[E]Y motif (67, 70, 72–74, 78) and ligand binding site (74, 79–82) (p -value 0.026 and 0.024), while the four discrete clusters are maintained, and the one centered around TM3 continues to expand. At 12% coverage, the cluster including the CWxP motif becomes connected to the central TM3 cluster (Fig. 4E, blue). At 14% coverage, CovET recovers the PIF[W] (70, 78) motif (p -value 0.05), and the NPxxY motif cluster joins the large network cluster centered on TM3. At 16% coverage, two main clusters are now formed. The larger one spans from TM3 to TM7, while the smaller one is localized between TM1 and TM2. At that coverage, CovET recovers residues from almost all motifs and conserved elements that have been implicated in allosteric signal transduction (p -value

CovET identifies protein structure–function modules

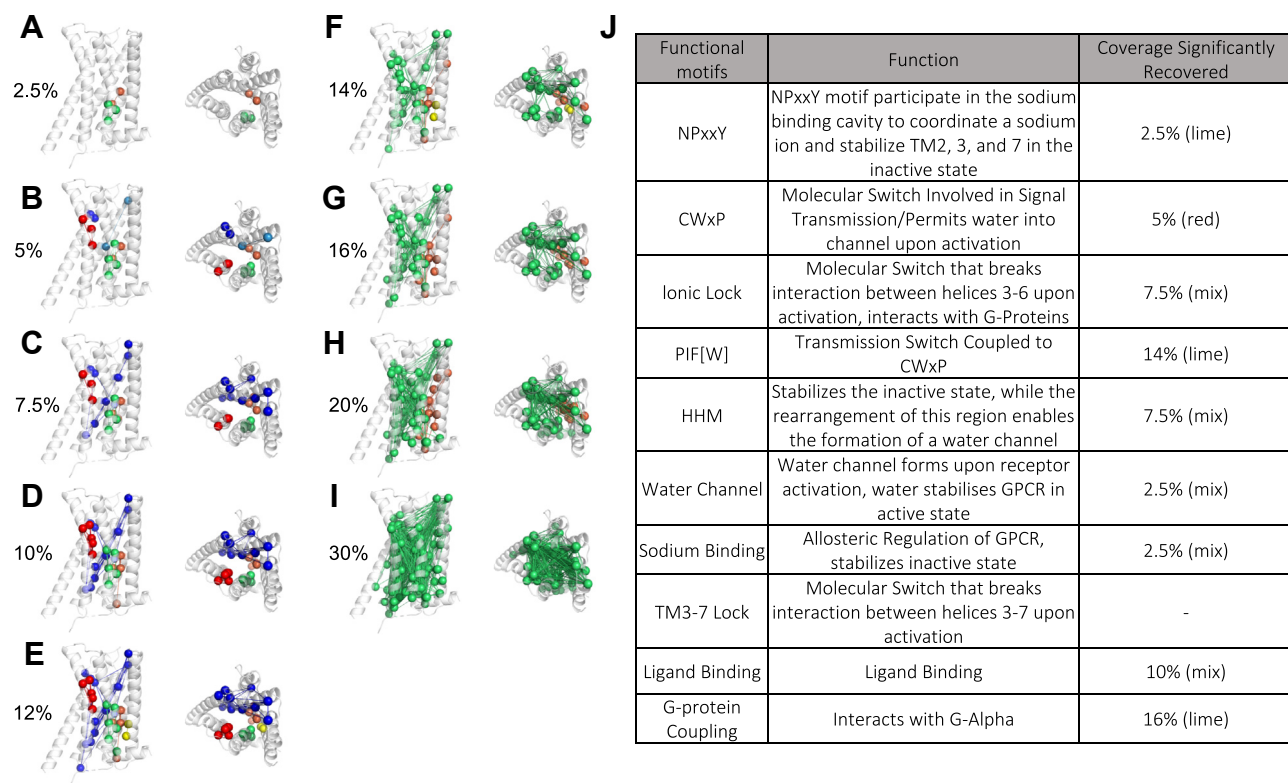


Figure 4. CovET predictions highlight functional network in the dopamine D2 receptor. *Top* CovET pairs at different coverages are displayed as *lines*. Residues involved in the pairs are shown as *spheres*. The *top* residues and pairs are colored based on the discrete clusters they belong to at each coverage cutoff. Increasing coverage shows a growing, highly interconnected network that covers the core of the protein and the GPCR allosteric pathway. The composition of the discrete clusters can be found in [Table S7](#). *A–I*, CovET network formed at various coverage cutoffs in the D2R. *J*, most functional motifs in the D2R are recovered by CovET network at various coverage cutoffs. GPCR, G protein-coupled receptor; TM3, transmembrane helix 3.

1.21E-12), except the TM3-7 (68) lock (not recovered by 30% coverage) and the G-protein coupling site (recovered at 30% coverage, p -value 0.028). At 16% coverage, 42 residues are identified by CovET, 31 (~74%) of which overlap with the ligand or G protein-binding sites, known motifs, state determinants, or members of the allosteric network. Interestingly, some residues are not described among previously recognized functional positions. These include: 65, 75^{2.45}, 89^{2.59}, 97, 100, 107^{3.25}, 139, 160^{4.50}, 191^{5.40}, 192^{5.41}, and 198^{5.47} (superscripts refer to the Ballesteros-Weinstein numbering system for GPCR TM segments) (83). Residues 97, 100, and 107^{3.25} form a small local cluster ($C\beta$ within 8 Å) in the structure, while the hydroxyl group in Ser75^{2.45} is hydrogen bonded with the side chain of Trp160^{4.50} (Fig. S8). Thus, these additional CovET residues are not randomly distributed in the structure and would be of interest for further mutational testing. The remaining two clusters continue to grow at 20% coverage and merge into a single connected network, spanning the 7TM from end to end at 30% coverage. This ranked order of appearance of functional modules that grow and coalesce to span the entire known allosteric pathway is in sharp contrast to the interaction networks identified by DCA, EVCouplings, and ET-MIP which localize to TM1 and do not overlap significantly with key sites (Fig. S6 and Table S6). Together, these results show that CovET recovers hierarchically and specifically the structural and functional

motifs most critical components to the allosteric pathway mediating signal transduction in Class A bioamine GPCR.

Pairs predicted by CovET correlate with epistatic interactions

To test if CovET predicts intraprotein epistatic interactions, we evaluated the correlation of the top predictions with large-scale mutational studies. We analyzed five high-throughput deep mutagenesis studies that measured the fitness of double mutants: the RRM domain (50), the WW domain (84), TEM-1 β -lactamase (85), the IgG-binding domain of protein G (GB1) (86), and the prion-like domain of TDP-43 (TAR DNA-binding protein 43) (87). Because the best epistasis model is unknown, we computed epistasis scores using four commonly applied models (Additive, Log, Min, and Product) (84) for each of the datasets. Since CovET and other covariation methods tested here do not consider the impact of specific mutations pairs and have no directionality, we took the mean and absolute value of all epistasis scores available for a given residue pair, resulting in a final score that represents the average deviation from wildtype behavior. In the RRM dataset, CovET predictions correlate the best with these experimental epistasis scores for all four commonly applied epistasis models (Pearson correlations per epistasis model: Min 0.239, Log 0.170, Product 0.177, Additive 0.460, Figure 5). The next best correlations,

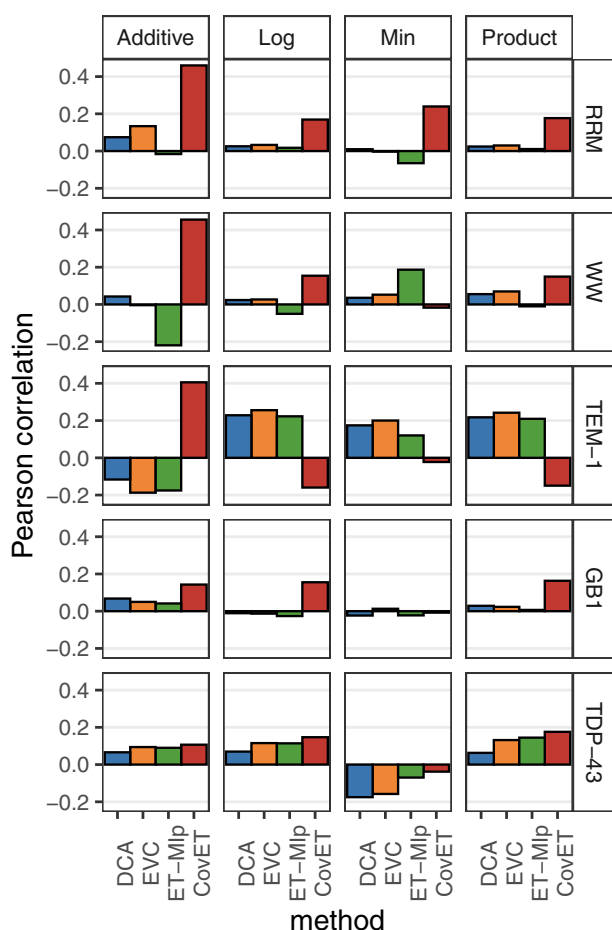


Figure 5. Correlation of covariation scores with epistasis scores for three proteins. Epistasis scores were calculated using four different models, and the correlation between the calculated and the average displacement from wildtype was calculated for each pair. These values and covariation scores from the DCA, EVCouplings (EVC), ET-Mip, and CovET methods were measured for their Pearson correlation. For both domains, the evaluated epistasis models were Min, Log, Product, and Additive. The negative raw score for CovET was used because unlike other methods, the higher CovET raw score means a less coupled pair. GB1, IgG-binding domain of protein G; RRM, RNA recognition motif; TAR DNA, prion-like domain of TDP-43.

from EVCouplings, were much lower across every epistasis model (Fig. 5). The WW domain dataset consisted of 47,000 variants, including 5010 double mutants (~2.5% of possible) (84). CovET correlated best with the experimentally determined epistasis scores using Log, Product, and Additive models, with Pearson correlation scores of 0.154, 0.149, and 0.456, respectively. For the minimum epistasis model, only ET-Mip correlates with experimental data with highest Pearson correlation of 0.187. For TEM-1, CovET gives the best correlation in the Additive model (Pearson correlation 0.406), which is also the best correlation among all epistasis models. In GB-1 and TDP-43, CovET correlates the best with epistasis scores computed using Additive, Log, and Product models, while none of the covariation methods correlates with the Min epistasis model. In addition, when considering all four epistasis models together, CovET best correlates with experimental data in all five proteins tested (Fig. 5, Pearson correlations: RRM Additive 0.460, WW Additive 0.456, TEM-1 Additive 0.406, GB-1 Product 0.163, and TDP-43 Product 0.176). These

results suggested that CovET better predicts protein residue functional couplings than DCA, EVCouplings, and ET-Mip and suggests that phylogenetic couplings are generally in better agreement with Additive epistasis.

Discussion

Our study introduces a new method to identify evolutionarily coupled sequence positions. CovET is distinct from existing methods (13, 14) in two ways: first, it explicitly accounts for phylogenetic history by applying the ET framework to score every pair of residues recursively along successive evolutionary tree partitions. Second, CovET exclusively penalizes uncoupled variations within a partition's branch, and pairs that are completely conserved or that covary within a branch are not penalized. We find that, unlike other methods, top CovET predictions form significant spatial clusters that overlap with ligand binding sites in a diverse set of protein families, as well as known functional sites in the RRM and WW domains and in the D2R. The functional relevance of top CovET pairs is further supported by deep mutational scans where CovET predicted epistatic interactions between residues better than other methods. Moreover, the structural clusters and networks defined by top-ranked CovET pairs are functionally relevant. In the D2R, for example, top CovET residues form dense clusters of mutually coupled pairs that overlap with key functional sites and reveal, at increasing coverage, nearly the entire canonical allosteric network. The hierarchical nature of this network, reminiscent of past ET analysis (36), suggests that CovET captures the functional and evolutionary architecture of the Class A GPCR transduction mechanism through couplings that define the allosteric pathway. Given how well CovET recovers the Class A GPCR allosteric activation pathway, it would be interesting to test the couplings predicted by CovET in the D2R. In addition, in our recent study, residues 180 and 181 in the metabotropic glutamate receptor 4 were predicted to be highly covariant by CovET. Indeed, residue position 180 and 181 demonstrated a strong epistatic interaction toward ligand binding in metabotropic glutamate receptor 4 (88). Similar experiments can be conducted on highly ranked CovET residues in the D2R. Further validating our coupling predictions in the D2R *in vitro* or *in vivo* could potentially strengthen our understanding of its physiologically (71, 72, 74, 78) and therapeutically relevant (70, 72, 74, 78) allosteric activation pathway.

The performance of CovET is rooted in the phylogenetic logic of the ET framework. Evolution proceeds from random sequence variations followed by functional selection, repeated at each generation (89–91). Therefore, the pressure for two residues to comutate, or not mutate at all, will be large when the interaction between the residues maintains an important protein function within a narrow range of fitness tolerance. Conversely, that pressure will be small, or nil, if a large change in protein function is well tolerated by an organism in its new adopted environment. Thus, covariation patterns between two sequences in a protein family must be interpreted with respect to the function each protein serves in its environment. While direct functional assessments of proteins are sparse, the ET

CovET identifies protein structure–function modules

approach instead uses distances among phylogenetic branches since these are defined through a sequence metric that approximates functional aptitude groupings (17, 89, 90). As a result, nonconcerted variations across short evolutionary ranges, where function is likely preserved, should strongly indicate a lack of coupling and be penalized more severely than nonconcerted variations across long ranges in the tree. CovET captures the functional context of covariation patterns by scoring the complete alignment at the tree's root node and each subalignment spawned by a divergence event. In contrast: traditional covariation methods are performed over a complete alignment, without consideration of phylogenetic structure (92). Such an approach intrinsically assumes that all sequences share the same structure and function, which is less and less likely as we consider proteins that are further and further apart in sequence identity and in an evolutionary tree. These methods may then be confounded by less meaningful covariations in functionally unrelated regions of the tree. In addition, most covariation models penalize pairs of residues that are invariant together (93), even though, presumably, many of these important positions are essential for protein function. CovET considers both conservation and covariation as positive signals and penalizes neither. By exclusively penalizing nonconcerted variations—the only signal that one can be sure does not support covariation—CovET avoids uncertainties regarding invariant and mostly invariant residues.

Given its unique approach to identifying coupled residues, CovET may provide an orthogonal feature set to current covariation methods in protein structure prediction machine learning systems (13, 14). CovET nearly recapitulates the performance of other methods in predicting structural contacts, but its predictions are fundamentally unique, they are further apart in sequence, yet significantly spatially clustered. CovET may also be used to improve genotype-phenotype predictions. Taking a more formal approach, we previously used ET to approximate the first derivative of the evolutionary landscape function $f(\gamma) = \Phi$, which maps genotypes γ to their fitness potential Φ (94). This approach allowed us to estimate the functional impact of coding variants with high accuracy (95) and was further validated with diverse practical applications (96–101). However, ET only evaluates the functional importance of single residues, thus missing higher order interactions among residues. CovET may be interpreted as the mixed second derivative of f with respect to residue pairs. In the future, with proper scaling and sign, the addition of this second-order epistatic term may improve our approximation of the evolutionary landscape function and our understanding of the genotype-phenotype relationship.

In summary, CovET predicts functionally coupled sequence positions by accounting explicitly for phylogenetic divergences during evolution. This approach enriches current views of residue couplings by informing whether variants occur in preserved or divergent functional contexts. Examples from functional sites and an allosteric pathway suggest this approach may provide additional insights to understand protein structure and function and machine learning features to predict them.

Experimental procedures

Sequence retrieval and multiple sequence alignment construction

To test our algorithm on the most diverse set of proteins possible, we turned to the Pfam database (42). The protein families used in this study are summarized in Table S1. The Pfam database contains a comprehensive list of protein families, each represented by a multiple sequence alignment. We extracted alignments for each family and removed those where no family member had an available experimental structure with at least 3.5 Å resolution. In the case where multiple structures were available for a family, the structure that best aligned to its respective linear sequence was assigned as the reference structure for the family; its linear sequence was assigned the query sequence. Alignments were then filtered using hhlblits (102) to minimum 70% coverage with query, minimum 30% sequence similarity to query, and 98% maximum pairwise sequence similarity. Larger alignments were filtered to contain the most diverse 2000 sequences. All filtering steps were done in tandem using the command: `hhfilter -i <alignment_path> -cov 70 -qid 30 -id 98 -diff 2000`. Due to the presence of small proteins with few contacts (e.g., single α helices), families were further filtered to have at least one short-range (6–11 amino acids apart in sequence), medium-range (12–24 amino acids apart in sequence), and long-range (24+ amino acids apart in sequence) structural contact (103, 104). The Pfam database is further grouped into clans of families that share an evolutionary origin. To ensure coverage of the entire known proteome, for each clan in the Pfam database that has at least one family that meets the above criteria, we used the family with the highest number of sequences to perform CovET and other methods.

The RRM domain (105), the WW domain (60), the D2R (70, 78), TEM-1 β -lactamase (85), the GB1 (86) and the prion-like domain of TDP-43 (87) were used as detailed examples in this study (Table S2). Homologous sequences for these proteins were retrieved by using the blastp utility of the BLAST+ tool (106) to search the UniProt90 sequence database (downloaded from <https://www.uniprot.org/downloads>) (107). This sequence database was first filtered to remove all sequences including the terms “Fragment” or “Low Quality”. Then, it was used to create a database compatible with the blastp tool using the makeblastdb utility also provided with the BLAST+ tool. Blastp (version 2.9.0, build May 27, 2019) was run using the filtered database as the target and a max e-value cutoff of 0.05 (41) and a max sequence return of 20,000. Sequences identified by the BLAST search were further filtered such that only sequences covering at least 70% of their query and with identity of $25\% \leq \text{query} \leq 98\%$ (41, 89) were kept. Sequences were also removed if they included amino acids other than the standard 20 or gaps, if their description included the terms “artificial”, “fragment”, “low quality”, “partial”, or “synthetic”, or if their taxonomy recorded in UniProt included the terms “synthetic” or “artificial”. The sequences passing these filters were aligned using the ClustalW tool (version 2.1) (47, 108) using the quicktree option. To reduce redundant sequences, all pairwise identities were calculated between sequences in the constructed alignments. Sequences were removed such that any cluster of

sequences with >98% sequence identity to each other were represented by only one sequence. The alignment was re-aligned using ClustalW with the same settings after this filtering step.

Phylogenetic reconstruction under ET framework

In the ET framework (17, 41, 109), the distance between any two sequences in the multiple sequence alignment is computed by:

$$\text{Dist}(seq_a, seq_b) = 1 - \frac{\sum_{i=1}^l f(seq_{a,i}, seq_{b,i})}{\min(\sum_{i=1}^l g(seq_{a,i}), \sum_{i=1}^l g(seq_{b,i}))} \quad (1)$$

where seq_a is the protein sequence at the a th row in the multiple sequence alignment. $seq_{a,i}$ is the amino acid in the i th position in a th sequence (character at [a, i] when considering MSA as a matrix), and l is the column count of the MSA.

$$f(x, y) = \begin{cases} 1, & \text{if } BLOSUM62(x, y) \geq 2 \\ 0, & \text{if } BLOSUM62(x, y) < 2 \end{cases}$$

where $BLOSUM62(x, y)$ is the log odds of between amino acid character x and y in the BLOSUM62 matrix. The log odds between gap character and any character (including itself) are manually set to 0.

$$g(x) = \begin{cases} 0, & \text{if } x \text{ is the gap character} \\ 1, & \text{otherwise} \end{cases}$$

A UPGMA tree is generated across the sequences using the distance matrix (110).

Covariation predictions

CovET is first described here, and the code is available in the accompanying codebase. The formula for the covariation of a pair of positions is given by:

$$\text{CovET}_{ij} = 1 + \sum_{n=1}^{N-1} \frac{1}{n} \sum_{g=1}^n e^{-\sum_{v \in NC} f_{ij,v} \ln(f_{ij,v})} \quad (2)$$

$$v_{ab,ij} \equiv (x_{a,i}, x_{a,j}, x_{b,i}, x_{b,j})$$

$$x_{a,i}, x_{a,j}, x_{b,i}, x_{b,j} \in \{20 \text{ amino acids, gap}\}$$

$$v_{ab,ij} \in NC, \text{ if } (x_{a,i} = x_{b,i} \ \& \ x_{a,j} \neq x_{b,j}) \text{ or } (x_{a,i} \neq x_{b,i} \ \& \ x_{a,j} = x_{b,j})$$

where ij is a pair of residue positions, N is the depth of the phylogenetic tree, n is a level in the phylogenetic tree, and g is

a group of sequences (branch) at that level in the tree. The term in the second sum is the diversity metric or perplexity, which is the exponential of the Shannon entropy. When comparing a pair of residues in two sequences in a 21-letter alphabet for all amino acids and a gap character, there are 194,481 (21^4) possible outcomes for the 20 standard amino acids plus a gap character. These possible outcomes can be classified as conservation ($AD > AD$), concerted variation ($AD > CE$), and nonconcerted variation ($AB > AE$). Four hundred one (21^2) of them are conservations. 176,400 ($21 \times 21 \times 20 \times 20$) of them are concerted variation. The remaining 17,640 of them are nonconcerted variations $\alpha^2 \times 2(\alpha - 1)$, respectively, where α is the number of characters used. v in the entropy calculation is any of the 17,640 possible non-concerted variations (NC). The process of characterizing a pair of positions for a given group, g , is demonstrated in Figure 1. By scoring these transitions, nonconcerted variation is penalized, while conservation and concerted variation, or covariation, are not.

ET-MIp (16) was reimplemented in Python and is available with the codebase distributed with this study. ET-MIp score for a pair of residues are given by the equation:

$$\text{ETMI}_{p,ij} = \sum_{n=1}^{N-1} \frac{1}{n} \sum_{g=1}^n \text{MI}_p^g(i, j) \quad (3)$$

where ij is a pair of residues, N is the depth of the phylogenetic tree, n is a level in the phylogenetic tree, and g is a group of sequences (branch) at that level in the tree. $\text{MI}_p^g(i, j)$ is mutual information with the product correction (111) given by:

$$\text{MI}_p(i, j) = \text{MI}(i, j) - \text{APC}(i, j) \quad (4)$$

where $\text{MI}(i, j)$ is the mutual information (112):

$$\text{MI}(i, j) = H_i + H_j - H_{ij} \quad (5)$$

where H_i and H_j are the Shannon entropy of positions i and j , respectively, and H_{ij} is the joint entropy of the pair of position. $\text{APC}(i, j)$ is the average product correction (111) given by:

$$\text{APC}(i, j) = \frac{\text{MI}(i, \bar{y}) \cdot \text{MI}(j, \bar{y})}{\overline{\text{MI}}} \quad (6)$$

where $\text{MI}(i, \bar{y})$ and $\text{MI}(j, \bar{y})$ are the mean MI with respect to position i or j , and $\overline{\text{MI}}$ is the mean mutual information over the pair of positions.

EVcouplings (13, 43) covariation scores were computed using the implementation provided at <https://github.com/debbiemarkslab/EVcouplings>. To perform EVcouplings predictions, the provided config file was used to set the following settings for each run. In the “global” settings: “region” = None, “theta” = 0.8. The “pipeline” was set to “protein_monomer”. “batch” was set to None. The “stages” were set to [“align”, “couplings”]. For the “align” settings: “protocol” = “existing”, “input_alignment” was set to the path of the alignment file used by all

CovET identifies protein structure–function modules

prediction methods, “first_index” = 1, “compute_num_effective_seqs” = False, “seq_id_filter” = None, “minimum_sequence_coverage” = 0, “minimum_column_coverage” = 0, “extract_annotation” = False. These settings were used to ensure that the provided alignment was used without alteration, so that results would be comparable across methods. For the “couplings” settings: “protocol” = “standard”, “iterations” = 100, “lambda_J” = 0.01, “lambda_J_times_Lq” = True, “lambda_h” = 0.01, “lambda_group” = None, “scale_clusters” = None, “alphabet” was set to the ‘-’ gap character and the 20 standard amino acids, “ignore_gaps” = False, “reuse_ecs” = True, and “min_sequence_distance” = 0. The “global” variables: “prefix”, “sequence_id”, and “sequence_file” were provided separately for each protein.

DCA was computed using the multivariate Gaussian approach implemented in Julia (44) and available at <https://github.com/carlobaldassi/GaussDCA.jl>.

Evaluation of structural contacts

To evaluate structural contacts, query sequences were first aligned with the corresponding structures (“PDB” column of Table S1), and only positions present in both were considered for evaluation. Contacts were then determined using the definition used in the CASP competitions, only residues at least six residues apart in sequence and whose C β (C α for glycine) were within 8 Å of each other were considered true contacts (103, 104, 113–118). Contacts were broken into three categories, also defined in CASP by number of amino acids of separation, which were short (6–11), medium (12–23), and long (≥ 23) (103, 104, 117). Finally, predictions were compared against the combined set and three subsets of contacts using the AUROC and AUPRC, as implemented in Sklearn (119). All predicted contacts were evaluated, not just top contacts, as suggested in a recent CASP competition (104). The AUPRC were then adjusted by subtracting the observed positive rate, which is the expected AUPRC for a random predictor. The AUROCs and adjusted AUPRCs of each method were compared using paired two-sided Wilcoxon Rank Sum test to determine if differences were statistically significant.

Evaluation of structural clustering using the selection cluster weighting z-score

The SCW z-score (41, 45–47, 120) was used to measure the nonrandomness of top covariation predictions mapped to the protein structure. As described in the previous section, residues were only considered when they could be aligned between the query sequence and target structure. We then converted the covariation rankings into single residue rankings. Each residue was ranked based on the score of the best covariation pair it forms between other residues. SCW z-scores were calculated for the top 30% of residues, a cutoff which has been used in previous studies (45, 46) and has been shown to correspond with the upper limit of clustering significance in previous studies (47). The SCW z-score can be computed using the code distributed with this study or using the

PyETViewer plugin for PyMol (121). It is described by the equation:

$$w = \sum_{i < j}^L S(i)S(j)A(i,j)b(i,j) \quad (7)$$

where L is the full set of pairs of residues present in a protein (counted only once per pair as specified by the term $i < j$), S is a selection function and returns one for a given residue (i or j) if that residue is in the set of pairs described by the 30% coverage cutoff, and A is an adjacency matrix for all residues in the structure where position i, j is one if the shortest distance between atoms of the two residues is < 4 Å and 0 otherwise. The term $b(i, j)$ is the bias coefficient, for unbiased analyses $b(i, j)$ evaluates to one for all pairs of residues, while for biased analyses $b(i, j)$ evaluates to the sequence separation between the two residues (*i.e.*, $|i - j|$). Differences between methods over all the proteins in the Pfam dataset were measured using paired two-sided Wilcoxon Rank Sum test.

Determination of average sequence separation

To determine the difference in pair biases for each of the covariation methods, the average sequence separation between residues predicted among the top-ranked pairs was determined. Top pairs were selected for evaluation until 30% of the residues in the structure was included. The sequence separation between the residues of the pair was calculated, *i.e.*, if a pair consists of residues i and j , the sequence separation is given by $|i - j|$. The average of all sequence separations in the set of top pairs for each protein was calculated, and methods were compared using paired two-sided Wilcoxon Rank Sum test to determine if there was a significant difference in the average sequence separations.

Recovery of functionally important residues

We obtained the coordinates for biological ligands from the BioLiP database (49) for each of our Pfam queries. A residue was considered as functionally important if the distance between any atom is smaller than 4 Å between that residue and the biological ligands. As with the calculation of SCW z-scores, we then converted the covariation rankings into single residue rankings. Each residue was ranked based on the score of the best covariation pair it forms between other residues. Single residue rankings were compared against the set of functional important residues using the AUROC and AUPRC. The AUPRCs were then adjusted by subtracting the observed positive rate, which is the expected AUPRC for a random predictor. The AUROCs and adjusted AUPRCs of each method were compared using paired two-sided Wilcoxon Rank Sum test to determine if differences were statistically significant.

Scoring of the identification of gold standard residues

Key conserved residues, motifs, functional sites, and other contributors to protein structure–function from the literature were identified for the RRM and WW domains and the D2R. The overlap of covariation predictions at different coverage cutoffs (10, 20, and 30% for the RRM and WW domains and 2.5, 5, 7.5, 10, 12, 14, 16, 20, and 30% for D2R) with these key sites as well as the union of all sites for each domain or protein were measured for significance using the one-sided hypergeometric test. For the WW domain, a set of known variable residues was also evaluated as a negative control.

Measuring the correlation of covariation predictions with experimental data

Single and double mutant data from large-scale mutagenesis screens of the RRM domain (50), the WW (84) domain, TEM-1 β -lactamase (85), the GB1 (86), and the prion-like domain of TDP-43 (87) were used to compute epistasis scores for these two domains. The log fitness scores ($\ln W$) of GB1 were obtained from Rollins *et al.* (122) and were exponentially transformed back to fitness before the epistasis scores calculation. The uncorrected toxicity scores for double mutants were used for TDP-43 (87). The toxicity scores for double and single mutants were also exponentially transformed back to have the *WT* toxicity score normalized to 1. Epistasis scores were computed based on the reported fitness values for single and double mutants using four different models of epistasis: Product, Additive, Log, and Min, the formulas for each of these models are provided below (84).

$$\epsilon_{ab}^{product} = M_{ab} - M_a \cdot M_b \quad (8)$$

$$\epsilon_{ab}^{additive} = (M_{ab} + WT) - (M_a + M_b) \quad (9)$$

$$\epsilon_{ab}^{log} = M_{ab} - \log_2((2^{M_a} - WT) \cdot (2^{M_b} - WT) + WT) \quad (10)$$

$$\epsilon_{ab}^{min} = M_{ab} - \min(M_a, M_b) \quad (11)$$

In each of these formulas, M_{ab} is the fitness value measured for double mutant ab , M_a is the fitness value of single mutant a , M_b is the fitness value measured for single mutant b , and WT is the fitness value measured for the wildtype domain. In these studies, the fitness values were normalized against wildtype fitness so WT is always 1.

Since the covariation metrics only provide a single score for each pair and the experimental studies provide many more, the mean of all epistasis scores for a given pair was taken. Similarly, covariation metrics evaluated here do not provide a sign to their prediction, so the absolute value of the mean calculated for each pair was taken. This means the covariation scores were compared to the average displacement from wildtype activity as measured by each epistasis model. Only pairs tested in the experimental studies were tested, and the

Pearson correlation coefficient between the raw score of each covariation method and each set of epistasis scores was computed to determine which method corresponded better with experimental observations. The negative raw score was used for CovET, because the CovET raw score has a different direction wherein a lower value means better covariation between a pair.

Data availability

The code for CovET is open and freely available on GitHub at: <https://github.com/LichtargeLab/Covariation-ET>.

Supporting information—This article contains supporting information (13,16,41,44,47,89,106–108,111,112,123–126).

Acknowledgments—The authors would like to thank Benu Atri and Angela Wilkins for further development of the ET-mIp method which led to this work, and Eunna Huh and David Marciano for their helpful discussions during the development of CovET.

Author contributions—D. M. K. and O. L. conceptualization; D. M. K. and O. L. methodology; D. M. K., S. H., and C. W. software; D. M. K., S. H., and C. W. validation; D. M. K., S. H., and C. W. formal analysis; D. M. K., S. H., C. W., and O. L. investigation; D. M. K., S. H., and C. W. writing—original draft; D. M. K., S. H., and C. W. visualization; M. A. A., T. G. W., and O. L. writing—review & editing; O. L. supervision; O. L. project administration; O. L. funding acquisition.

Funding and additional information—This work was supported by the National Institutes of Health (AG068214-01, AG061105, GM066099, and AG074009 to O. L.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of interest—The authors declare no conflict of interest with the contents of this article.

Abbreviations—The abbreviations used are: AUPRC, area under the precision recall curve; AUROC, area under the receiver operating curve; D2R, dopamine D2 receptor; ET, Evolutionary Trace; GB1, IgG-binding domain of protein G; GPCR, G protein-coupled receptor; RRM, RNA recognition motif; SCW, selection cluster weighting; TDP-43, TAR DNA-binding protein 43; TM3, transmembrane helix 3.

References

- Levinthal, C. (1969) How To Fold Graciously. In: DeBrunner, J. T. P., Munck, E., eds. *Mossbauer Spectroscopy in Biological Systems: Proceedings of a Meeting Held at Allerton House, Monticello, Illinois*, University of Illinois Press, Champaign, IL: 22–24
- Tüdös, E., Fiser, A., and Simon, I. (1994) Different sequence environments of amino acid residues involved and not involved in long-range interactions in proteins. *Int. J. Pept. Protein Res.* **43**, 205–208
- Dosztányi, Z., Fiser, A., and Simon, I. (1997) Stabilization centers in proteins: identification, characterization and predictions. *J. Mol. Biol.* **272**, 597–612
- Altschuh, D., Lesk, A. M., Bloomer, A. C., and Klug, A. (1987) Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J. Mol. Biol.* **193**, 693–707

CovET identifies protein structure–function modules

- Finkelstein, A. V., and Ptitsyn, O. B. (1987) Why do globular proteins fit the limited set of folding patterns? *Prog. Biophys. Mol. Biol.* **50**, 171–190
- Murzin, A. G., and Finkelstein, A. V. (1988) General architecture of the α -helical globule. *J. Mol. Biol.* **204**, 749–769
- Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., *et al.* (2020) Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589
- Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., *et al.* (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876
- Lockless, S. W., and Ranganathan, R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295–299
- Tsai, C.-J., and Nussinov, R. (2014) A unified view of “how allostery works”. *PLoS Comput. Biol.* **10**, e1003394
- Hatley, M. E., Lockless, S. W., Gibson, S. K., Gilman, A. G., and Ranganathan, R. (2003) Allosteric determinants in guanine nucleotide-binding proteins. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 14445–14450
- Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., *et al.* (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6**, e28766
- Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., *et al.* (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U. S. A.* **108**, E1293–E1301
- Terrón-Díaz, M. E., Wright, S. J., Agosto, M. A., Lichtarge, O., and Wensel, T. G. (2019) Residues and residue pairs of evolutionary importance differentially direct signaling bias of D2 dopamine receptors. *J. Biol. Chem.* **294**, 19279–19291
- Sung, Y.-M., Wilkins, A. D., Rodriguez, G. J., Wensel, T. G., and Lichtarge, O. (2016) Intramolecular allosteric communication in dopamine D2 receptor revealed by evolutionary amino acid covariation. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 3539–3544
- Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358
- Onrust, R., Herzmark, P., Chi, P., Garcia, P. D., Lichtarge, O., Kingsley, C., *et al.* (1997) Receptor and β binding sites in the α subunit of the retinal G protein transducin. *Science* **275**, 381–384
- Cushman, I., Bowman, B. R., Sowa, M. E., Lichtarge, O., Quijcho, F. A., and Moore, M. S. (2004) Computational and biochemical identification of a nuclear pore complex binding site on the nuclear transport carrier NTF2. *J. Mol. Biol.* **344**, 303–310
- Ribes-Zamora, A., Mihalek, I., Lichtarge, O., and Bertuch, A. A. (2007) Distinct faces of the Ku heterodimer mediate DNA repair and telomeric functions. *Nat. Struct. Mol. Biol.* **14**, 301–307
- Adikesavan, A. K., Katsonis, P., Marciano, D. C., Lua, R., Herman, C., and Lichtarge, O. (2011) Separation of recombination and SOS response in *Escherichia coli* RecA suggests LexA interaction sites. *PLoS Genet.* **7**, 1002244
- Raviscioni, M., He, Q., Salicru, E. M., Smith, C. L., and Lichtarge, O. (2006) Evolutionary identification of a subtype specific functional site in the ligand binding domain of steroid receptors. *Proteins* **64**, 1046–1057
- Yang, M., Wang, W., Zhong, M., Philippi, A., Lichtarge, O., and Sanborn, B. M. (2002) Lysine 270 in the third intracellular domain of the oxytocin receptor is an important determinant for $G\alpha_q$ coupling specificity. *Mol. Endocrinol.* **16**, 814–823
- Lin, C. Y., Varma, M. G., Joubel, A., Madabushi, S., Lichtarge, O., and Barber, D. L. (2003) Conserved motifs in somatostatin, D2-dopamine, and α 2B-adrenergic receptors for inhibiting the Na-H exchanger, NHE1. *J. Biol. Chem.* **278**, 15128–15135
- Shenoy, S. K., Drake, M. T., Nelson, C. D., Houtz, D. A., Xiao, K., Madabushi, S., *et al.* (2006) β -arrestin-dependent, G protein-independent ERK1/2 activation by the β 2 adrenergic receptor. *J. Biol. Chem.* **281**, 1261–1273
- Kobayashi, H., Ogawa, K., Yao, R., Lichtarge, O., and Bouvier, M. (2009) Functional rescue of β 1-adrenoceptor dimerization and trafficking by pharmacological chaperones. *Traffic* **10**, 1019–1033
- Bonde, M. M., Yao, R., Ma, J. N., Madabushi, S., Haunsø, S., Burstein, E. S., *et al.* (2010) An Angiotensin II type 1 receptor activation switch patch revealed through evolutionary trace analysis. *Biochem. Pharmacol.* **80**, 86–94
- Sowa, M. E., He, W., Wensel, T. G., and Lichtarge, O. (2000) A regulator of G protein signaling interaction surface linked to effector specificity. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 1483–1488
- Raviscioni, M., Gu, P., Sattar, M., Cooney, A. J., and Lichtarge, O. (2005) Correlated evolutionary pressure at interacting transcription factors and DNA response elements can guide the rational engineering of DNA binding specificity. *J. Mol. Biol.* **350**, 402–415
- Kristensen, D. M., Matthew, R. M., Lisewski, A., Erdin, S., Chen, B. Y., Fofanov, V. Y., *et al.* (2008) Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. *BMC Bioinformatics* **9**, 17
- Venner, E., Lisewski, A. M., Erdin, S., Matthew Ward, R., Amin, S. R., and Lichtarge, O. (2010) Accurate protein structure annotation through competitive diffusion of enzymatic functions over a network of local evolutionary similarities. *PLoS One* **5**, 14286
- Sowa, M. E., He, W., Slep, K. C., Kercher, M. A., Lichtarge, O., and Wensel, T. G. (2001) Prediction and confirmation of a site critical for effector regulation of RGS domain activity. *Nat. Struct. Biol.* **8**, 234–237
- Quan, X. J., Denayer, T., Yan, J., Jafar-Nejad, H., Philippi, A., Lichtarge, O., *et al.* (2004) Evolution of neural precursor selection: functional divergence of proneural proteins. *Development* **131**, 1679–1689
- Ward, R. M., Venner, E., Daines, B., Murray, S., Erdin, S., Kristensen, D. M., *et al.* (2009) Evolutionary trace annotation server: automated enzyme function prediction in protein structures using 3D templates. *Bioinformatics* **25**, 1426–1427
- Erdin, S., Ward, R. M., Venner, E., and Lichtarge, O. (2010) Evolutionary trace annotation of protein function in the structural proteome. *J. Mol. Biol.* **396**, 1451–1473
- Madabushi, S., Gross, A. K., Philippi, A., Meng, E. C., Wensel, T. G., and Lichtarge, O. (2004) Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions. *J. Biol. Chem.* **279**, 8126–8132
- Gu, P., Morgan, D. H., Sattar, M., Xu, X., Wagner, R., Raviscioni, M., *et al.* (2005) Evolutionary trace-based peptides identify a novel asymmetric interaction that mediates oligomerization in nuclear receptors. *J. Biol. Chem.* **280**, 31818–31829
- Baameur, F., Morgan, D. H., Yao, H., Tran, T. M., Hammit, R. A., Sabui, S., *et al.* (2010) Role for the regulator of G-protein signaling homology domain of G protein-coupled receptor kinases 5 and 6 in β 2-adrenergic receptor and rhodopsin phosphorylation. *Mol. Pharmacol.* **77**, 405–415
- Item, C. B., Mihalek, I., Lichtarge, O., Jalan, A., Vodopiutz, J., Muhl, A., *et al.* (2007) Manifestation of hawkinsinuria in a patient compound heterozygous for hawkinsinuria and tyrosinemia III. *Mol. Genet. Metab.* **91**, 379–383
- Shaibani, A., Shchelochkov, O. A., Zhang, S., Katsonis, P., Lichtarge, O., Wong, L. J., *et al.* (2009) Mitochondrial neurogastrointestinal encephalopathy due to mutations in RRM2B. *Arch. Neurol.* **66**, 1028–1032
- Mihalek, I., Res, I., and Lichtarge, O. (2004) A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.* **336**, 1265–1282
- Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419
- Hopf, T. A., Green, A. G., Schubert, B., Mersmann, S., Schärfe, C. P. I., Ingraham, J. B., *et al.* (2019) The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* **35**, 1582–1584
- Baldassi, C., Zamparo, M., Feinauer, C., Procaccini, A., Zecchina, R., Weigt, M., *et al.* (2014) Fast and accurate multivariate Gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. *PLoS One* **9**, e9272

45. Wilkins, A. D., Lua, R., Erdin, S., Ward, R. M., and Lichtarge, O. (2010) Sequence and structure continuity of evolutionary importance improves protein functional site discovery and annotation. *Protein Sci.* **19**, 1296–1311
46. Wilkins, A. D., Venner, E., Marciano, D. C., Erdin, S., Atri, B., Lua, R. C., *et al.* (2013) Accounting for epistatic interactions improves the functional analysis of protein structures. *Bioinformatics* **29**, 2714–2721
47. Madabushi, S., Yao, H., Marsh, M., Kristensen, D. M., Philippi, A., Sowa, M. E., *et al.* (2002) Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.* **316**, 139–154
48. Wang, C., Konecki, D. M., Marciano, D. C., Govindarajan, H., Williams, A. M., Wastuwidyaningtyas, B., *et al.* (2021) Identification of evolutionarily stable functional and immunogenic sites across the SARS-CoV-2 proteome and greater coronavirus family. *Bioinformatics* **37**, 4033–4040
49. Yang, J., Roy, A., and Zhang, Y. (2013) BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.* **41**, D1096–D1103
50. Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R., and Fields, S. (2013) Deep mutational scanning of an RRM domain of the *Saccharomyces cerevisiae* poly(A)-binding protein. *RNA* **19**, 1537–1551
51. Mangus, D. A., Evans, M. C., and Jacobson, A. (2003) Poly(A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression. *Genome Biol.* **4**, 223
52. Imataka, H., Gradi, A., and Sonenberg, N. (1998) A newly identified N-terminal amino acid sequence of human eIF4G binds poly(A)-binding protein and functions in poly(A)-dependent translation. *EMBO J.* **17**, 7480–7489
53. Maris, C., Dominguez, C., and Allain, F. H.-T. (2005) The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J.* **272**, 2118–2131
54. Dreyfuss, G., Swanson, M. S., and Piñol-Roma, S. (1988) Heterogeneous nuclear ribonucleoprotein particles and the pathway of mRNA formation. *Trends Biochem. Sci.* **13**, 86–91
55. Muto, Y., and Yokoyama, S. (2012) Structural insight into RNA recognition motifs: versatile molecular Lego building blocks for biological systems. *Wiley Interdiscip. Rev. RNA* **3**, 229–246
56. Sachs, A. B., Davis, R. W., and Kornberg, R. D. (1987) A single domain of yeast poly(A)-binding protein is necessary and sufficient for RNA binding and cell viability. *Mol. Cell. Biol.* **7**, 3268–3276
57. Kühn, U., Gündel, M., Knoth, A., Kerwitz, Y., Rüdell, S., and Wahle, E. (2009) Poly(A) tail length is controlled by the nuclear Poly(A)-binding protein regulating the interaction between Poly(A) polymerase and the cleavage and polyadenylation specificity factor. *J. Biol. Chem.* **284**, 22803–22814
58. Lunde, B. M., Moore, C., and Varani, G. (2007) RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.* **8**, 479–490
59. Koepf, E. K., Petrassi, H. M., Sudol, M., and Kelly, J. W. (2008) WW: an isolated three-stranded antiparallel β -sheet domain that unfolds and refolds reversibly; evidence for a structured hydrophobic cluster in urea and GdnHCl and a disordered thermal unfolded state. *Protein Sci.* **8**, 841–853
60. Sudol, M., Chen, H. I., Bougeret, C., Einbond, A., and Bork, P. (1995) Characterization of a novel protein-binding module - the WW domain. *FEBS Lett.* **369**, 67–71
61. Fowler, D. M., Araya, C. L., Fleishman, S. J., Kellogg, E. H., Stephany, J. J., Baker, D., *et al.* (2010) High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7**, 741–746
62. Jäger, M., Dendle, M., and Kelly, J. W. (2009) Sequence determinants of thermodynamic stability in a WW domain - An all- β -sheet protein. *Protein Sci.* **18**, 1806–1813
63. Yanagida, H., Matsuura, T., and Yomo, T. (2008) Compensatory evolution of a WW domain variant lacking the strictly conserved Trp residue. *J. Mol. Evol.* **66**, 61–71
64. Jiang, X., Kowalski, J., and Kelly, J. W. (2008) Increasing protein stability using a rational approach combining sequence homology and structural alignment: stabilizing the WW domain. *Protein Sci.* **10**, 1454–1465
65. Toepert, F., Pires, J. R., Landgraf, C., Oschkinat, H., and Schneider-Mergener, J. (2001) Synthesis of an Array Comprising 837 Variants of the hYAP WW Protein Domain. *Angew. Chem. Int. Ed. Engl.* **40**, 897–900
66. Pires, J. R., Taha-Nejad, F., Toepert, F., Ast, T., Hoffmüller, U., Schneider-Mergener, J., *et al.* (2001) Solution structures of the YAP65 WW domain and the variant L30 K in complex with the peptides GTPPPPYTVG, N-(n-octyl)-GPPPY and PLPPY and the application of peptide libraries reveal a minimal binding epitope. *J. Mol. Biol.* **314**, 1147–1156
67. Baldessari, F., Capelli, R., Carloni, P., and Giorgetti, A. (2020) Coevolutionary data-based interaction networks approach highlighting key residues across protein families: the case of the G-protein coupled receptors. *Comput. Struct. Biotechnol. J.* **18**, 1153–1159
68. Filipek, S. (2019) Molecular switches in GPCRs. *Curr. Opin. Struct. Biol.* **55**, 114–120
69. Venkatakrishnan, A. J., Deupi, X., Lebon, G., Tate, C. G., Schertler, G. F., and Babu, M. M. (2013) Molecular signatures of G-protein-coupled receptors. *Nature* **494**, 185–194
70. Hauser, A. S., Kooistra, A. J., Munk, C., Babu, M. M., Bouvier, M., Gloriam, D. E., *et al.* (2021) GPCR activation mechanisms across classes and macro/microscales. *Nat. Struct. Mol. Biol.* **28**, 879–888
71. Wingler, L. M., and Lefkowitz, R. J. (2020) Conformational basis of G protein-coupled receptor signaling versatility. *Trends Cell Biol.* **30**, 736–747
72. Tehan, B. G., Bortolato, A., Blaney, F. E., Weir, M. P., and Mason, J. S. (2014) Unifying Family A GPCR Theories of Activation. *Pharmacol. Ther.* **143**, 51–60
73. Audet, M., and Bouvier, M. (2012) Restructuring G-protein-coupled receptor activation. *Cell* **151**, 14–23
74. Rodriguez, G. J., Yao, R., Lichtarge, O., and Wensel, T. G. (2010) Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 7787–7792
75. Snyder, S. H., Taylor, K. M., Coyle, J. T., and Meyerhoff, J. L. (1970) The role of brain dopamine in behavioral regulation and the actions of psychotropic drugs. *Am. J. Psychiatry* **127**, 199–207
76. Roth, B. L., Sheffler, D. J., and Kroeze, W. K. (2004) Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discov.* **3**, 353–359
77. Comai, S., Tau, M., Pavlovic, Z., and Gobbi, G. (2012) The psychopharmacology of aggressive behavior. *J. Clin. Psychopharmacol.* **32**, 237–260
78. Zhou, Q., Yang, D., Wu, M., Guo, Y., Guo, W., Zhong, L., *et al.* (2019) Common activation mechanism of class A GPCRs. *Elife* **8**, e50279
79. Javitch, J. A., Fu, D., Chen, J., and Karlin, A. (1995) Mapping the binding-site crevice of the dopamine D2 receptor by the substituted-cysteine accessibility method. *Neuron* **14**, 825–831
80. Lan, H., DuRand, C. J., Teeter, M. M., and Neve, K. A. (2006) Structural determinants of pharmacological specificity between D1 and D2 dopamine receptors. *Mol. Pharmacol.* **69**, 185–194
81. Simpson, M. M., Ballesteros, J. A., Chiappa, V., Chen, J., Suehiro, M., Hartman, D. S., *et al.* (1999) Dopamine D4/D2 receptor selectivity is determined by a divergent aromatic microdomain contained within the second, third, and seventh membrane-spanning segments. *Mol. Pharmacol.* **56**, 1116–1126
82. Kalani, M. Y. S., Vaidehi, N., Hall, S. E., Trabanino, R. J., Freddolino, P. L., Kalani, M. A., *et al.* (2004) The predicted 3D structure of the human D2 dopamine receptor and the binding site and binding affinities for agonists and antagonists. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 3815–3820
83. Visiers, I., Ballesteros, J. A., and Weinstein, H. (2002) Three-dimensional representations of G protein-coupled receptor structures and mechanisms. *Methods Enzymol.* **343**, 329–371
84. Araya, C. L., Fowler, D. M., Chen, W., Muniez, I., Kelly, J. W., and Fields, S. (2012) A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 16858–16863

CovET identifies protein structure–function modules

85. Gonzalez, C. E., and Ostermeier, M. (2019) Pervasive pairwise intragenic epistasis among sequential mutations in TEM-1 β -lactamase. *J. Mol. Biol.* **431**, 1981–1992
86. Olson, C. A., Wu, N. C., and Sun, R. (2014) A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr. Biol.* **24**, 2643–2651
87. Bolognesi, B., Faure, A. J., Seuma, M., Schmiedel, J. M., Tartaglia, G. G., and Lehner, B. (2019) The mutational landscape of a prion-like domain. *Nat. Commun.* **10**, 4162
88. Huh, E., Agosto, M. A., Wensel, T. G., and Lichtarge, O. (2023) Co-evolutionary signals in metabotropic glutamate receptors capture residue contacts and long-range functional interactions. *J. Biol. Chem.* **299**, 103030
89. Lichtarge, O., Sowa, M. E., and Philippi, A. (2002) Evolutionary traces of functional surfaces along G protein signaling pathway. *Methods Enzymol.* **344**, 536–556
90. Lichtarge, O., and Sowa, M. E. (2002) Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.* **12**, 21–27
91. Lichtarge, O., Yao, H., Kristensen, D. M., Madabushi, S., and Mihalek, I. (2003) Accurate and scalable identification of functional sites by evolutionary tracing. *J. Struct. Funct. Genomics* **4**, 159–166
92. Talavera, D., Lovell, S. C., and Whelan, S. (2015) Covariation is a poor measure of molecular coevolution. *Mol. Biol. Evol.* **32**, 2456–2468
93. Fodor, A. A., and Aldrich, R. W. (2004) Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins* **56**, 211–221
94. Katsonis, P., and Lichtarge, O. (2014) A formal perturbation equation between genotype and phenotype determines the evolutionary action of protein-coding variations on fitness. *Genome Res.* **24**, 2050–2058
95. Katsonis, P., and Lichtarge, O. (2019) CAGI5: objective performance assessments of predictions based on the evolutionary action equation. *Hum. Mutat.* **40**, 1436–1454
96. Osman, A. A., Neskey, D. M., Katsonis, P., Patel, A. A., Ward, A. M., Hsu, T.-K., et al. (2015) Evolutionary action score of TP53 coding variants is predictive of platinum response in head and neck cancer patients. *Cancer Res.* **75**, 1205–1215
97. Neskey, D. M., Osman, A. A., Ow, T. J., Katsonis, P., McDonald, T., Hicks, S. C., et al. (2015) Evolutionary action score of TP53 identifies high-risk mutations associated with decreased survival and increased distant metastases in head and neck cancer. *Cancer Res.* **75**, 1527–1536
98. Chun, Y. S., Passot, G., Yamashita, S., Nusrat, M., Katsonis, P., Loree, J. M., et al. (2019) Deleterious effect of RAS and evolutionary high-risk TP53 double mutation in colorectal liver metastases. *Ann. Surg.* **269**, 917–923
99. Koire, A., Katsonis, P., Kim, Y. W., Buchovecky, C., Wilson, S. J., and Lichtarge, O. (2021) A method to delineate de novo missense variants across pathways prioritizes genes linked to autism. *Sci. Transl. Med.* **13**, eabc1739
100. Kim, Y. W., Al-Ramahi, I., Koire, A., Wilson, S. J., Konecki, D. M., Mota, S., et al. (2021) Harnessing the paradoxical phenotypes of APOE ϵ 2 and APOE ϵ 4 to identify genetic modifiers in Alzheimer's disease. *Alzheimers Dement.* **17**, 831–846
101. Marciano, D. C., Wang, C., Hsu, T.-K., Bourquard, T., Atri, B., Nehring, R. B., et al. (2022) Evolutionary action of mutations reveals antimicrobial resistance genes in *Escherichia coli*. *Nat. Commun.* **13**, 3189
102. Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J., and Söding, J. (2019) HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473
103. Schaarschmidt, J., Monastyrskyy, B., Kryshtafovych, A., and Bonvin, A. M. J. J. (2018) Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age. *Proteins* **86**, 51–66
104. Shrestha, R., Fajardo, E., Gil, N., Fidelis, K., Kryshtafovych, A., Monastyrskyy, B., et al. (2019) Assessing the accuracy of contact predictions in CASP13. *Proteins* **87**, 1058–1068
105. Cléry, A., Blatter, M., and Allain, F. H. T. (2008) RNA recognition motifs: boring? not quite. *Curr. Opin. Struct. Biol.* **18**, 290–298
106. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009) BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421
107. Bateman, A. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515
108. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., Mcgettigan, P. A., McWilliam, H., et al. (2007) Clustal W and clustal X version 2.0. *Bioinformatics* **23**, 2947–2948
109. Lua, R. C., Wilson, S. J., Konecki, D. M., Wilkins, A. D., Venner, E., Morgan, D. H., et al. (2016) UET: a database of evolutionarily-predicted functional determinants of protein sequences that cluster as functional sites in protein structures. *Nucleic Acids Res.* **44**, D308–D312
110. Sokal, R. R., and Michener, C. D. (1958) A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.* **38**, 1409–1438
111. Dunn, S. D., Wahl, L. M., and Gloor, G. B. (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **24**, 333–340
112. Martin, L. C., Gloor, G. B., Dunn, S. D., and Wahl, L. M. (2005) Using information theory to search for co-evolving residues in proteins. *Bioinformatics* **21**, 4116–4124
113. Graña, O., Baker, D., MacCallum, R. M., Meiler, J., Punta, M., Rost, B., et al. (2005) CASP6 assessment of contact prediction. *Proteins* **61**, 214–224
114. Izarzugaza, J. M. G., Graña, O., Tress, M. L., Valencia, A., and Clarke, N. D. (2007) Assessment of intramolecular contact predictions for CASP7. *Proteins* **69**, 152–158
115. Ezkurdia, L., Grana, O., Izarzugaza, J. M. G., and Tress, M. L. (2009) Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. *Proteins* **77**, 196–209
116. Monastyrskyy, B., Fidelis, K., Tramontano, A., and Kryshtafovych, A. (2011) Evaluation of residue-residue contact predictions in CASP9. *Proteins* **79**, 119–125
117. Monastyrskyy, B., D'Andrea, D., Fidelis, K., Tramontano, A., and Kryshtafovych, A. (2016) New encouraging developments in contact prediction: assessment of the CASP11 results. *Proteins* **84**, 131–144
118. Monastyrskyy, B., D'Andrea, D., Fidelis, K., Tramontano, A., and Kryshtafovych, A. (2014) Evaluation of residue-residue contact prediction in CASP10. *Proteins* **82**, 138–153
119. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011) Scikit-learn: machine learning in python. *J. Machine Learn. Res.* **12**, 2825–2830
120. Mihalek, I., Reš, I., and Lichtarge, O. (2007) Background frequencies for residue variability estimates: BLOSUM revisited. *BMC Bioinformatics* **8**, 488
121. Lua, R. C., and Lichtarge, O. (2010) PyETV: a PyMOL evolutionary trace viewer to analyze functional site predictions in protein complexes. *Bioinformatics* **26**, 2981–2982
122. Rollins, N. J., Brock, K. P., Poelwijk, F. J., Stiffler, M. A., Gauthier, N. P., Sander, C., et al. (2019) Inferring protein 3D structure from deep mutation scans. *Nat. Genet.* **51**, 1170–1176
123. Hopf, T. A., Colwell, L. J., Sheridan, R., Rost, B., Sander, C., and Marks, D. S. (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* **149**, 1607–1621
124. Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P. I., Springer, M., Sander, C., et al. (2017) Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135
125. Hopf, T. A., Schärfe, C. P. I., Rodrigues, J. P. G. L. M., Green, A. G., Kohlbacher, O., Sander, C., et al. (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* **3**, e03430
126. Marks, D. S., Hopf, T. A., and Sander, C. (2012) Protein structure prediction from sequence variation. *Nat. Biotechnol.* **30**, 1072–1080