eLife

Reviewed Preprint
v1 • April 9, 2026
Not revised

For correspondence:
cocco@phys.ens.fr

† These authors contributed equally.

¶ Co-last authors.

Competing interests: No competing interests declared

Funding: See page 21

Reviewing editor: Qiang Cui, Boston University, United States

Design and experimental characterization of specificity-switching mutational paths of WW domains

Ahmed Rehan[*,†], **Eugenio Mauri**[‡,†], **Jorge Fernandez-de-Cossio-Diaz**[‡,§,§], **Pierre-Guillaume Brun**[‡], **Remi Monasson**[‡], **Marco Ribezzi-Crivellari**[*,¶], **Simona Cocco**[‡,¶]

[*]École Supérieure de Physique et de Chimie Industrielles-ESPCI Laboratoire de Biochimie (LBC), Paris, France • [‡]Laboratory of Physics of the Ecole Normale Supérieure, CNRS UMR 8023 and PSL Research, Sorbonne Université, Paris, France • [§]Université Paris-Saclay, CNRS, CEA, Institut de Physique Théorique, Gif-sur-Yvette, France

eLife Assessment

In this **important** study, the authors demonstrate that generative AI techniques (restricted Boltzmann machine) can be used effectively to design and characterize mutational pathways of WW domains with different binding specificities. The computational studies are complemented by experimental validations, and the results provide **solid** evidence supporting the idea that sequence landscape holds significance in understanding protein evolution from a transition path perspective. The minor weakness of the study in the current form concerns limited success in designing variants with smoothly varying binding specificities. Nevertheless, the work will likely have a major impact on research aimed at understanding how evolution navigates fitness landscapes as well as reconstructing ancestral sequences.

https://doi.org/10.7554/eLife.110491.1.sa2

Abstract

Specific interactions between proteins and other biomolecules are ubiquitous in cellular processes. How specificity is encoded in the protein sequence and can be modified through a minimal set of concerted mutations is a complex issue. In this work, we focus on the WW protein domain, whose variants specifically bind to different classes of proline-rich peptides. Combining unsupervised learning of homologous WW sequence data with Restricted Boltzmann Machines (RBM) and path-sampling methods, we design mutational paths of putative WW domains interpolating between two natural WW domains with either distinct or similar specificities. Sequences along the designed paths are then experimentally validated with high-throughput in-vitro binding assays against 3 peptides of different classes. The vast majority (93%) of intermediate sequences along the designed paths are responsive to the initial or/and final peptides. On the contrary, domains along scrambled paths, in which the same mutations are introduced in random order are not functional, emphasizing how successful design crucially depends on the ability to model epistatic interactions. Interestingly, switch in specificity between classes I and IV whose representative peptides bind to different pockets on the WW domain appears to be smooth, with intermediates displaying some level of binding cross-reactivity with all tested peptides. We finally show that the RBM paths share a high identity with internal nodes obtained from ancestral sequence reconstruction based on the seed WW domains.

1 Introduction

As living organisms evolve, the proteins they produce may need to adapt and specialize to a new task while preserving their fundamental biological function, such as catalyzing certain classes of reactions or binding to specific targets [1]. An illustration of this flexibility is provided by hemoglobin and myoglobin: these proteins share a common ancestor and the conserved mechanism of binding oxygen to the heme group, however hemoglobin transports oxygen from the lungs to the tissues, and myoglobin stores oxygen for use during muscle contraction [2]. Another interesting example is the WW domain, a family of small protein domains (30-40 amino acids long), whose name comes from the presence of two strongly conserved tryptophan residues [3]. Found in proteins across multiple species, WW domains are essential to key cellular pathways, such as the Hippo pathway, which controls cell proliferation, apoptosis, and organ size [4]. WW domains are widely distributed among proteins and are commonly found both in the cytoplasm and in the nucleus [5]. Following their initial identification, these domains became a primary focus of scientific inquiry due to their involvement in signaling complexes linked to a range of human disorders, including Alzheimer's disease, Liddle's syndrome, muscular dystrophy, Huntington's disease, and some types of cancers [6].

Most WW domains bind proline-rich peptides [7]. However, they can differ significantly in the specific amino-acid motifs they recognize, defining three broad specificity classes (I, II/III, IV) [8, 6, 9], which determine the biological pathways their interactions affect. Two examples of WW domains containing proteins present in the human proteome are the yes-associated protein 1 (YAP1) and the Peptidyl-prolyl cis-trans isomerase NIMA interacting 1 (Pin1). While the former acts as a transcriptional co-activator of genes involved in cellular proliferation (with a major role in oncogenic activity [10]), the latter isomerizes only phospho-Serine/Threonine-Proline motifs and has many roles, among which cell cycle regulation.

The involvement of mutations or dysregulation of WW domain-containing proteins in the pathogenesis of various diseases, including cancer, muscular dystrophy, and cardiovascular disorders, has been well documented. For instance, aberrant activity of the YAP/TAZ transcriptional regulators, which is mediated by their WW domains, is a hallmark of many cancers due to its role in promoting unchecked cell growth [11]. Similarly, mutations in the WW domain of Pin1 are involved in the emergence of Alzheimer's disease [12], while NEDD4-2 are associated with Liddle syndrome, which disrupts sodium channel regulation and is a risk factor for hypertension [13]. These examples illustrate the significance of WW domains in health and disease, establishing them as crucial targets for therapeutic intervention. Rational design of WW domains has been proposed to change their specificity and recognize mutated ligands. As an example, in Liddle's syndrome, the PPXY motif of a sodium channel subunit, β EnaC is mutated, which prevents sodium channel degradation by Nedd4; changing specificity of the WW domain from class I to II/III would restore recognition [14].

In addition to their clinical relevance, WW domains serve as valuable models in structural biology and protein evolution due to their small size, conserved structure, and diverse binding partners. Their modular nature and capacity to mediate specific interactions provide insights into protein network dynamics and evolutionary adaptation [15, 8, 6, 9, 16]. In this context, a natural question is to understand how homologous proteins with different ligand specificities evolved from ancestral, and possibly promiscuous proteins [17], while remaining functional at all stages. This question dates back to the seminal works of J. Maynard Smith, who described evolution as a word game in which one can change one letter at a time while keeping meaningful intermediate words, as illustrated by the path "word→wore→gore→gone→gene" [18]. Ancestral sequence reconstruction, by constructing putative intermediates linking the diverse modern proteins, is a practical way to address this problem [19]. However, bridging the gap between the theoretical and conceptual study of evolution through paths [20] and realistic fitness landscapes is difficult [21, 22].

From an experimental point of view, high-throughput experiments allow for building mutational paths between phenotypically distinct proteins separated by D mutations [22, 23, 21, 24, 25], when D is small enough. Examples include Ref. [25], in which mutational paths in a region of 4 residues increasing antibiotic resistance in β-lactamase were investigated, and Ref. [21], where all 2^{13} direct paths involving 13 residues allowing to change the fluorescence properties of a protein, were tested. Let us also mention Ref. [26] where all 20^4 paths over a 4-residue region in the protein GB1 were characterized and Ref. [27], where about 1000 paths in the fitness landscape of transcription factors for short (9 nucleotides) DNA sequences were tested and characterized. When the number D of mutations separating the two proteins becomes large, the space of possible sequences cannot be explored in an exhaustive manner any longer. Evolutionary methods such as ancestral protein reconstruction (ASR) can help to isolate sets of important mutations involved in the switch of affinity in protein systems [28, 29], although the phenotypic characterization of the mutations remains essentially experimental. To this end, computational models approximating the genotype-phenotype mapping are necessary to prune this huge space and retain only good putative paths leading to the change of function.

Recently, we proposed such a computational approach to design mutational paths [30]. Briefly speaking, the approach relies on two ingredients. First, we use an unsupervised machine-learning architecture, trained from homologous sequence data, to model the global fitness landscape of the protein family of interest, hereafter WW domains. Various generative models were recently proposed based on deep architectures [31, 32], e.g. diffusion processes [33]. Simpler unsupervised architectures such as Boltzmann Machines (BM) or Variational Auto-Encoders have been shown to be good generative models, with the ability to design novel proteins with functionalities comparable to natural ones [16, 9, 34, 35, 36], while being easier to interpret. Hereafter, we use Restricted Boltzmann Machines (RBM), which are bipartite probabilistic graphical models learning relevant amino-acid motifs with key role for structure stability and function [37] and able to design new functional proteins [38] or other biomolecules [39, 40]. Second, we sample the space of mutational paths, defined as chains of sequences that (1) have high scores according to the RBM generative model and (2) differ from the previous and next ones along the path by a small (and controllable) number of mutations. At their extremities the chains are anchored by two sequences of interest, hereafter two natural WW domains with distinct specificities.

This path design approach was shown to successfully generate high-quality mutational paths of in silico (lattice-based) proteins, for which the ground-truth was known [30]. In addition, we showed it could be applied to the WW domain family, using as training data the multiple-sequence alignment from the PFAM/InterPro database (PFAM ID: PF00397). Intermediate configurations had high scores according to the trained model and to computational methods based on protein structure information, such as AlphaFold and ProteinMPNN [41, 42].

In this paper, we go a step further in designing mutational paths for switching specificity in WW domains. We present a high-throughput fully in-vitro binding assay that allows us to simultaneously evaluate the activity of the designed WW domain paths against multiple peptides attached to different specificity classes. These experimental results not only validate the computational approach but also allow us to study in great details how the switching of specificity takes place along the path. Lastly, we show that the RBM designed paths share some similarity with the internal nodes of phylogenetic trees inferred from representative WW domains sequences (the PFAM seed).

2 Computational models for path generation and experimental verification

Our pipeline is described in Figure 1. We consider the space of putative WW domains (Figure 1A), whose sequences contain a variable portion of $L = 31$ (Figure 1B) amino acids plus appropriate constant flanking regions or 'spacers' (Figure 1D). The probabilistic model for the variable part is provided by a two-layer unsupervised neural network, called Restricted Boltzmann Machine (RBM), trained on homologous natural sequence data, see Figures 1B,C

(Methods 5.1 and [43, 30]). RBM learns a distribution over the set of all possible sequences, which allows us to score putative WW sequences, and, in turn, to generate new sequences with high scores.

Natural WW domains are broadly classified according to three specificity classes (I, II/III, IV) [8, 6, 9]. We choose peptides representative of each class to assess the binding specificity of the designed variants (Table 1). Some latent variables inferred by the RBM have been shown to be informative about the speci-ficity classes [43, 30]. More precisely, high-dimensional sequences cluster in the 2-dimensional plane defined by the inputs attached to two particular latent units, I_1 and I_2 (see Ref. [43], Methods Section and Figure 2A). Experimentally tested sequences reported in the literature appear to consistently belong to the same clusters depending on their specificities. This clustering property is sketched in Figure 1C.

We learn global RBM models, trained on all available WW data irrespectively of their binding specificity, or class-specific models, trained from sequences attached to one class only, see Methods 5.1.1; notice that the attribution of sequences to one of the three functional classes for training is mostly based on the 2-dimensional clustering mentioned above and not on experimental binding experiments, and is therefore subject to errors. Specific models are, by construction, informative about the binding specificity of the WW sequences but are generally of lesser quality than the global model, due to the smaller number of available annotated WW data, in particular for class IV. As a consequence, they cannot be used to generate de novo high-quality sequences from one class or another, but rather to score the specificity of sequences generated from the global model.

We use path-sampling algorithms, developed in [30], to design paths of sequences fulfilling the following conditions:

- the initial and final sequences are natural WW domains, possibly with different specificities, e.g. triangle and diamond in the sketch of Figure 1, and identity (fraction of identical residues) ranging from 30% to 50%, which are typical values within a homologous family.
- Intermediate sequences along the path have high scores (probabilities) according to the RBM global model. Amino acids in the intermediate sequences can take any value.
- The number of mutations from one sequence to the next along the path is small, typically one or two. The length of the designed path is typically of 1.5 times the hamming distance between the initial and final sequences.

To sample paths we have employed a Monte Carlo algorithm, which randomly draws paths according to the RBM probability distribution (Methods 5.2). We have designed four batches of sequences, listed in Appendix 8.1 together with the parameters used for the models. Intermediate sequences along the paths are scored according to the global RBM model, and to class-specific models to predict their specificities. In our sampled paths amino acids can “transiently” mutate into third-party values (underlined in blue in Figure 2A). Transient mutations can yield higher fitness scores along the evolutionary path, particularly when epistatic interactions are strong and compensatory mutations are expected [44].

To test the sequences generated by the model in high-throughput, we developed a fully *in-vitro*, rapid and accurate assay: “fluorescence-Activated Bead Counting (fABC)”. The assay assesses interactions between proteins and target molecules immobilized on magnetic beads, see Figure 1D. In the present case our targets are proline-rich peptides specific of the three different subclasses of WW domains (I, II/III, IV). Briefly, synthetic genes coding for SNAP-tagged WW domains[45] are expressed using cell-free *in-vitro* Transcription and Translation (IVTT) [46]. The SNAP-tag is leveraged for the fluorescent labelling of the WW domains using a benzyl-guanine modified fluorescent dye. Fluorescent binders are then mixed with magnetic beads coated with the target peptide and allowed to equilibrate. Binding activity is then measured by quantifying bead fluorescence on a flow cytometry instrument[47]. We demonstrate how, by using spectrally distinct fluorescently-labelled DNA linkers to coat the magnetic beads, we are able to multiplex different targets within a single run, increasing speed and reducing the cost of the assay. This method provides a robust and versatile platform for studying protein/protein or protein/peptide interactions and sequence/function relationship.

Figure 1. Experimental and computational pipeline for designing WW domains of variable specificities.

A. Sketched of the activity vs. sequence landscape of WW domains. Highly active domains may bind to different peptides (black shapes) and are shown with different colours. **B**. Homologous sequences define the training data for our restricted Boltzmann machine (RBM) model. The binding specificities of some sequences are annotated, while others are not available. **C**. After training, the latent variables of the RBM define low-dimensional projections that identify clusters of sequences sharing the same specificity. Sequences along paths may cross regions deprived of natural sequences, of unknown specificity. **D**. Sketch of the experiment. In vitro transcription and translation of our construct result in the expression of a fusion protein including a WW domain (green), a linker (yellow), and the SNAP tag (blue), labelled with BG-AF647 dye. The WW domain of the fusion protein may bind to peptide-coated magnetic beads (brown). The fluorescence intensity on the bead surfaces is then measured using flow cytometry, assessing the strength of the interaction between the labelled protein and its target peptide.

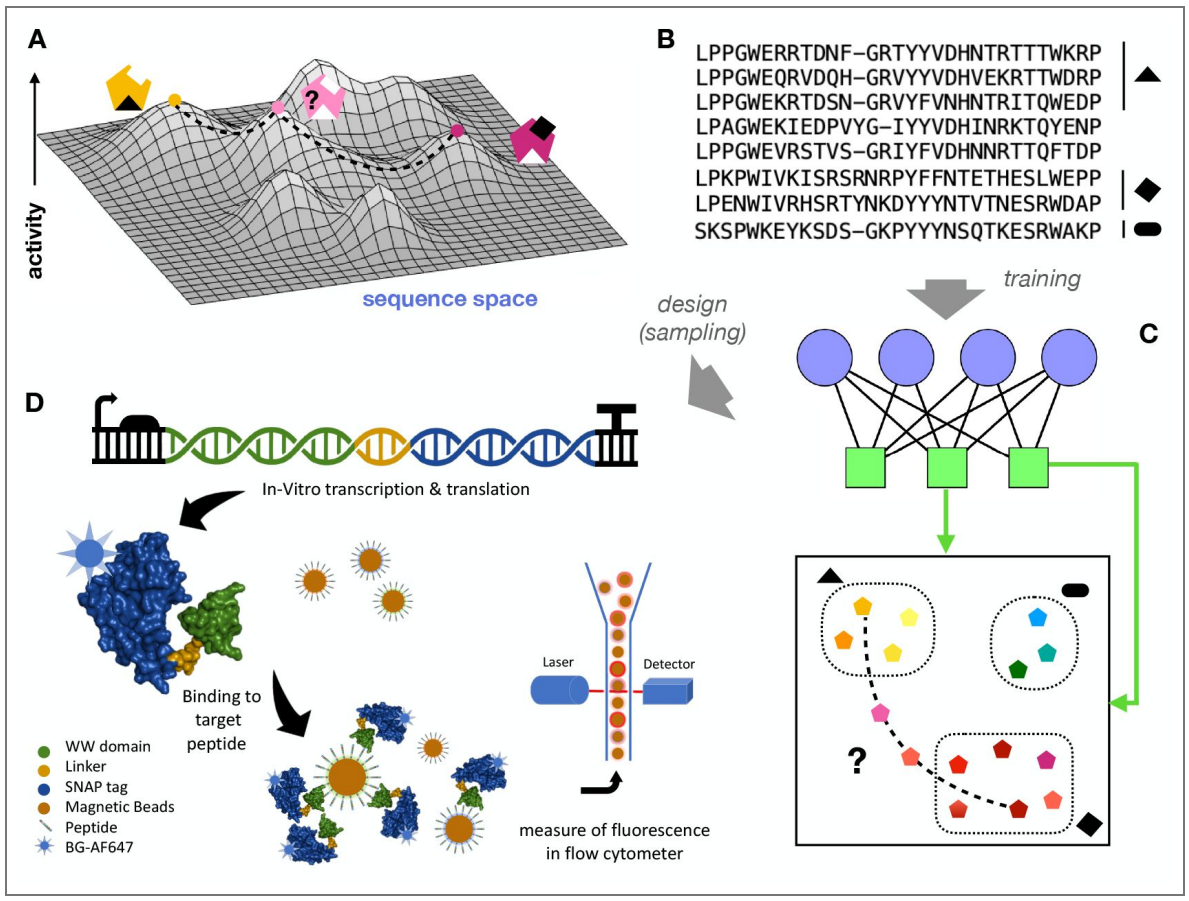


Table 1. Peptides used for the experimental validation of mutational paths in WW domains.

Highlighted in red are residues that bind to the WW domain. The phosphorylated Threonine in the peptide of class IV is indicated by pT.

Class	Sequence
I	GESPPPPYSRYPMD
II	APPTPPPLPPD
IV	EQPLpTPVTDL

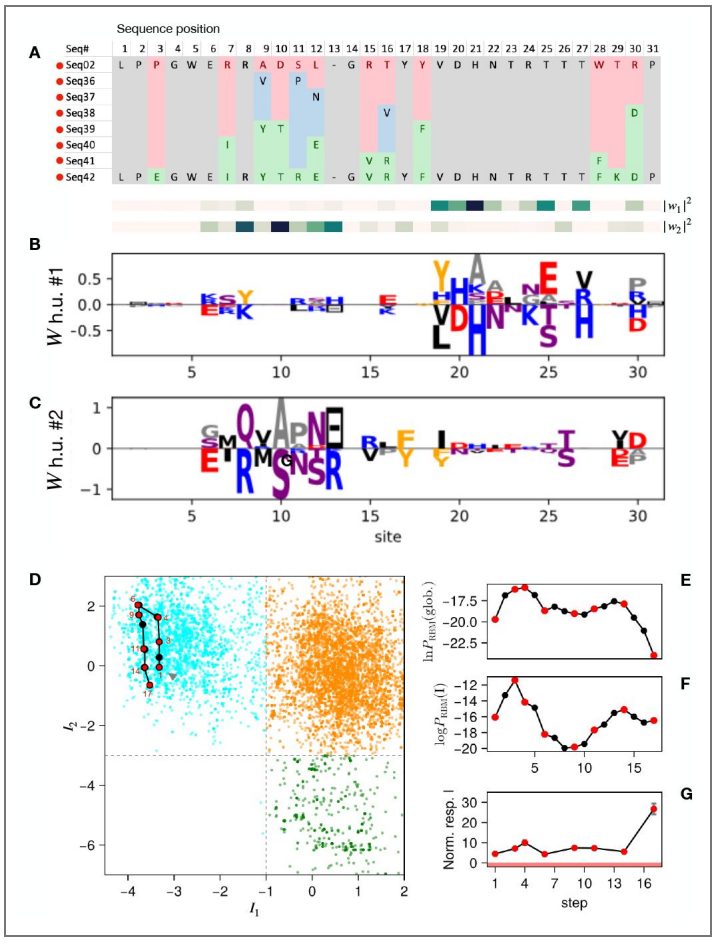


Figure 2. Path within class I WW domains.

A. Alignment of the tested sequences showing the mutated amino acids. **B, C**. Logo representation of the weight vectors attached to the two RBM latent units most correlated with the specificity classes. **D**. Projection of the path on the latent weights, clustering sequences according to their binding specificities. The specificity thresholds indicated in the figure (Class I: $I_1 < -1$ and $I_2 > -3$; Class II/III: $I_1 > -1$ and $I_2 > -3$; Class IV: $I_1 > -1$ and $I_2 < -3$) are informed by previous specificity assays on natural sequences [9, 8] (see Figure 3 in [43] and Figure 2a in [30]). **E, F**. Global and Class I specific RBM scores for the sequences in the path. **G**. Experimental binding responses of the tested sequences. The red band, defined by the experimental noise limit, denotes the non-binding zone.

3 Results

We present and discuss hereafter a path connecting two natural WWs of type I, three paths connecting natural WWs of types I and IV, two paths connecting natural WWs of types I and II/III. Along each designed path, we pick up the sequences at regular intervals for experimental measurements of their binding affinities to ligands of classes I, II/III and IV. The density of experimentally assessed sequences is higher close to specificity switching. To test the presence of epistatic effects in specificity switching paths we compare our designed paths with scrambled paths, in which the same list of mutations present in the paths generated by the model are introduced in a different order, hence breaking epistatic interactions with the background sequence.

A complete list of the designed paths and tested sequences is provided in Tables in Appendix 8.1.

3.1 Assessment of natural WW activity levels

Several WW domains perform their function in vivo either as multi-domains or as part of a complex architecture [48, 49] and may require additional activating, post-translational modifications. Our in-vitro assay is based on the expression of individual WW domains and, as a consequence it is possible that binding activity may be lost. To explore and avert these issues, we tested a set of 28 wild-type sequences taken from the homologous sequence data (12 Class I, 15 Class II/III, 1 Class IV) against three peptides, representative of the targets of the three classes. Out of these sequences, a total of 8 did not give any response in our assay (2 classified as Class I and 6 classified as Class II/III). This is in particular the case of the YAP1 WW domain that we considered in [30]. Half of the remaining sequences gave a clear and specific response for the peptides associated to the classes (8 sequences for Class I, 5 for Class II, and 1 for Class IV). Two sequences showed cross-specificity to peptides associated to Classes I and II. In the following we discuss paths starting and ending at sequences with clear and specific responses (Appendix 8.1)

3.2 Path within class I WW domains

We start by testing a path between two Class I wild-types (2LTW and NP595793.1). Supp. Figure 7A shows one of these natural WW domains in complex with the cognate peptide. We show in Figure 2A the sequences of the two natural WWs of type I we considered, as well as the six tested intermediate sequences, sampled by the RBM-path algorithm connecting the two wild-types that were experimentally tested. The full sampled path is given in Suppl. Fig 14). Residues that change along the path are colored. As shown on the structures in Supp. Figure 7A, a large fraction ($\sim 40\%$) of amino acids is mutated along the path. In Figure 2B,C we show the weights attached to the two specificity-related latent units of the RBM (I_1, I_2). Projections of the sequences in the path are consistently confined within the cluster attached to class I (Figure 2D). Most mutated amino acids are not proximal with the ligand, and not located on the sites with large entries in the RBM-weight attached to the latent units detecting class I specificity (I_2) (Figure 2B), in agreement with the fact that the specificity remains unchanged along the path. Twelve positions in total are altered across the path, four of which (9, 11, 12, 16) sample an intermediate residue, allowing to improve the global score of the artificial intermediate proteins with respect to the initial and final natural wild-type.

Figure 2E shows the measurement of the binding to a class I peptide of the tested sequences in the path. All tested sequences show specificity responses much above noise level. The global RBM scores of the sequences along the path and the class I specific RBM score are shown in Figure 2E,F. Interestingly, designed sequences with the largest scores, in particular the second and the third sampled ones, have larger responses than the initial natural WW.

The presence of indirect (transient) mutations (9, 11, 12, 16) improves the global score. In particular, several pairs involving site 9 (9-13, 9-14, 9-16) have a large epistatic score according to the RBM model, and correspond to contacts on the 3D fold, see Suppl. Figure 9. Indirect mutations may be therefore important for stabilizing the designed WWs along the path by exploiting epistatic interactions. As an illustration, the transient mutation A9V appearing on the second

tested WW (Seq. 36) increases the stability via a favorable contact interaction, as seen from the value of the Miyazawa-Jernigan energy [50] between amino acids V9 and L12. This interaction is lost on the third tested sequence (Seq. 37) due to the mutation L12N compensated by the mutation T16V through a favorable contact interaction between amino acids V9 and V16. The Class I natural sequence anchoring the path (Seq. 42) has the smallest global score according to the RBM, but a good Class I local score and a very good binding response, illustrating the difference between the global score, a proxy for stability, and the specificity scores. In conclusion, the experimental probing of a designed path connecting two Class I wild types allows us to assess the binding response of the designed sequences along the path and to validate the RBM model predictions, in particular the epistatic interactions and the global and local scores.

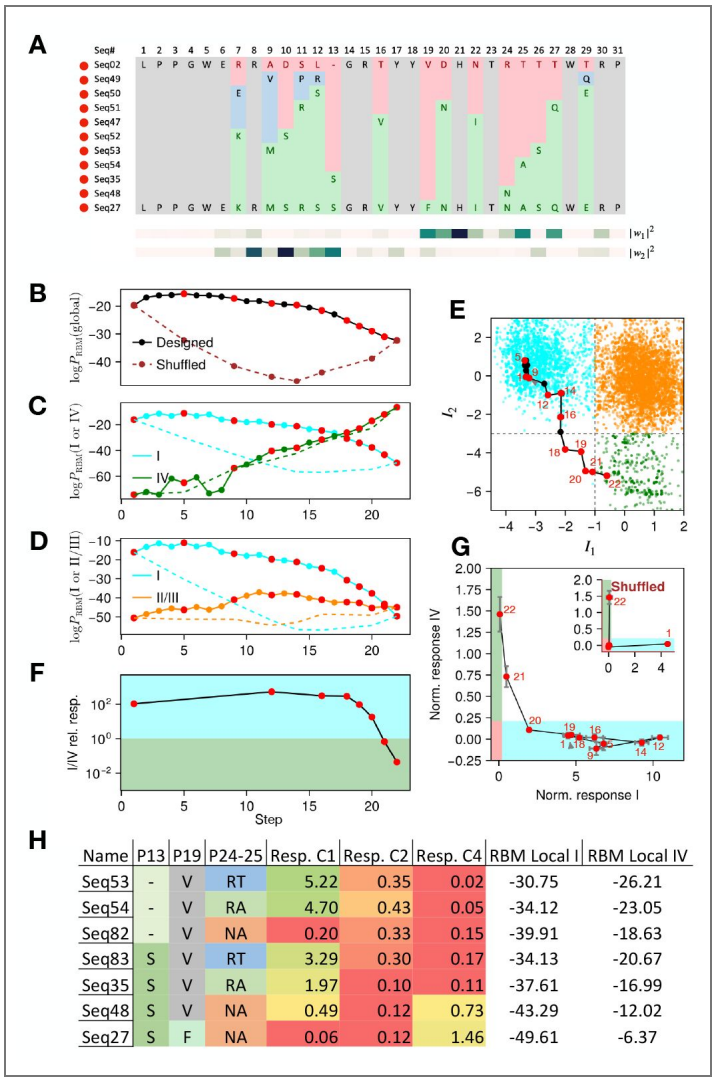
3.3 Paths from Class I to Class IV

We now consider a path designed to interpolate between two different specificity classes, I (wild-type NP595793.1) and IV (wild-type PIN1). As shown in Supp. Figure 7B and previously reported [51], ligands to Class IV domains bind in a different pocket compared to their Class I counterparts. This alternative binding mode requires a longer loop between the β_1 and β_2 strands. Position 13 carries a gap in the Class I WW domain, and amino acid S in the Class IV natural domain (Figure 3A). The existence of two binding pockets may favor switching specificity and binding cross-ractivity as was hypothesized in [30].

Class IV poses challenges from the computational point of view: the number of natural sequences in the cluster associated to Class IV is small compared to the other specificity classes, see green dots in Figure 2D. The number of experimentally validated Class IV sequences is even smaller ([8, 9] and Figure 3 in [43]). Due to the lack of training data, we expect RBM models to be less predictive for Class IV sequences..

Figure 3A shows the list of experimentally tested sequences obtained by subsampling the full path (sampled sequences are represented by red dots; the full path is given in Suppl. Fig 15). About 50% of the protein sequence is modified along the path. Most of the mutated residues are within or close to the two binding pockets, which can be recognized based on the large norm of the RBM weights (w_2^2 and w_1^2 in Figure 3A) of the two specificity-related latent units. The importance of the residues in positions 8 and 13 and their changes in the path to achieve Class IV specificity agree with the weights (previously shown in Figure 2B) attached to unit 1 (w_1^2 in Figure 3A). On the contrary, sites 19, 24 and 25 carry large entries in the weight (Figure 2C) associated to latent unit 2 (w_2^2 in Figure 3A) [43, 30]. Positions 7, 9, 11, 12, 29 carry indirect mutations in the path (shown in blue in Figure 3A). The global RBM score (Figure 3B) is slightly smaller near the end of the path than at the beginning, perhaps due to the difficulty of properly modeling Class IV-like sequences or to a larger stability of Class I molecules with respect to the other classes, as shown by the MSA alignment in Suppl. Figure 10. We observe the expected trends for the specific RBM models: Classes I and IV have, respectively, lower and higher scores at the end of the path than at the beginning (Figure 3C). The two scores cross after the 6th sampled sequence in the path (Seq. 52). The path then goes, in the low-dimensional projection, across a region devoid of natural sequences (Figure 3E). Interestingly, in this region, the Class II/III specific RBM score gets closer its Class I counterpart (Figure 3D). Experiments confirm that sequences in this region keep a significant activity towards both Class I and Class II ligands (Suppl. Table 4). Furthermore, the last designed sequence along the transition path (WW150) shows cross-reactivity to Class I and IV ligands, consistently with our argument given in [30] (Figure 3G). Taken altogether, these results points towards a substantial level of promiscuity [30] along the path.

To confirm that the order in which mutations are introduced along the designed path is dictated by epistatic interactions, and thus depends on the amino-acid background, we study “scrambled” paths, in which mutations are introduced in reverse order (Suppl. Table 7). We observe that scrambled paths have global scores smaller than designed paths (Figure 3B), and do not contain any functional sequence (apart from the anchoring natural WWs), see inset of Figure 3G. Of notice, epistasis effects are correctly detected by Class I specific scores, but not by Class IV specific scores that assign a similar score to scrambled and designed paths. This observation is compatible



Name	P13	P19	P24-25	Resp. C1	Resp. C2	Resp. C4	RBM Local I	RBM Local IV
Seq53	-	V	RT	5.22	0.35	0.02	-30.75	-26.21
Seq54	-	V	RA	4.70	0.43	0.05	-34.12	-23.05
Seq82	-	V	NA	0.20	0.33	0.15	-39.91	-18.63
Seq83	S	V	RT	3.29	0.30	0.17	-34.13	-20.67
Seq35	S	V	RA	1.97	0.10	0.11	-37.61	-16.99
Seq48	S	V	NA	0.49	0.12	0.73	-43.29	-12.02
Seq27	S	F	NA	0.06	0.12	1.46	-49.61	-6.37

Figure 3. Path from Class I to Class IV.

A. Alignment of tested sequences along the path showing mutated amino acids. **B**,**C**,**D**. Scores of sequences in the path predicted by the global (top), Class I or IV (middle) and Class I or II/III (bottom) RBM models. **E**. Two-dimensional projection of the path in the plane of specificity-related latent units. **F**. Relative binding responses to Class I and IV peptides of sequences along the path. **G**. Normalizes experimental binding responses to Class I and IV peptides. **H**. Responses of Seq.53, 54, 35, 48, 27 compared with Seq.82, 83 differing by mutations on residues 13 (carrying a gap for a shorter $\beta_1 - \beta_2$ loop or an S), and class I specificity-determining positions 24-25.

with the absence of epistasis (Suppl. Figure 10) in the Class IV specific RBM model, probably due to the the scarcity of training data, but could also be related to a lower stability than Class I domains. To better understand the correlation between the two binding modes towards Class I and Class IV ligands, we study the specificity of sequences with adequate pairs of amino acids in position 24 & 25 for Class I specificity, namely RA or RT, or a mutated pair, NA, decreasing the binding affinity to Class I peptide, as well as long or short $\beta_1 - \beta_2$ loop. We find that extending the $\beta_1 - \beta_2$ loop is not sufficient to achieve Class IV specificity, e.g. the sequence with RT in P24-25 and a long loop is characterized as being in Class I. Moreover, decreasing the affinity of Class I with the NA mutations and extending the loop is necessary to bind class IV ligands, underlining the presence of a long-range interaction between the two binding modes.

As shown in Suppl. Table we have tested 3 additional paths interpolating between Classes I and IV. Remarkably, even on the path starting from the YAP1 wild-type, classified as non binding by our setup, all the designed sequences are functional (Supp. Table 3). All paths are found to follow similar trajectories in the two-dimensional plane, and show cross-reactivity in the region of the input space devoid of natural sequences, see Suppl. Figure 13: several sequences (Seqs. 52, 53, 54, 34, 35) even show mild cross-reactivity to Class II peptide. All paths pass through sequence 'WW150' before going directly to 'WW10' in 2 mutations, showing how constrained is the convergence towards Class IV reactivity.

We have also designed a direct path interpolating between the two classes (Suppl. Figure 13), along which transient mutations are forbidden. This path shows smaller RBM global score, especially for Class I sequences, in agreement with Section 3.2. Experimentally measured responses are very similar to the one found for indirect paths (red: no response; green: large response). In the region lacking natural sequences, the direct path goes through the same sequences (Seqs. 53, 35, 48) as the indirect ones, again underlying how constrained are sequences when approaching Class IV.

3.4 Paths from Class I to Class II/III

We now report results on paths interpolating between Classes I (wild type NP595793.1) and II/III (Wild type 1YWI or formin binding protein), see Figure 4. This path critically differs from the Class I → IV case above as ligands of Classes I and II bind to the same pocket of the WW domain [9, 43, 8, 14] (Supp. Figure 7C).

Figure 4A gives the list of experimentally tested sequences obtained by subsampling the complete path given in Suppl. Fig 16. Note that the natural Class II/III domain at the end of the path (Seq.64) has two flanking sequences (2 amino acids in the N-terminus and 6 in the C-terminus), which are not covered by the RBM model. To study a path ending at Seq. 64 we thus added the same flanking regions to Seq. 2, leading to Seq. 2x with no loss in specificity, as well as to all the designed sequences along the path. Supp. Figure 7C shows the complexes formed by the WW domain and the two ligands; the majority of sites (∼ 75%) in the molecule undergo mutations. Several residues are transiently mutated at the beginning of the path. Some of them (9-11-12) were already encountered on the path within Class I above, and increase the global scores (Figure 4B) with respect to the wild-type WW domain.

The global and specific scores relative to Classes I and II/III of the sequences are shown in Figure 4B,C. Notice that the natural WW domain of Class II/III at the extremity of the path has much lower global and specific scores than its Class I counterpart. This may be due to the fact that Class II/III domains are generally less stable than in Class I, as suggested by the contact map obtained from Class II/III specific RBM (Suppl. Figure 10) and by the lower conservation of residues in Class II/III domains, see sequence logo in Suppl. Figure and [48] (Table 3). We predict a change in specificity after the fifth sampled sequence (Seq. 61) from the low-dimensional representation of Figure 4D and from the crossing of the specificity scores in Figure 4C.This switch requires mutations of all specificity-determining residues: V19W, H21T [9, 48, 43] and R24N, compatible with the weight vector associated to I_2, and sequentially introduced from Seq.84 to Seq.86 along the path.

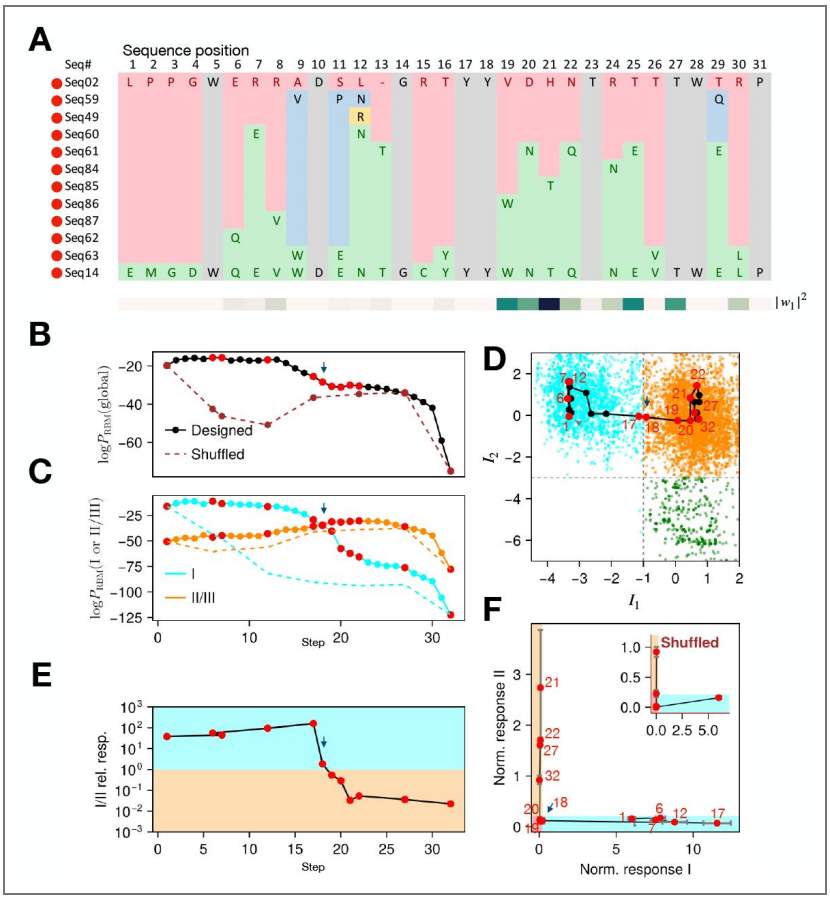


Figure 4. **Path from Class I to Class II/III.**

A. Alignment of the tested sequences, showing the mutated amino acids. Same color code as in Figure 2. **B, C**. Global and class-specific RBM scores of the sequences in the path. **D**. Projection of the path one the plane of RBM latent units relevant for class identification. **E**. Ratio of experimentally measured responses to peptides of class II over Class I for intermediate sequences along the path. **F**. Normalized responses of the designed sequences against peptides of Classes I and II. The red band, defined by the experimental noise limit, indicates the non-binding zone.

The normalized responses of the tested molecules are shown in Figure 4F. To better characterize the switch, we investigate the responses of the four designed sequences following Seq. 61 in the path (crossing the specificity threshold in Figure 4E) to the three peptides (C1, C2, C3). No response to C2 peptide is found until the complete mutations of the three key residues above in Seq. 87, showing the largest C2 binding response, coinciding with the maximum of the Class II/III score. Previous rational design-based experiments for switching the specificity of Class I YAP (with peptide PPPPXP) WW to Class II/III FE65 (with peptide PPPPPP) have already pointed out that the two substitutions L19W and H21G were sufficient to perform specificity switching in the YAP1 background having Q24. As already mentioned, YAP1 is not responsive in our assay. Moreover, given the large fraction (6/15) of non responding wild-type sequences of class II, the response to the C2 peptide seems harder to detect. Lastly, we do not exclude that other non-tested ligands could be able to bind to such sequences, eventually in a less specific and more cross-reactive way.

We next experimentally analyze a set of 'scrambled' sequences (Suppl. Table 7). These scrambled sequences have RBM specific and global scores smaller than the designed ones, and none of them is functional, see Figure 4F inset.

We report experimental results obtained for a second path switching specificity from classes I to II/III in Suppl. Table 6, see Suppl. Figure for a low-dimensional visualization of the path. All tested sequences but one are responsive and bind peptides of Classes I or as predicted by the class-specific scores.

3.5 Comparison of model scores and binding measurements

The analysis of the mutational paths above has shown the ability of our computational model to generate new sequences behaving as WW domains with desired binding specificities. To further quantify this capability, we study the correlations between the global RBM scores and the experimentally assessed binding responses in Figure 5. The scores of the tested sequences, either natural or designed, have large values, biased towards the right tail of the histogram of natural WW sequences (Figure 5A). In particular, most designed sequences are, according to the global RBM model, as good as, or better than the majority of natural sequences. This result was expected as we set the effective temperature to the value 1/3 (lower than the standard value 1) to design sequences with high scores [34].

Sequences along the paths obtained through reshuffling of the mutations in Figures 4 and 3 have, on the contrary, smaller global scores, in agreement with the experimentally observed lack of functionality. The scrambling of mutations, irrespectively of the amino-acid background, is expected to be detrimental in the presence of epistatic interactions.

Histograms of scores for the natural sequences associated to specific classes are given in Suppl. Figure 8 and confirm the RBM scores of Class I natural sequences have typically larger values than the ones of Classes II/III and IV, which could be indicative of stronger structural stability.

In Figure 5B,C,D, we compare, for each class, the scores assigned by the specific RBM models to the experimental binding measures of the corresponding class ligand. We observe positive correlations, see Figure 5E for the three models, and less so for the global score. Class I is an exception, as the global score also shows a strong correlation; this may be due to the fact that class I WW have the larger scores as shown in Suppl. Figure 8 and several contacts are kept by conserved amino acids in Class I domains [48], see logo representations and contact map for Class I in Suppl. Figure 10.

3.6 Comparison with Ancestral Sequence Reconstruction on WW domains

Lastly, we compare the mutational paths designed with our method between wild-type domains with the pathways linking these domains on the phylogenetic tree derived from the WW domains present in the PFAM seed. To do so, we reconstruct the tree connecting all these domains and perform Ancestral Sequence Reconstruction (ASR) on the intermediate nodes. This tree topology is inferred directly from the WW domains only [52] and not from the proteins they belong to.

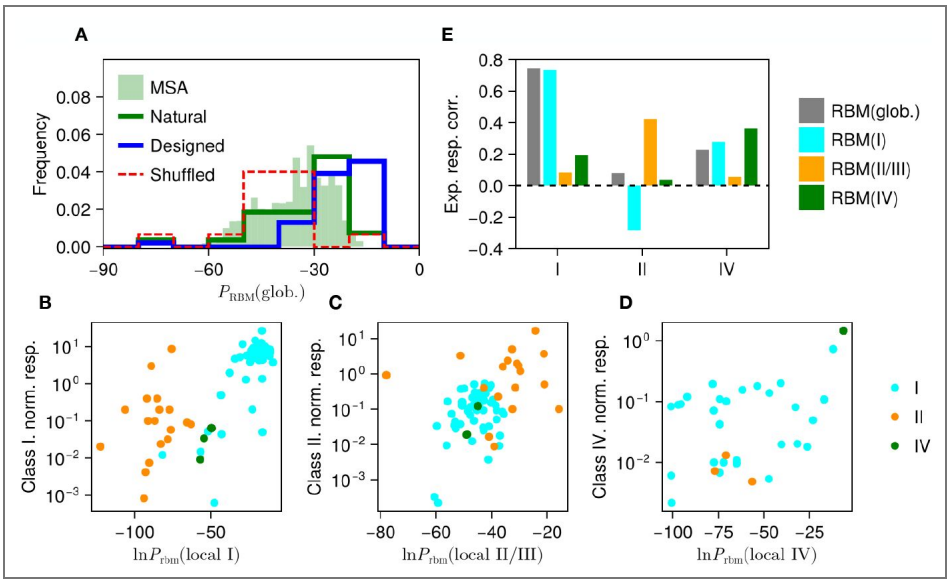


Figure 5. Model scores and experimental responses.

A. Histogram of global RBM scores of probed sequences (Natural, Designed, and Shuffled). The histogram of global RBM scores of the full sequence data is also shown for comparison. **B, C, D**. Comparison between RBM local scores and binding responses, for Class I (B), II/III (C), and IV (D) ligands. Sequences in each panel are colored according to the predicted classes, *i.e*. to the quadrants in the 2D projection plane defined by the specificity-related latent units. **E**. Pearson correlation coefficient between experimental responses to the three ligands (I, II, IV), and the RBM scores (specific and global).

Multiple pathways connect the domain subtypes on the tree (see supplementary figure 19). As illustrated in figure 6A and with the collapsed tree topology in figure 6A, we observe that all type I and IV domains are connected through a narrow path, which crosses the empty quadrant in the 2D-latent space defined by the two specificity-related RBM hidden units. This result is similar to what is obtained with our RBM-based method.

While ASR paths have in median a large number = 113 of mutations (Supplementary Figure 20, middle bottom panel), the RBM paths were constrained to be much shorter, with 14 to 25 mutations, a length shared by few ASR paths only, see Supplementary Figure 13A, B, C. This observation could explain why RBM-designed paths cross a region devoid of contemporary sequences. Remarkably, while ASR paths are not constrained in length, the pathways connecting type I and type IV domains also follow the same narrow path. Notice that while on the phylogenetic tree inferred from WW domain sequences only, the topology directly connects domains of type I with those of type IV, different results are found with trees obtained from families of homologous proteins containing WW domains as a subpart of their sequences. In the latter case, paths going from I to IV pass through region II/III (Supplementary Figure 21) In particular, the phylogeny of the protein NP001351424 (peptidyl-prolyl cis-trans isomerase from *Mus musculus*), which contains the WT class IV domain used as anchor for path sampling, shows intermediate ancestors between class I and II/III homologs belonging to the class II/III (Supplementary Figure 21D).

The funnel connecting directly WW domains of classes I and IV is therefore an interesting specific feature of WW-domain reconstructed trees, not locked to the evolution of the rest of the protein, and possibly better reflecting the selective pressure associated with binding specificity. The differences between the topologies of the trees found from the WW domains and from the proteins containing them are compatible with previous observations [52]; in particular, domains with different specificity types in the same protein may be less related to domains of the same type on different proteins.

Figures 6B shows that the ASR sequences along the pathways (in purple), connecting the wild-type sequences anchoring the RBM paths are strikingly similar with the RBM intermediates. Their residue identity exceed 90%, as evidenced in Figure 6E and F. Conversely, the median identity between randomly selected modern domains in the seed is 33%. However, the scores of the ancestral sequences along the phylogenetic pathways assigned by the RBM are significantly lower than the ones of the the RBM-designed sequences (Figure 6C and D). This result is expected as ASR reconstruction does not take into account epistasis, differently from RBM, and we expect ASR sequences to generally be of lesser quality.

Incorporating the RBM-designed sequences into the phylogenetic tree further corroborates their similarity with the ancestral sequences. These RBM sequences collapse on the trajectory connecting the wild-type sequences (Figure 6B, with the dotted purple ASR paths overlapped on the blue synthetic paths).

4 Discussion

In this work, we propose and test a computational method to design mutational paths allowing for specificity switching [18, 53]. We focus on WW domains, a small binding unit of many proteins, whose variants specifically bind to different classes of proline-rich peptides. Our approach relies on a model of the fitness landscape of WW domains learned by Restricted Boltzmann Machine (RBM) from homologous WW sequence data and on Monte Carlo path sampling [30, 44] of functional paths [26] that connect extant WW domains across sequence space. Using a high-throughput in-vitro approach, we test the binding response of the designed sequences toward three peptides representative of the three main WW specificity classes. Our study provides experimental validation of multiple designed paths both within the same WW specificity class (binding the same peptide) and across different specificity classes, capable of switching specificity through accumulation of mutations. These experiments, in turn, give insights on the main

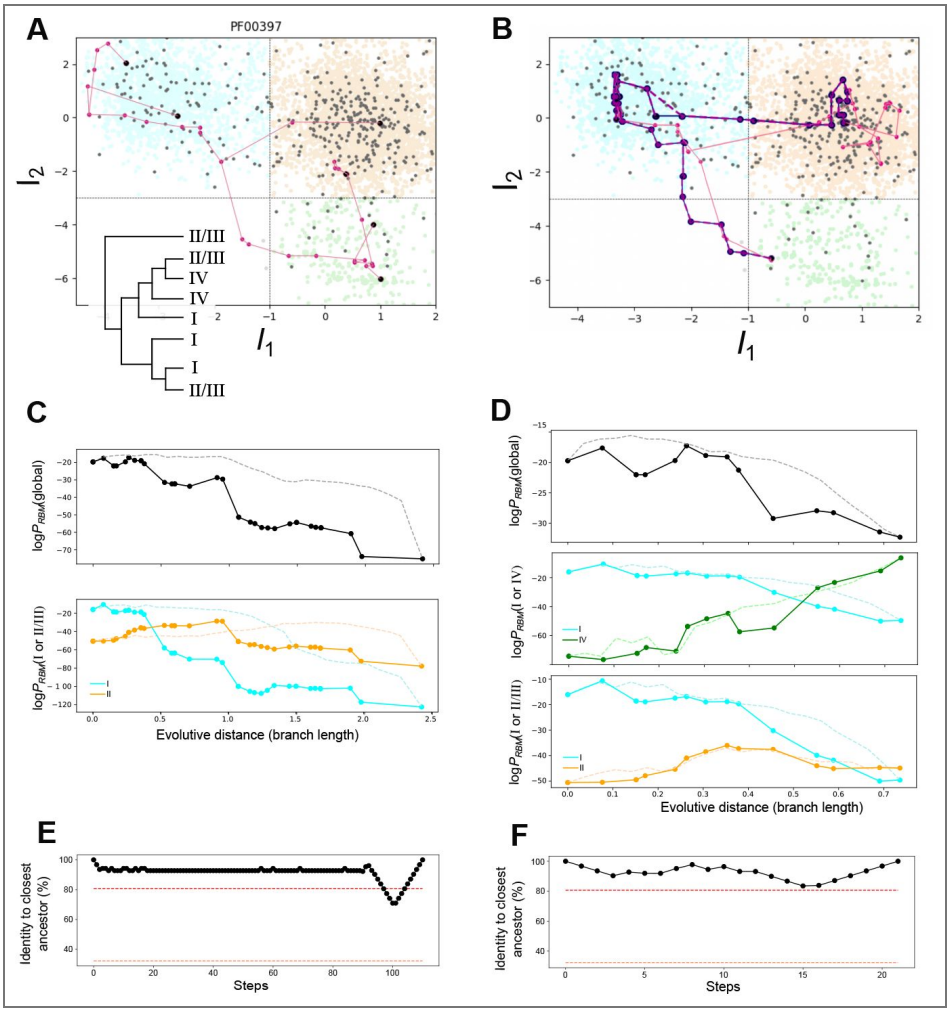


Figure 6. Pathways derived from ancestral sequence reconstruction (ASR) on the WW domains in PFAM's seed.

A. 2D projections of selected paths linking domains of different types (I, II/III and IV) on the two specificity-related RBM hidden units, and simplified representation of the seed tree topology. Black dots: modern sequences; Purple dots: average ancestral sequences sampled from the posterior distribution. **B**. 2D projections of paths linking WT domains presented in Figures 3 and 4 along the tree. Black: mutational paths designed with the RBM; Purple: paths obtained with ASR; Dotted purple: ASR path obtained when adding to the WW tree the RBM intermediates. Blue and purple paths are largely overlapped with one another. **C**., **D**. Scores of ancestors obtained with ASR (solid lines) and with the global (top) and specific (middle panels) RBM models. Scores obtained along RBM-sampled paths are shown as dashed lines for comparison. **C**. Path I to II/III. **D**. Path I to IV. **E**., **F**. Identity between intermediate sequences along the RBM path and the closest ancestors reconstructed by ASR; median and maximum identity between seed WW domains are shown as dashed red line for reference. **E**. Path I to II. **F**. Path I to IV.

requirements for the change of specificity in a WW protein domain. We finally compare the RBM-designed paths with pathways obtained by ancestral sequence reconstruction (ASR) on a phylogenetic tree built on the WW domains of the PFAM seed.

Navigability in the fitness landscape was first studied in evolutionary biology using theoretical models [18, 54, 55, 56]. In recent years, the availability of massive experimental data began to allow for partial fitness landscape reconstructions. The approach introduced here for the path design is, in a way, a synthesis between these two approaches. Unlike theoretical frameworks such as the 'house of cards' [57] or the NK [54] models, our fitness landscape is directly inferred from sequence data and is therefore, at some level of approximation, representative of the complexity of a real protein domain [43, 30, 58]. Moreover, our approach borrows to statistical physics some conceptual and methodological ingredients, such as Monte-Carlo sampling of paths [54]; in addition, the bipartite nature of the RBM networkmakes it possible to use mean-field approximations to theoretically characterize the topology and accessibility of the fitness landscape [54, 23], e.g. the mean trajectory of paths and their entropy [44, 59]. Computational approaches enormously reduce the huge number of a priori paths to be experimentally tested, and could speed up the rational design of protein selectivity, attempted since the 1990s in the field of drug discovery, in which few mutations targeting key specificity residues were probed, often ignoring the presence of complex epistatic effects between residues [14, 60, 61, 62, 63].

Overall the designed paths are rich in functional sequences: only 4 out of 58 tested sequences were non responsive to any of the three tested ligands, while all the other sequences responded to at least one of the peptide representative of the specificity class corresponding to the natural WW domain at the extremities of the path. This result proves the general validity of combining generative models for the fitness landscapes inferred from sequence data and Monte-Carlo path sampling to propose viable paths. It extends the recent success in protein design with computational methods [64] to the design of full mutational paths. Our RBM model is capable of capturing important epistatic effects constraining the order in which mutations can be introduced along the path. As shown by scrambling the mutational order along the path, sequences in which mutations take place independently from the background content are not functional, as they do not take into account epistasis. We stress that the modeling of epistatic interactions with the RBM is global over all WW domains, and not attached to any specificity class. It is not possible to learn accurate landscapes for each single specificity as only ~50 out of 17,000 available WW sequences in databases have been experimentally characterized in terms of binding specificity [8].

Being shallow, the RBM parameters of the model are interpretable, and unveil both key specificity determining residues, similarly to methods based on principal component analysis ([65, 66]) and epistatic interactions as in the Direct Coupling Analysis [67, 68]. Moreover, the tested specificities along the paths are in good agreement with the approximate class-specific RBM scores we have introduced; specificity switches take place through changes in determining residues and amino acids identified by the RBM weights attached to the two hidden units identified as most informative about specificity classes [43].

Our designed paths allow for the presence of "transient" substitutions with amino acids different from the ones in the anchoring wild-type WW sequences [69]. In all the paths we tested, we observe such mutations to transient amino acids on variable residues away from the main specificity determinant. These reversed mutations are important: they transiently improve the global RBM score in the initial part of the path, starting from a class I wild-type, and possibly increase the stability of the molecules with respect to the wild-type, while preparing the background for the specificity change. However, class switching typically involves mutating the specificity determining residues to the amino acids found in the final wild-type sequence. The above findings are in full agreement with the stability-specificity tradeoff previously introduced in evolutionary path characterization [56, 70, 71] and in the specific characterization of Pin WW variants [72].

Mutational paths connecting specificity Classes I and IV are particularly interesting, because they cross a region deprived of natural sequences in the low-dimensional projection defined by the two RBM specificity hidden units, see Figure 3E [30]. Synthetic WW domains along the I-IV paths go

through a bottleneck in the space of sequences: several designed and tested paths go through the same intermediates, and are characterized by cross-specificity to peptides of Classes I, II and IV. Ancestral promiscuity and switch of ligands have been proposed for several enzymatic families based on ASR [17, 73, 29]. Promiscuous WW domains could be related to ancestral states, which could have specialized during evolution in favor of more stringent ligand selectivity. In Humans, WW domains exist also in tandem and multi-domains, and gene duplication is a plausible mechanism for such a specialization [49]. Reconstruction of ancestral sequences (ASR) from domains contained in PFAM's seed supports this hypothesis, unveiling a similar narrow pathways connecting type I and IV domains, with ancestral nodes showing a large residue identity with the RBM sequences along the path. However, phylogenetic reconstructions on several families of homologous proteins containing WW domains rather show transitions between classes I and II/III and classes II/III and IV. It is therefore difficult to assess the evolutionary significance of the putative ancestral sequences along the I-IV path found in the WW-domain tree.

ASR and RBM-path sampling are complementary tools to explore hypothetical paths between distantly related modern sequences, both based on hypothesizing evolution as a parsimonious path in a fitness landscape. This analogy between the fitness landscape inferred from sequence data and the phylogenetic paths has already been used in protein engineering as a method to infer novel proteins with enhanced properties such as affinity, stability, and solubility [74, 75] and to predict mutational effects in [76], where the distance between sequences on a Neighbor-Joining tree serves as a proxy to assess the likelihood and, consequently, the fitness of novel mutations in a protein.

However ASR offers much lower resolution in the sequence space compared to the RBM, as the computational complexity in reconstructing a tree makes impossible to use all sequences available in the databases; it does not provide a true quantification of the likelihood of the intermediate sequences and does not model epistatic interactions between residues. The similarity of ASR and RBM paths proves that the sampling of paths in the fitness landscape with the RBM could help to explore likely intermediate sequences between groups of distantly related proteins, a task for which classical ASR is not well suited;which could prove useful both for protein engineering and phylogenetic studies.

Our work could be extended in several directions. First, we plan to extend our approach to sample paths attached not at their two extremities, but anchored to one wild-type WW domain only. Imposing that the affinity to a ligand different from the initial one increases along the path would make the setting more closely matches a natural evolutionary path.

Second, the in-vitro assay could be complemented in several ways. Our current assay does not allow for cooperativity between WW domains, which may be important in vivo. This could explain why some natural WW domains tested alone, such as YAP, do not display affinity to their cognate ligands. It has been previously shown that multi tandem WW domain confer higher ligand affinity [49]. In addition, our current experimental setup is limited to a small number of peptides. It is possible that the four non-responsive sequences we have sampled on the path connecting Class I and Class II/III wild-type WW domain may bind to other non-tested peptides representative of the same classes, possibly in a less specific and more cross-reactive way. To better investigate the above points we plan to extend the experimental binding measures to in vivo binding assays, and to measure binding against a library of peptides through competitive growth in yeast cells [9, 77].

Lastly, while we have here used experiments as a way to validate the paths proposed by the computational method, it would be interesting to iteratively improve our estimate of the fitness landscape inference by integrating the new experimentally labeled data using active learning methods [78, 79, 80, 64].

5 Materials and Methods

5.1 Restricted Boltzmann Machines

We train our unsupervised model on a data-set of homologous proteins, presented as a list of aligned sequences $\{\mathbf{v}^1, \mathbf{v}^2, ..., \mathbf{v}^k, ..., \mathbf{v}^B\}$ (B is the total number of sequences in the data set), where each entry is an array $\mathbf{v}^k = \{v_1^k, \ldots, v_i^k, \ldots v_N^k\}$ in which each one of the N entries may be in one of the possible 21 states (20 amino acids + the gap site). As a model, we use Restricted Boltzmann Machines [81], a neural network consisting of a visible layer $\mathbf{v}$ representing the data and an hidden layer of M neurons $\mathbf{h} = \{h_\mu\}_{\mu=1}^M$. RBMs define a joint probability distribution over the visible and hidden layer as

$$P_{\rm RBM}(\mathbf{v},\mathbf{h}) = \frac{1}{Z_{\rm RBM}} \exp\left[\sum_{i=1}^N g_i(v_i) + \sum_\mu I_\mu(\mathbf{v}) h_\mu - \sum_\mu \mathcal{U}(h_\mu)\right], \qquad (1)$$

where $Z_{\rm RBM}$ is the normalisation constant, the input to an hidden unit is the projection of the sequence on the weigth matrix $w_{i,\mu}$: $I_\mu(\mathbf{v}) = \sum_i w_{i,\mu}(v_i)$ and $\mathcal{U}_\mu$ are potential energy functions over the real-valued hidden unit activations h_μ. We choose $\mathcal{U}_\mu$ to be double Rectified Linear Unit (dReLU) potentials of the form

$$\mathcal{U}_\mu(h) = \frac{1}{2}\gamma_{\mu,+}h_+^2 + \frac{1}{2}\gamma_{\mu,-}h_-^2 + \theta_{\mu,+}h_+ + \theta_{\mu,-}h_-, \text{ where } h_+ = \max(h,0),\ h_- = \min(h,0), \qquad (2)$$

where we have defined the hyper-parameters $\gamma_{\mu,\pm}$, $\theta_{\mu,\pm}$. Marginalising over the hidden units we obtain the probability distribution over the sequence space

$$P_{\rm RBM}(\mathbf{v}) = \frac{1}{Z_{\rm RBM}} \exp\left[\sum_{i=1}^N g_i(v_i) + N\sum_{\mu=1}^M \Gamma_\mu(I_\mu(\mathbf{v})/N)\right], \qquad (3)$$

Where $\Gamma_\mu(I/N) = \frac{1}{N}\ln\int \mathrm{d}h\, e^{I\,h - \mathcal{U}_\mu(h)}$. We use Persistent Contrastive Divergence [82] to train the model over the multi-sequence alignment of the protein family of interest in order to maximise the likelihood. This training algorithm has been shown to be sufficiently robust under cautious regularization [83]. The code and the data used to train our RBMs for WW domains can be found in [84], while the hyperparameters for the training can be found in the Supplementary Materials of [30].

5.1.1 Global and specific RBM models

We learned two classes of RBM models, depending on the train data. A global RBM model used to design the paths was learned on all the sequence data, while three class-specific RBM models, for binding classes of type I, II/III, or IV were learned on sub-set of sequences (containing 8304, 8292, 637 sequences respectively). To define the sub-sets we clusterized the sequences according to the value of their inputs $I_\mu(\mathbf{v})$ on the two hidden units $\mu = 1$, $\mu = 2$ chosen as the most informative about the specificity classes, shown in Figs. 2, 4, 3. The specificity threshold indicated in the figure ($I_2 < -1$; $I_2 > -1, I_1 < -3$; $I_1 > -3$) have been informed in [30] by some previous experimental tests [9, 8] (see Figure 3 in [43] and Figure 2a in [30]) on the WW specificity classes. This results in three class-specific RBM models (one for class I, one for class II/III and one for class IV) which can be used as predictor of the binding affinity of a natural or artificial sequence.

5.2 Path sampling algorithm

We define the probability a path of T steps, $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2 ..., \mathbf{v}_{T-1}\}$, connecting two sequences $\mathbf{v}_{\rm start}$ and $\mathbf{v}_{\rm end}$, as follows:

$$\mathcal{P}[\mathcal{V}|\mathbf{v}_{\rm start}, \mathbf{v}_{\rm end}] = \frac{1}{Z_{\rm path}} \pi(\mathbf{v}_{\rm start}, \mathbf{v}_1) \times \prod_{t=1}^{T-1} P_{\rm RBM}(\mathbf{v}_t) \times \prod_{t=1}^{T-2} \pi(\mathbf{v}_t, \mathbf{v}_{t+1}) \times \pi(\mathbf{v}_{T-1}, \mathbf{v}_{\rm end}). \qquad (4)$$

Here π is an interaction terms that is used to minimised the number of mutations at each step. Our goal is to sample paths that do at most one mutation at each time step. In practice we set $\pi(\mathbf{v}, \mathbf{v}') = 1$ if the two sequences are equal, $\pi(\mathbf{v}, \mathbf{v}') = e^{-\wedge}$ (with $\wedge > 0$) if the two sequences differ by one

mutation and $\pi(\mathbf{v}, \mathbf{v}') = 0$ otherwise. In [30] we propose a Monte-Carlo (coupled with simulated annealing) procedure to sample paths from $\mathcal{P}(\mathcal{V})^{\beta}$ at a given inverse temperature $\beta = 1/T$. Other choices of the transition term π are discussed in the following subsection.

Here we stress that our path sampling approach is independent of the restricted Boltzmann machine used to train the model and one can use in principle different probability distribution $P(\mathbf{v})$. An alternative choice would be to define our path sampling algorithm over the space of nucleotide sequences that encode for the specific protein. Each sequence would be scored according to the RBM model after translation into the corresponding amino acids sequence according to genetic code. In this way one can introduce codon biases in the sampling approach. We applied this approach to the case of WW domain and we obtained similar results to the sampling procedure carried out without codon bias.

5.3 Ancestral protein reconstruction on seed domains

Phylogenetic tree inference and Ancestral Sequence Reconstruction (ASR) were conducted to obtain an unrooted tree topology and ancestral node sequences for all 408 isolated WW domain in PF00397. We used IQ-Tree [85] and an evolutionary model defined by the amino-acid mutation matrix LG [86] and Gamma parameters discretized in four classes [87] (LG+G4, identified as the best model to describe domain evolution, see Supplementary Material Section 8.5.2). Gaps were reconstructed from the sequence alignment converted into a binary presence/absence matrix from binary sequences, using the GTR2 model [88]. We calculated inputs to the hidden units of the RBM for modern WW domains, as well as their mean values for ancestral sequences based on the posterior probabilities of ancestral residues. Scripts to perform the ASR are available on the project's Github.

We also performed phylogenetic reconstruction from the 24 seed domains within or near cluster associated with type IV domains, as well as the NP00131424 protein used as a type IV anchor for the RBM path sampling, as described in Supplementary Material Section 8.5.2.

5.4 Experimental Materials

Cell free expression system (PURExpress ® In Vitro Protein Synthesis Kit, New England Biolabs, E6800), SNAP labeling Dye (SNAP-Surface® Alexa Fluor® 647, New England Biolabs, S9136), Anti-SNAP antibody (SNAP-tag Polyclonal Antibody, Thermo Fisher Scientific, CAB4255), Streptavidin Magnetic beads (Dynabeads™ M-280 Streptavidin, Thermo Fisher Scientific, 11205D), Protein A magnetic beads (Pierce™ Protein A Magnetic Beads, Thermo Fisher Scientific, 88845), oligonucleotide linkers and eBlocks™ Gene Fragment provider (Integrated DNA Technologies), synthetic peptides provider (GenScript Biotech), Gibson assembly (Gibson Assembly® Master Mix, New England Biolabs, E2611), Polymerase kit (Q5® High-Fidelity DNA Polymerase, New England Biolabs, M0491) DNA purification beads kit (SPRIselect®, Beckman Coulter, B23318) and phosphate-buffered saline buffer (DPBS 10X, Thermo Fisher Scientific, 14200075), DTT (DL-Dithiothreitol solution, Sigma-Aldrich, E2611), Tris-HCl (UltraPure™ 1 M Tris-HCI Buffer, pH 7.5, Invitrogen, 15567-027), Tween-20 (TWEEN® 20, Sigma-Aldrich, P2287), EDTA (0.5M EDTA pH 8.0, Invitrogen, AM9260G), Binding buffer (Tris-HCl 50 mM pH 7.5, EDTA 10 mM, DTT1 mM and 0.5 % Tween 20, prepared in-house), Wash buffer (1X DPBS with 0.5% Tween-20).

5.5 Experimental protocol

Our experimental setup is sketched in Figure 1D. Double-stranded DNA gene fragments encoding the WW domain sequences were synthesized and inserted by Gibson assembly into a plasmid containing a T7 promoter and a ribosome binding site (RBS) upstream of the cloning site, these elements being essential for initiation of the expression process. A sequence encoding a linker and a SNAP-tag has been fused downstream of the WW domain sequence and a stop codon has been placed at the end of the SNAPtag sequence followed by a T7 terminator sequence. The SNAPtag is a protein capable of reacting irreversibly with the benzylguanine (BG) group. This can be used to label the protein fluorescently, using dyes coupled to BG.

After insertion of the WW domain genes, a polymerase chain reaction (PCR) was conducted to amplify the entire construct, producing a linear product encompassing the complete sequence between the promoter and the terminator. The resulting PCR product was then utilized for expression in a cell-free transcription and translation system, with the fluorescent substrate.

Concurrently with the expression of the WW protein, magnetic beads coated with the target peptides, which are presumed to bind with the WW domains, are prepared. These comprise a metallic bead displaying a multitude of copies of the peptide of interest on its surface, as well as a fluorescent dye for the identification of the beads. At the same time the protein A-coated magnetic beads are incubated with a polyclonal anti-SNAPtag antibody for the purpose of assessing the efficacy of the expression process.

Following the expression of the fusion proteins and labelling with the fluorescent dye, a mixture of color-coded beads coated with different peptides and anti-SNAPtag antibody was added to the expression mixture. Upon binding of the WW domain to the peptide, the fluorescent dye attached to the SNAP tag will report the amount of protein bound to the bead. The beads were then washed and passed through a flow cytometer, where the average dye fluorescence intensity on the bead surfaces was measured. This indicated the amount of WW domain fusion proteins bound to each peptide-coated bead and the expression assessment beads, which was then normalized for expression. Each WW domain protein was evaluated against three peptides, each representing a distinct class of WW domain. Further details are available in the supplementary Material 5.5.

Fusion Protein assembly, expression and labelling

eBlocks™ gene fragments for genes comprised in the different pathways were designed, ordered, and cloned into the SNAP fusion plasmid using Gibson Assembly® Master Mix. Using Q5® High-Fidelity DNA Polymerase, PCR reactions were performed to amplify linear products containing all the necessary parts for the expression of the fusion proteins. PCR products were purified using SPRIselect® beads. Expression of the fusion proteins were performed in a PURExpress® in vitro protein synthesis reactions with the purified linear PCR products at 15 nM and SNAP-Surface® Alexa Fluor® 647 at 5 µM final concentration. The cell-free *in-vitro* transcription and translation (IVTT) reactions were incubated for 5 minutes at 37°C followed by 4 hours of incubation at 30°C.

Expression control beads preparation

Expression control beads were prepared using 5 µl of protein A magnetic beads (50 ug) from Thermo Fisher Scientific, washed twice with 1X DPBS from Thermo Fisher Scientific with 0.5% Tween-20 from Sigma-Aldrich, resuspended in 50 µl of 100 ng/µl anti-SNAP polyclonal antibody diluted in 1X DPBS from Thermo Fisher Scientific with 0.5% Tween-20 from Sigma-Aldrich, incubated for 30 min at 37°C with shaking, then washed three times with binding buffer.

Streptavidin beads preparation with target peptides

beads were prepared using 5 µl of strepta-vidin magnetic beads from Thermo Fisher Scientific, washed twice with 1X DPBS from Thermo Fisher Scientific with 0.5% Tween-20 from Sigma-Aldrich, then resuspended in 50 µl of 1uM DNA linkers from Integrated DNA Technologies diluted in 1X DPBS from Thermo Fisher Scientific with 0.5% Tween-20 from Sigma-Aldrich. The beads were incubated with the DNA linkers for 30 minutes at 37°C with shaking at 600 RPM in ThermoMixer® C from Eppendorf, then washed twice with wash buffer and resuspended in 20 ng/µl of target azide peptide from GenScript Biotech diluted in wash buffer, then incubated overnight at room temperature with rotation. After washing the beads twice with binding buffer, the beads were resuspended in 50 µl of binding buffer, after which beads with different targets and identification fluorescent dyes were mixed.

Binding

10 µl of labeled proteins were mixed with 4 µl of beads mixture then incubated for 1 hour at 25°C with shaking at 600 RPM in ThermoMixer® C from Eppendorf. After binding, the beads were washed twice with wash buffer and resuspended in 200 µl of wash buffer. Finally, the fluorescence intensity on different bead species was measured Guava® easyCyte flow cytometer (Millipore).

Experimental Design

Each WW domain was tested against three peptides, each one corresponding to a different class of WW domains. The three peptides are given in Table 1. IVTT No expression control and no binding controls (peptides with different sequences; IAKLDEESILKQ and PALWKSKGAD) was performed for each experiment to confirm the specificity of WW domains binding to the target peptides. All binding experiments, including target and control peptide measurements, were conducted in triplicate to ensure reproducibility and statistical robustness.

Data availability

All generated and tested sequences are provided in the Supplementary Tables 2-7. All codes will be accessibles in *https://github.com/orgs/CoccoMonassonLab/* and *https://github.com/cossio/TransitionP*.

Acknowledgements

This work was supported by the grants ANR-19 Decrypted CE300021-01, ANR-24 CE15-Methaflu and, ANR-23 CE45-0034 ProDiGen.

Additional files

Supplementary material.

Additional information

Funding

Funder	Grant reference number	Author
Agence Nationale de la Recherche (ANR)	Decrypted CE300021-01	Simona Cocco
Agence Nationale de la Recherche (ANR)	ANR-23 CE45-0034 ProDiGen	Simona Cocco
Agence Nationale de la Recherche (ANR)	ANR-24 CE15- Methaflu	Simona Cocco

Author ORCID iDs

Jorge Fernandez-de-Cossio-Diaz: https://orcid.org/0000-0002-4476-805X
Pierre-Guillaume Brun: https://orcid.org/0009-0007-7536-9987
Remi Monasson: https://orcid.org/0000-0002-4459-0204
Marco Ribezzi-Crivellari: https://orcid.org/0000-0003-1217-5572
Simona Cocco: https://orcid.org/0000-0002-1852-7789

References

[1] **Murzin Alexey G** (1998) How far divergent evolution goes in proteins. *Current opinion in structural biology* **8**:380-387 https://doi.org/10.1016/s0959-440x(98)80073-0 | PubMed

[2] **Fanelli Alessandro Rossi,** Antonini Eraldo, Caputo Antonio (1964) Hemoglobin and myoglobin. *Advances in protein chemistry* **19**:73-222 https://doi.org/10.1016/s0065-3233(08)60189-8 | PubMed

[3] **Sudol Marius** (1996) Structure and function of the ww domain. *Progress in biophysics and molecular biology* **65**:113-132 https://doi.org/10.1016/s0079-6107(96)00008-9 | PubMed

[4] **Salah Z,** Aqeilan RI (2011) WW domain interactions regulate the hippo tumor suppressor pathway. *Cell Death & Disease* **2**:e172-e172 https://doi.org/10.1038/cddis.2011.53 | PubMed

[5] **Sudol Marius,** Sliwa Krzysztof, Russo Tommaso (2001) Functions of WW domains in the nucleus. *FEBS Letters* **490**:190-195 https://doi.org/10.1016/s0014-5793(01)02122-6 | PubMed

[6] **Sudol Marius,** Hunter Tony (2002) NeW wrinkles for an old domain. *Cell* **103**:1001-1004

[7] **Zarrinpar Ali,** Lim Wendell A (2000) Converging on proline: the mechanism of ww domain peptide recognition. *Nature structural biology* **7**:611-613 https://doi.org/10.1038/77891 | PubMed

[8] **Otte Livia,** Wiedemann Urs, Schlegel Brigitte, Pires Jose Ricardo, Beyermann Michael, Schmieder Peter, Krause Gerd, Volkmer-Engert Rudolf, Schneider-Mergener Jens, Oschkinat Hartmut (2003) Ww domain sequence activity relationships identified using ligand recognition propensities of 42 ww domains. *Protein Science* **12**:491-500 https://doi.org/10.1110/ps.0233203 | PubMed

[9] **Russ William P,** Lowery Drew M, Mishra Prashant, Yaffe Michael B, Ranganathan Rama (2005) Natural-like function in artificial ww domains. *Nature* **437**:579-583 https://doi.org/10.1038/nature03990 | PubMed

[10] **Shibata Masahiro,** Ham Kendall, Hoque Mohammad Obaidul (2018) A time for yap1: Tumorigenesis, immunosuppression and targeted therapy. *International journal of cancer* **143**:2133-2144 https://doi.org/10.1002/ijc.31561 | PubMed

[11] **Piccolo Stefano,** Dupont Sirio, Cordenonsi Michelangelo (2014) The biology of YAP/TAZ: Hippo signaling and beyond. *Physiological Reviews* **94**:1287-1312 https://doi.org/10.1152/physrev.00005.2014 | PubMed

[12] **Butterfield D. Allan,** Abdul Hafiz Mohmmad, Opii Wycliffe, Newman Shelley F., Joshi Gururaj, Ansari Mubeen Ahmad, Sultana Rukhsana (2006) Review: Pin1 in alzheimer's disease. *Journal of Neurochemistry* **98**:1697-1706 https://doi.org/10.1111/j.1471-4159.2006.03995.x | PubMed

[13] **Fotia Andrew B.,** Dinudom Anuwat, Shearwin Keith E., Koch Jan-Peter, Korbmacher Christoph, Cook David I., Kumar Sharad (2003) The role of individual nedd4-2 (KIAA0439) WW domains in binding and regulating epithelial sodium channels. *The FASEB Journal* **17**:70-72 https://doi.org/10.1096/fj.02-0497fje | PubMed

[14] **Espanel Xavier,** Sudol Marius (1999) A single point mutation in a group i ww domain shifts its specificity to that of group ii ww domains. *Journal of Biological Chemistry* **274**:17284-17289 https://doi.org/10.1074/jbc.274.24.17284 | PubMed

[15] **Macias Maria J.,** Wiesner Silke, Sudol Marius (2002) WW and SH3 domains, two different scaffolds to recognize proline-rich ligands. *FEBS Letters* **513**:30-37 https://doi.org/10.1016/s0014-5793(01)03290-2 | PubMed

[16] **Socolich Michael,** Lockless Steve W, Russ William P, Lee Heather, Gardner Kevin H, Ranganathan Rama (2005) Evolutionary information for specifying a protein fold. *Nature* **437**:512-518 https://doi.org/10.1038/nature03991 | PubMed

[17] **Khersonsky Olga,** Tawfik Dan Salah (2010) Enzyme promiscuity: A mechanistic and evolutionary perspective. *Annual Review of Biochemistry* **79**:471-505 https://doi.org/10.1146/annurev-biochem-030409-143718 | PubMed

[18] **Smith John Maynard** (1970) Natural selection and the concept of a protein space. *Nature* **225**:563-564 https://doi.org/10.1038/225563a0 | PubMed

[19] **Pauling Linus,** Zuckerkandl Emile, Henriksen Thormod, Lövstad Rolf (1963) Chemical Paleogenetics. Molecular "Restoration Studies" of Extinct Forms of Life. *Acta Chemica Scandinavica* **17**:9-16 https://doi.org/10.3891/acta.chem.scand.17s-0009

[20] **Wright Sewall** (1931) Evolution in mendelian populations. *Genetics* **16**:97 https://doi.org/10.1093/genetics/16.2.97 | PubMed

[21] **Poelwijk Frank J.,** Socolich Michael, Ranganathan Rama (2019) Learning the pattern of epistasis linking genotype and phenotype in a protein -Nature Communications. *Nat. Commun* **10**:1-11 https://doi.org/10.1038/s41467-019-12130-8 | PubMed

[22] **De Visser J. Arjan G.M.,** Krug Joachim (2014) Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics* **15**:480-490 https://doi.org/10.1038/nrg3744 | PubMed

[23] **Franke Jasper,** Klozer Alexander, De Visser J. Arjan G. M., Krug Joachim (2011) Evolutionary accessibility of mutational pathways. *PLoS Computational Biology* **7**:e1002134 https://doi.org/10.1371/journal.pcbi.1002134 | PubMed

[24] **Moulana Alief,** Dupic Thomas, Phillips Angela M, Chang Jeffrey, Nieves Serafina, Roffler Anne A, Greaney Allison J, Starr Tyler N, Bloom Jesse D, Desai Michael M (2022) Compensatory epistasis maintains ace2 affinity in sars-cov-2 omicron ba. 1. *Nature Communications* **13**:7011 https://doi.org/10.1038/s41467-022-34506-z | PubMed

[25] **Weinreich Daniel M.,** Delaney Nigel F., DePristo Mark A., Hartl Daniel L. (2006) Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**:111-114 https://doi.org/10.1126/science.1123539 | PubMed

[26] **Wu Nicholas C,** Dai Lei, Olson Anders, Lloyd-Smith James O, Sun Ren (2016) Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife* **5**:e16965 https://doi.org/10.7554/eLife.16965 | PubMed

[27] **Aguilar-Rodriguez Jose,** Payne Joshua L., Wagner Andreas (2017) A thousand empirical adaptive landscapes and their navigability. *Nature Ecology & Evolution* **1**:0045 https://doi.org/10.1038/s41559-016-0045 | PubMed

[28] **Nocedal Isabel,** Laub Michael T (2022) Ancestral reconstruction of duplicated signaling proteins reveals the evolution of signaling specificity. *eLife* **11**:e77346 https://doi.org/10.7554/eLife.77346 | PubMed

[29] **Brochier-Armanet Celine,** Madern Dominique (2021) Phylogenetics and biochemistry elucidate the evolutionary link between l-malate and l-lactate dehydrogenases and disclose an intermediate group of sequences with mix functional properties. *Biochimie* **191**:140-153 https://doi.org/10.1016/j.biochi.2021.08.004 | PubMed

[30] **Mauri Eugenio,** Cocco Simona, Monasson Riemi (2023) Mutational paths with sequence-based models of proteins: From sampling to mean-field characterization. *Phys. Rev. Lett* **130**:158402 https://doi.org/10.1103/physrevlett.130.158402 | PubMed

[31] **Chu AE,** Huang PS (2024) Sparks of function by de novo protein design. *Nature Biotechnology* **42**:203-215 https://doi.org/10.1038/s41587-024-02133-2 | PubMed

[32] **Zambaldi Vinicius,** La David, Chu Alexander E., Patani Harshnira, Danson Amy E., Kwan Tristan O. C., Frerix Thomas, Schneider Rosalia G., Saxton David, Thillaisundaram Ashok, *et al.* (2024) De novo design of high-affinity protein binders with alphaproteo. *arXiv* https://doi.org/10.48550/arXiv.2409.08022

[33] **Watson JL,** Bennett NR, et al. (2023) De novo design of protein structure and function with rfdiffusion. *Nature* **620**:1089-1100 https://doi.org/10.1038/s41586-023-06415-8 | PubMed

[34] **Russ William P.,** Figliuzzi Matteo, Stocker Christian, Barrat-Charlaix Pierre, Socolich Michael, Kast Peter, Hilvert Donald, Monasson Remi, Cocco Simona, Weigt Martin, *et al.* (2020) An evolution-based model for designing chorismate mutase enzymes. *Science* **369**:440-5 https://doi.org/10.1126/science.aba3304 | PubMed

[35] **Hawkins-Hooker Alex,** Depardieu Florence, Baur Sebastien, Couairon Guillaume, Chen Arthur, Bikard David (2021) Generating functional protein variants with variational autoencoders. *PLOS Computational Biology* **17**:1-23 https://doi.org/10.1371/journal.pcbi.1008736 | PubMed

[36] **Malbranke Cyril,** Bikard David, Cocco Simona, Monasson Remi, Tubiana Jerome (2023) Machine learning for evolutionary-based and physics-inspired protein design: Current and future synergies. *Current Opinion in Structural Biology* **80**:102571 https://doi.org/10.1016/j.sbi.2023.102571 | PubMed

[37] **Tubiana Jerome,** Cocco Simona, Monasson Remi (2019) Learning protein constitutive motifs from sequence data. *eLife* **8**:e39397 https://doi.org/10.7554/eLife.39397 | PubMed

[38] **Malbranke Cyril,** Rostain William, Depardieu Florence, Cocco Simona, Monasson Riemi, Bikard David (2023) Computational design of novel cas9 pam-interacting domains using evolution-based modelling and structural quality assessment. *PLOS Computational Biology* **19**:e1011621 https://doi.org/10.1371/journal.pcbi.1011621 | PubMed

[39] **Di Gioacchino Andrea,** Procyk Jonah, Molari Marco, Schreck John S, Zhou Yu, Liu Yan, Monasson Riemi, Cocco Simona, Sulc Petr (2022) Generative and interpretable machine learning for aptamer design and analysis of in vitro sequence selection. *PLoS computational biology* **18**:e1010561 https://doi.org/10.1371/journal.pcbi.1010561 | PubMed

[40] **Fernandez-De-Cossio-Diaz Jorge,** Hardouin Pierre, du Moutier Francois-Xavier Lyonnet, Di Gioacchino Andrea, Marchand Bertrand, Ponty Yann, Sargueil Bruno, Monasson Remi, Cocco Simona (2023) Designing molecular rna switches with restricted boltzmann machines. *bioRxiv* https://doi.org/10.1101/2023.05.10.540155

[41] **Jumper John,** Evans Richard, Pritzel Alexander, Green Tim, Figurnov Michael, Ronneberger Olaf, Tunyasuvunakool Kathryn, Bates Russ, Zidek Augustin, Potapenko Anna, *et al.* (2021) Highly accurate protein structure prediction with alphafold. *Nature* **596**:583-589 https://doi.org/10.1038/s41586-021-03819-2 | PubMed

[42] **Dauparas Justas,** Anishchenko Ivan, Bennett Nathaniel, Bai Hua, Ragotte Robert J, Milles Lukas F, Wicky Basile IM, Courbet Alexis, de Haas Rob J, Bethel Neville, *et al.* (2022) Robust deep learning-based protein sequence design using proteinmpnn. *Science* **378**:49-56 https://doi.org/10.1126/science.add2187 | PubMed

[43] **Tubiana Jerome,** Cocco Simona, Monasson Remi (2019) Learning protein constitutive motifs from sequence data. *eLife* **8**:e39397 https://doi.org/10.7554/eLife.39397 | PubMed

[44] **Mauri Eugenio,** Cocco Simona, Monasson Remi (2023) Direct vs. global transition paths in potts-like energy landscapes. *arXiv* https://doi.org/10.48550/arXiv.2304.03128

[45] **Keppler Antje,** Gendreizig Susanne, Gronemeyer Thomas, Pick Horst, Vogel Horst, Johnsson Kai (2003) A general method for the covalent labeling of fusion proteins with small molecules in vivo. *Nature Biotechnology* **21**:86-89 https://doi.org/10.1038/nbt765 | PubMed

[46] **Shimizu Yoshihiro,** Inoue Akio, Tomari Yukihide, Suzuki Tsutomu, Yokogawa Takashi, Nishikawa Kazuya, Ueda Takuya (2001) Cell-free translation reconstituted with purified components. *Nature Biotechnology* **19**:751-755 https://doi.org/10.1038/90802 | PubMed

[47] **Givan Alice Longobardi** (2001) *Flow Cytometry: First Principles* Wiley.

[48] **Kasanov Jeremy,** Pirozzi Gregorio, Uveges Albert J, Kay Brian K (2001) Characterizing class i ww domains defines key specificity determinants and generates mutant domains with novel specificities. *Chemistry & biology* **8**:231-241 https://doi.org/10.1016/s1074-5521(01)00005-9 | PubMed

[49] **Lin Zhijie,** Yang Zhou, Xie Ruiling, Ji Zeyang, Guan Kunliang, Zhang Mingjie (2019) Decoding ww domain tandem-mediated target recognitions in tissue growth and cell polarity. *eLife* **8**:e49439 https://doi.org/10.7554/eLife.49439 | PubMed

[50] **Miyazawa Sanzo,** Jernigan Robert L (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**:534-552 https://doi.org/10.1021/ma00145a039

[51] **Kato Yusuke,** Ito Mie, Kawai Kunji, Nagata Koji, Tanokura Masaru (2002) Determinants of ligand specificity in groups i and iv ww domains as studied by surface plasmon resonance and model building. *Journal of Biological Chemistry* **277**:10173-10177 https://doi.org/10.1074/jbc.m110490200 | PubMed

[52] **Sudol Marius,** Chen Henry I., Bougeret Cecile, Einbond Aaron, Bork Peer (1995) Characterization of a novel protein-binding module — the WW domain. *FEBS Letters* **369**:67-71 https://doi.org/10.1016/0014-5793(95)00550-s | PubMed

[53] **de Visser J. Arjan G.M.,** Krug Joachim (2014) Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics* **15**:480-490 https://doi.org/10.1038/nrg3744 | PubMed

[54] **Kauffman Stuart,** Levin Simon (1987) Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology* **128**:11-45 https://doi.org/10.1016/s0022-5193(87)80029-2 | PubMed

[55] **Kimura Motoo** (1983) *The neutral theory of molecular evolution* Cambridge University Press.

[56] **Das Suman G,** Direito Susana Ol, Waclaw Bartlomiej, Allen Rosalind J, Krug Joachim (2020) Predictable properties of fitness landscapes induced by adaptational tradeoff's. *eLife* **9**:e55155 https://doi.org/10.7554/eLife.55155 | PubMed

[57] **Kingman J. F. C.** (1978) A simple model for the balance between selection and mutation. *Journal of Applied Probability* **15**:1-12 https://doi.org/10.2307/3213231

[58] **Di Bari Leonardo,** Bisardi Matteo, Cotogno Sabrina, Weigt Martin, Zamponi Francesco (2024) Emergent time scales of epistasis in protein evolution. *Proceedings of the National Academy of Sciences* **121**:e2406807121 https://doi.org/10.1073/pnas.2406807121 | PubMed

[59] **Huot Marian,** Wang Dianzhuo, Shakhnovich Eugene, Monasson Rémi, Cocco Simona (2025) Constrained evolutionary funnels shape viral immune escape. *bioRxiv* https://doi.org/10.1101/2025.10.26.684604

[60] **Kurth Torsten,** Ullmann Dirk, Jakubke Hans-Dieter, Hedstrom Lizbeth (1997) Converting trypsin to chymotrypsin: Structural determinants of s1' specificity. *Biochemistry* **36**:10098 https://doi.org/10.1021/bi970937l | PubMed

[61] **Korendovych Ivan V.** (2018) *Rational and Semirational Protein Design, volume 1685 of Methods in Molecular Biology* New York, NY: Springer New York. pp. 15-23

[62] **Venekei Istvan,** Szilagyi Laszlo, Graf Laszlo, Rutter William J. (1996) Attempts to convert chymotrypsin to trypsin. *FEBS Letters* **379**:143-147 https://doi.org/10.1016/0014-5793(95)01484-5 | PubMed

[63] **Yang Kevin K.,** Wu Zachary, Arnold Frances H. (2019) Machine-learning-guided directed evolution for protein engineering. *Nature Methods* **16**:687-694 https://doi.org/10.1038/s41592-019-0496-6 | PubMed

[64] **Listov Dina,** Goverde Casper A., Correia Bruno E., Fleishman Sarel Jacob (2024) Opportunities and challenges in design and optimization of protein function. *Nature Reviews Molecular Cell Biology* **25**:639-653 https://doi.org/10.1038/s41580-024-00718-y | PubMed

[65] **Casari Georg,** Sander Chris, Valencia Alfonso (1995) A method to predict functional residues in proteins. *Nature structural biology* **2**:171 https://doi.org/10.1038/nsb0295-171 | PubMed

[66] **Halabi Najeeb,** Rivoire Olivier, Leibler Stanislas, Ranganathan Rama (2009) Protein sectors: evolutionary units of three-dimensional structure. *Cell* **138**:774-786 https://doi.org/10.1016/j.cell.2009.07.038 | PubMed

[67] **Weigt Martin,** White Robert A, Szurmant Hendrik, Hoch James A, Hwa Terence (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences* **106**:67-72 https://doi.org/10.1073/pnas.0805923106 | PubMed

[68] **Cocco Simona,** Feinauer Christoph, Figliuzzi Matteo, Monasson Remi, Weigt Martin (2018) Inverse statistical physics of protein sequences: a key issues review. *Reports on Progress in Physics* **81**:032601 https://doi.org/10.1088/1361-6633/aa9965 | PubMed

[69] **Wu Nicholas C,** Dai Lei, Olson Anders, Lloyd-Smith James O, Sun Ren (2016) Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife* **5**:e16965 https://doi.org/10.7554/eLife.16965 | PubMed

[70] **Tokuriki Nobuhiko,** Stricher Francois, Serrano Luis, Tawfik Dan S. (2008) How protein stability and new functions trade off. *PLoS Computational Biology* **4**:e1000002 https://doi.org/10.1371/journal.pcbi.1000002 | PubMed

[71] **Bloom Jesse D.,** Labthavikul Sy T., Otey Christopher R., Arnold Frances H. (2006) Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences* **103**:5869-5874 https://doi.org/10.1073/pnas.0510098103 | PubMed

[72] **Jager Marcus,** Zhang Yan, Bieschke Jan, Nguyen Houbi, Dendle Maria, Bowman Marianne E, Noel Joseph P, Gruebele Martin, Kelly Jeffery W (2006) Structure-function-folding relationship in a ww domain. *Proceedings of the National Academy of Sciences* **103**:10648-10653 https://doi.org/10.1073/pnas.0600511103 | PubMed

[73] **Wheeler Lucas C,** Harms Michael J (2021) Were Ancestral Proteins Less Specific?. *Molecular Biology and Evolution* **38**:2227-2239 https://doi.org/10.1093/molbev/msab019 | PubMed

[74] **Spence Matthew A.,** Kaczmarski Joe A., Saunders Jake W., Jackson Colin J. (2021) Ancestral sequence reconstruction for protein engineers. *Current Opinion in Structural Biology* **69**:131-141 https://doi.org/10.1016/j.sbi.2021.04.001 | PubMed

[75] **Prakinee Kridsadakorn,** Phaisan Suppalak, Kongjaroon Sirus, Chaiyen Pimchai (2024) Ancestral Sequence Reconstruction for Designing Biocatalysts and Investigating their Functional Mechanisms. *JACS Au* **4**:4571-4591 https://doi.org/10.1021/jacsau.4c00653 | PubMed

[76] **Laine Elodie,** Karami Yasaman, Carbone Alessandra (2019) Gemme: a simple and fast global epistatic model predicting mutational effects. *bioRxiv* 543587 https://doi.org/10.1101/543587

[77] **Subbanna Mythili S.,** Winters Matthew J., Davey Norman E., Pryciak Peter M. (2025) A quantitative intracellular peptide-binding assay reveals recognition determinants and context dependence of short linear motifs. *Journal of Biological Chemistry* **301**:108225 https://doi.org/10.1016/j.jbc.2025.108225 | PubMed

[78] **Cocco Simona,** Posani Lorenzo, Monasson Rémi (2019) Functional couplings from sequence and mutational data. In Preparation.

[79] **Rotrattanadumrong Rachapun,** Yokobayashi Yohei (2022) Experimental exploration of a ribozyme neutral network using evolutionary algorithm and deep learning. *Nature Communications* **13**:4847 https://doi.org/10.1038/s41467-022-32538-z | PubMed

[80] **Yang Jason,** Lal Ravi G., Bowden James C., Astudillo Raul, Hameedi Mikhail A., Kaur Sukhvinder, Hill Matthew, Yue Yisong, Arnold Frances H. (2025) Active learning-assisted directed evolution. *Nature Communications* **16**:714 https://doi.org/10.1038/s41467-025-55987-8 | PubMed

[81] **Fischer Asja,** Igel Christian (2012) An introduction to restricted Boltzmann machines. In: Iberoamerican Congress on Pattern Recognition. Springer. pp. 14-36 https://doi.org/10.1007/978-3-642-33275-3_2

[82] **Tieleman Tijmen** (2008) Training restricted Boltzmann machines using approximations to the likelihood gradient. In: ICML '08: Proceedings of the 25th International Conference on Machine learning. pp. 1064-1071 https://doi.org/10.1145/1390156.1390290

[83] **Tubiana Jerome** (2018) Restricted Boltzmann machines : from compositional representations to protein sequence analysis. PhD thesis, Universite Paris sciences et lettres, Paris, France. https://doi.org/10.70675/a6d497cfz68a4z42a8za422zbbc81f218ea5

[84] **Tubiana Jerome** (2018) Probabilistic graphical models (pgm). https://github.com/jertubiana/PGM

[85] **Minh Bui Quang,** Schmidt Heiko A, Chernomor Olga, Schrempf Dominik, Woodhams Michael D, Von Haeseler Arndt, Lanfear Robert (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**:1530-1534 https://doi.org/10.1093/molbev/msaa015 | PubMed

[86] **Gascuel S. Q. Le and O.** (2008) An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution* **25**:1307-1320 https://doi.org/10.1093/molbev/msn067 | PubMed

[87] **Yang Ziheng** (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution* **39**:306-314 https://doi.org/10.1007/bf00160154 | PubMed

[88] **Aadland Kelsey,** Pugh Charles, Kolaczkowski Bryan (2019) High-Throughput Reconstruction of Ancestral Protein Sequence, Structure, and Molecular Function. In: Sikosek T. (Ed). *Computational Methods in Protein Evolution* **1851** New York, New York, NY: Springer. pp. 135-170 https://doi.org/10.1007/978-1-4939-8736-8_8 | PubMed

[89] **Finn Robert D,** Clements Jody, Eddy Sean R (2011) Hmmer web server: interactive sequence similarity searching. *Nucleic Acids Research* **39**:W29-W37 https://doi.org/10.1093/nar/gkr367 | PubMed

[90] **Kaminski Kamil,** Ludwiczak Jan, Pawlicki Kamil, Alva Vikram, Dunin-Horkawicz Stanislaw (2023) pLM-BLAST: Distant homology detection based on direct comparison of sequence representations from protein language models. *Bioinformatics* **39**:btad579 https://doi.org/10.1093/bioinformatics/btad579 | PubMed

[91] **Katoh K.** (2002) MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**:3059-3066 https://doi.org/10.1093/nar/gkf436 | PubMed

[92] **Kalyaanamoorthy Subha,** Minh Bui Quang, Wong Thomas K F, Von Haeseler Arndt, Jermiin Lars S (2017) ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**:587-589 https://doi.org/10.1038/nmeth.4285 | PubMed

[93] **Hesselberth Jay R,** Miller John P, Golob Anna, Stajich Jason E, Michaud Gregory A, Fields Stanley (2006) Comparative analysis of Saccharomyces cerevisiaeWW domains and their interacting proteins. *Genome Biology* **7**:R30 https://doi.org/10.1186/gb-2006-7-4-r30 | PubMed

[94] **Eme L.,** Sharpe S. C., Brown M. W., Roger A. J. (2014) On the Age of Eukaryotes: Evaluating Evidence from Fossils and Molecular Clocks. *Cold Spring Harbor Perspectives in Biology* **6**:a016139-a016139 https://doi.org/10.1101/cshperspect.a016139 | PubMed

[95] **Gaucher Eric A.,** jan Sridhar Govindara, Ganesh Omjoy K. (2008) Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* **451**:704-707 https://doi.org/10.1038/nature06510 | PubMed

J.Mistry, Chuguransky S., Williams L., Qureshi M., Salazar G.A., Sonnhammer E.L.L., Tosatto S.C.E., Paladin L., Raj S., Richardson L.J., *et al.* (2021) The WW domain is a protein module with two highly conserved tryptophans that binds proline-rich peptide motifs in vitro. PFAM. ID PF00397 https://www.ebi.ac.uk/interpro/entry/pfam/PF00397/

Peer reviews

Reviewer #1 (Public review):

Summary:

The authors aim to study mutational paths connecting WW domains with different binding specificities. Their approach combines an unsupervised sequence generative model based on RBMs with a path-sampling algorithm. The key result is that most intermediate sequences along the designed transition paths retain measurable binding activity in wet-lab assays, whereas paths containing the same mutations introduced in a randomized order are largely non-functional. This difference is attributed to epistatic interactions captured by the RBM model.

Strengths:

Exploring mutational paths in high-dimensional protein sequence space is a challenging problem. The computational framework used here is state-of-the-art and is strengthened by systematic experimental characterization of binding activity. The study is comprehensive in scope, including multiple transition paths both within and across WW specificity classes, and the integration of modeling with high-throughput experimental validation is a clear strength.

Weaknesses:

A major concern is whether the stated goal of specificity switching is fully achieved. Along the sampled transition paths, most intermediate variants appear to retain specificity close to either the initial or the final class, rather than exhibiting gradually shifting specificity. For example, in Figure 4G (Class I to Class II/III), binding appears largely binary, with intermediates behaving similarly to one of the endpoints. A similar pattern is observed in Figure 3H for the Class I to Class IV transition, where binding responses are close to 0 or 1. In this sense, the specificity-switching objective is only partially realized by assigning two endpoints with different specificity. This raises a broader conceptual question: is it possible that different WW specificities evolved from a common ancestor without passing through intermediates that exhibit mixed or intermediate specificity? If so, then inferring specificity-

switching pathways purely from extant natural sequences may be fundamentally challenging.

https://doi.org/10.7554/eLife.110491.1.sa1

Reviewer #2 (Public review):

This is an extremely important work that shows how one can use generative models to construct specificity-switching mutational paths in complex fitness landscapes. The experimental evidence is very clear, and the theoretical tools are innovative.

The work will likely have a deep impact on future research aimed at understanding how evolution navigates fitness landscapes as well as reconstructing ancestral sequences.

The manuscript is extremely clear and well written, the experimental evidence is strong, and the methods are clearly described, so I do not have major issues to raise. A few minor issues are listed below.

(1) I consider the WW domain as an 'easy' case from the point of view of generative modelling. The domain is rather short, epistatic effects are not very strong (e.g. Boltzmann learning usually converges very quickly to a very paramagnetic state), and the resulting models are well interpretable (e.g. the hidden units of the RBM correlate well with subclasses).

This is not always (not often?) the case, however. In more complex proteins, the learning procedures can be slower and the resulting models less interpretable. Just for completeness, perhaps the authors could comment on the generality of the results and what they would expect for other systems based on their experience.

(2) In Section 3.3, the authors say that direct paths connecting Class I and Class IV behave similarly to indirect paths, despite having lower scores according to the RBM. How generic is this? Does it also happen for other classes? This might be an important point to address, as direct paths are easier to sample.

(3) The path shown in Figure 4 goes through a region of non-functionality around sequences 18-19. It seems that the sample path is basically exploring the functional regions for Class I and Class II/III separately, trying to approach the other class, but then it can't really make the switch.

By contrast, the path going from Class I to Class IV seems able to perform the functional switch in a single step (20-21) without losing too much of the function.

Perhaps the authors could better comment on this? Is this a limitation of the sampling method, or a fundamental biological fact?

(4) On page 12, it is stated that the temperature was chosen to 1/3 to maximize the score. This is important and should be mentioned earlier (I didn't notice it until that point).

(5) On page 13, it is stated that: "However, the scores of the ancestral sequences along the phylogenetic pathways assigned by the RBM are significantly lower than the ones of the RBM-designed sequences. This result is expected as ASR reconstruction does not take into account epistasis, differently from RBM, and we expect ASR sequences to generally be of lesser quality."

I was very surprised by this result. My own experience with ASR shows that, on the contrary, sequences found by ASR (via maximum likelihood) tend to have high scores in the (R)BM, and tend to be more stable than extant sequences. I attribute this to the fact that ASR typically finds a "consensus" sequence that maximizes the contribution to the score coming from the

fields (the profile), which is typically dominant over the epistatic signal, resulting in a bigger score. Maybe the authors did not use maximum likelihood in the ASR? Some clarification might be useful here.

https://doi.org/10.7554/eLife.110491.1.sa0

Author response:

> ***Public Reviews:***
>
> ***Reviewer #1:***
>
> *Summary:*
>
> *The authors aim to study mutational paths connecting WW domains with different binding specificities. Their approach combines an unsupervised sequence generative model based on RBMs with a path-sampling algorithm. The key result is that most intermediate sequences along the designed transition paths retain measurable binding activity in wet-lab assays, whereas paths containing the same mutations introduced in a randomized order are largely nonfunctional. This difference is attributed to epistatic interactions captured by the RBM model.*
>
> *Strengths:*
>
> *Exploring mutational paths in high-dimensional protein sequence space is a challenging problem. The computational framework used here is state-of-the-art and is strengthened by systematic experimental characterization of binding activity. The study is comprehensive in scope, including multiple transition paths both within and across WW specificity classes, and the integration of modeling with high-throughput experimental validation is a clear strength.*
>
> *Weaknesses:*
>
> *A major concern is whether the stated goal of specificity switching is fully achieved. Along the sampled transition paths, most intermediate variants appear to retain specificity close to either the initial or the final class, rather than exhibiting gradually shifting specificity. For example, in Figure 4G (Class I to Class II/III), binding appears largely binary, with intermediates behaving similarly to one of the endpoints. A similar pattern is observed in Figure 3H for the Class I to Class IV transition, where binding responses are close to 0 or 1. In this sense, the specificityswitching objective is only partially realized by assigning two endpoints with different specificity. This raises a broader conceptual question: is it possible that different WW specificities evolved from a common ancestor without passing through intermediates that exhibit mixed or intermediate specificity? If so, then inferring specificity-switching pathways purely from extant natural sequences may be fundamentally challenging.*

This is a key question, which was one of the original motivations of our work. Both hypothesis of 'abrupt switches' (punctuated equilibria, corresponding to distinct specificities) and more gradual changes (smooth transition, through intermediate that exhibit mixed or intermediate specificity) are possible.

Many natural specificity-switching events have probably resulted from the need to adapt to environmental change and selection for a different specificity, which can be compatible with an abrupt change in specificity. Others may reflect the gradual evolution of promiscuous ancestral sequences to more specialized ones, loosing cross-reactivity. A molecular mechanism that could allow abrupt switching is gene duplication, a frequent mechanism for WW domain diversification, beyond standard mutational-driven evolution processes.

As for the specificity-switching paths for WW domains found in this work, the presence of weakly responsive cross-reactive intermediates along the designed paths for I<->IV, and their absence in the I<->II path, suggests that designing promiscuous domains is hard (see also related response to point 3 of Reviewer 2) and generally not selected by natural evolution (as seen from the clear clustering of extant proteins in different specificity classes).

For a small domain such as WW, mutations that favor some specificity classes are known to have detrimental effects on fundamental properties, such as folding kinetics and stability, see Ref [72]. It is possible that larger, less constrained protein domains could allow for more crossreactive variants and smoother specifity switching. However, experiments on fluorescent proteins looking for interpolation between two wave-lengths have shown that the switch was abrupt [Poelwijk et al. Nature Communications (2019)].

Our scope was to achieve a functional switch (imposed by the two extant end-points) through a path of designed, functional intermediates and to correctly predict, with our RBM model, the location of the specificity transition and of the cross-reactivity region (which we expected only along the I-IV path). This scope was successfully reached as demonstrated by experiments.

> ***Reviewer #2:***
>
> *This is an extremely important work that shows how one can use generative models to construct specificity-switching mutational paths in complex fitness landscapes. The experimental evidence is very clear, and the theoretical tools are innovative.*
>
> *The work will likely have a deep impact on future research aimed at understanding how evolution navigates fitness landscapes as well as reconstructing ancestral sequences.*
>
> *The manuscript is extremely clear and well written, the experimental evidence is strong, and the methods are clearly described, so I do not have major issues to raise. A few minor issues are listed below.*
>
> *(1) I consider the WW domain as an 'easy' case from the point of view of generative modelling. The domain is rather short, epistatic effects are not very strong (e.g. Boltzmann learning usually converges very quickly to a very paramagnetic state), and the resulting models are well interpretable (e.g. the hidden units of the RBM correlate well with subclasses).*
>
> *This is not always (not often?) the case, however. In more complex proteins, the learning procedures can be slower and the resulting models less interpretable. Just for completeness, perhaps the authors could comment on the generality of the results and what they would expect for other systems based on their experience.*

We agree with Reviewer 2 that WW sequences are short and simple to handle from a computational point of view, and was chosen for this reason to test the design of full mutational paths (after having benchmarked it to lattice-protein models, see Refs. [30] and [44]). Our work gives additional support to the effectiveness of generative models learned from sequence data. This said, from a biological point of view, WW is a highly constrained domain, see comment by Reviewer 1 above and our answer.

In longer and more complex proteins, we expect it will be more difficult to disentangle specificityswitching latent units, see Fernandez-de-Cossio-Diaz et al., Physical Review X 2023 for a discussion and a possible computational approach to this issue. Notice that, while relating the latent units to specificity classes was convenient, it was not used to generate the paths themselves. Therefore, we believe that our method is quite robust and easily generalizable to applications to more complex and longer proteins. As an illustration, we have recently used it to sample viral trajectories (more precisely, variants of the Receptor

Binding Domain of the SARSCoV-2 spike protein) capable of escaping antibody recognition, see Huot et al., PNAS 2026. In this recent work, we projected the paths onto the principal antigenic space, defined by the top two Principal Components of the viral variant binding affinities to 32 antibodies. In this representation, sampled paths displayed trends similar to natural paths, drawn from the sequences sampled during the pandemics. This finding supports the applicability and interpretation of our method for more complex proteins.

> *(2) In Section 3.3, the authors say that direct paths connecting Class I and Class IV behave similarly to indirect paths, despite having lower scores according to the RBM. How generic is this? Does it also happen for other classes? This might be an important point to address, as direct paths are easier to sample.*

We think that this finding, true for paths connecting classes I and IV, is not general. In a previous paper we have benchmarked our path-designing approach on simple models of insilico lattice proteins and shown that indirect path led to gains in the overall fitness (computed according with the ground-truth model) [Mauri, Cocco, Monasson, Physical Review E 2023, fig. 9-12].

In general, we would expect that indirect paths could explore alternative mutations, important to compensate for transitory destabilizing mutations that could occur along the path. We speculate that these stabilizing mutations happen for non-direct paths at its extremity near class-I wildtype. A slightly decrease in binding response to peptide C1 for direct path is nevertheless observed (see Suppl Table 4), but our experimental detection, focused on binding response, is not tailored to directly detect a difference in stability. When approaching the class-IV anchoring point, we observe that paths interpolating between classes I and IV are very constrained and show limited diversity, going through a funnel in sequence space corresponding to the direct path. We agree with Reviewer 2 that a more exhaustive comparison with direct paths would be interesting, and will add a sentence in conclusion.

> *(3) The path shown in Figure 4 goes through a region of non-functionality around sequences 1819. It seems that the sample path is basically exploring the functional regions for Class I and Class II/III separately, trying to approach the other class, but then it can't really make the switch.*
>
> *By contrast, the path going from Class I to Class IV seems able to perform the functional switch in a single step (20-21) without losing too much of the function.*
>
> *Perhaps the authors could better comment on this? Is this a limitation of the sampling method, or a fundamental biological fact?*

Class I to Class IV paths and Class I to Class II paths fundamentally differ because the binding pocket in Class I WW domains is different from the one of Class IV WWs, while Classes I and II/III share the same binding region. This important difference may explain why class I specificity can switch to class IV specificity (steps 20-21), without completely loosing affinity to the peptide of class I. To investigate if the two binding regions are really independent or not, we have tested some additional specific mutations along the I-IV mutational paths. In our attempts to engineer cross-reactivity, we have observed that it is important to substantially lower affinity to class I peptide to acquire class IV specificity, in agreement with previous studies [72]. Moreover, the I to IV path seems to go through a funnel-like part in the region with no natural sequences, with the same transition intermediates obtained in several designed paths. This indicates that the Class I to Class IV functional switch is more constrained than the Class I to II switch. Let us also emphasize that our assessment of class specificity is based on one peptide for each class. It would be interesting to test multiple WW-binding peptides with similar biochemical properties to acquire a more complete view of the specificities.

> *(4) On page 12, it is stated that the temperature was chosen to 1/3 to maximize the score. This is important and should be mentioned earlier (I didn't notice it until that point).*

Section 3.5 explains that RBM samples can be biased, by lowering the sampling temperature to 1/3 to obtain high-scores sequences, which are more likely to be functional as proven in [Russ et al., Science 2020]. We acknowledge (as also noted by Reviewer 1) that this section comes at the end of the manuscript, while differences in scores along the path are shown before, so the discussion of this important point is somewhat delayed. We will add a sentence earlier in Results to explain this point.

> *(5) On page 13, it is stated that: "However, the scores of the ancestral sequences along the phylogenetic pathways assigned by the RBM are significantly lower than the ones of the RBMdesigned sequences. This result is expected as ASR reconstruction does not take into account epistasis, differently from RBM, and we expect ASR sequences to generally be of lesser quality."*
>
> *I was very surprised by this result. My own experience with ASR shows that, on the contrary, sequences found by ASR (via maximum likelihood) tend to have high scores in the (R)BM, and tend to be more stable than extant sequences. I attribute this to the fact that ASR typically finds a "consensus" sequence that maximizes the contribution to the score coming from the fields (the profile), which is typically dominant over the epistatic signal, resulting in a bigger score. Maybe the authors did not use maximum likelihood in the ASR? Some clarification might be useful here.*

We agree with Reviewer 2 that the consensus sequence is an atypical sequence for an independent model with a large RBM score. We will update Figure 5 of the manuscript to show that this is also happening in our case.

We use Maximum Likelihood in ASR but our ASR path corresponds to all internal nodes of the reconstructed tree joining the two extant sequences, not only to the most ancestral node. Overall, the ancestral sequences along the ASR paths are different from the consensus sequence (mean identity of 76% and 60% respectively). The most ancestral nodes in the paths are also different from the consensus having 81% (paths between type I and IV domains) or 54%(paths between type I and II/III domains) similarity, and an RBM score of -21, or -58, respectively. We agree that some ASR internal-node sequence have a higher score than the natural wild-types (extant sequences). This is shown in Fig. 6: several points have larger RBM score than the two anchoring points at the extremities of the path, possibly due to the fact that natural sequences are not always the most stable ones. As discussed in conclusion, ASR nodes have moreover generally better scores than the sequences obtained by sampling an independent model. Phylogenetic reconstruction implicitly takes into account some degree of co-variation between sites in natural sequences, as shown by the success of the use of the phylogenetic distance of a mutated sequence to the wild-type for predicting the fitness effect of these mutations [Laine, Mol. Biol. Evol. 2019].

To better show this effect we will update Figure 6, reporting also the scores of the « scrambled » sequences, which do not respect potential epistasis extracted by the RBM. It appears that ASR sequences generally have better scores than the scrambled sequences, and lower than RBM sequences (sampled at T=1/3). RBM models takes into account multiple-residues correlations, which could contribute to reaching better scores than ASR and BM models. Ongoing studies on larger proteins show that the score of sequences sampled from ASR reconstruction, including the Maximum Likelihood one, can still be improved according to the RBM score by a few mutations consistent with the ASR posterior probabilities (unpublished).

Mistakes in the reference list will be amended in the updated version.

https://doi.org/10.7554/eLife.110491.1.sa3