Mutational Paths with Sequence-Based Models of Proteins: From Sampling to Mean-Field Characterization

Eugenio Mauri[⬤],[*] Simona Cocco[⬤], and Rémi Monasson[⬤][†]

Laboratory of Physics of the Ecole Normale Supérieure, CNRS UMR 8023 and PSL Research, Sorbonne Université, 24 rue Lhomond, 75231 Paris cedex 05, France

Identifying and characterizing mutational paths is an important issue in evolutionary biology, with potential applications to bioengineering. We here propose an algorithm to sample mutational paths, which we benchmark on exactly solvable models of proteins in silico, and apply to data-driven models of natural proteins learned from sequence data with restricted Boltzmann machines. We then use mean-field theory to characterize paths for different mutational dynamics of interest, and to extend Kimura's estimate of evolutionary distances to sequence-based epistatic models of selection.

Introduction.—Obtaining proteins with controlled properties, such as stability, binding affinity, and specificity is a central goal in bioengineering [1]. Over the past years, much progress on design was made using data-driven models, intended to capture the relation between protein sequences and functionalities. In particular, unsupervised machine-learning approaches such as Boltzmann machines (BM) or variational auto-encoders trained on homologous sequence data (defining a protein family) were able to design new proteins with functionalities comparable to natural proteins [2,3]. By comparison, the (even) harder problem of designing paths of sequences, interpolating between two homologous proteins has received little attention, see, however, Ref. [4]. Yet solving this problem would shed light on the navigability of the sequence landscape [5] and on how functional specificity, such as binding to distinct substrates could have emerged from ancestral, promiscuous proteins in the course of evolution [6]. In turn, it could help design new proteins interpolating between functional classes.

While various methods exist for building transition paths between the minima of a multidimensional continuous landscape [7,8], dealing with discrete configurations requires the development of specific procedures [9]. We hereafter propose a Monte Carlo algorithm to sample mutational paths in protein landscapes, e.g., obtained by restricted Boltzmann machines trained on sequence data. We first benchmark our sampling procedure on an exactly solvable model of lattice proteins [10], and demonstrate its capability to find high-quality paths between two proteins belonging to different subfamilies. We then apply our algorithm to the WW domain, a binding module involved in the regulation of protein complexes [11,12]. The functionality of the sequences along the paths is validated with structure (ligand + protein)-informed software [13]. Last of all we derive a mean-field characterization of paths,

tailored to the mutational dynamics of interest. This mean-field theory allows us to efficiently estimate evolutionary distances in the presence of strong epistatis in the selection process, which is not possible with profile models at the basis of most phylogenetic studies [14].

Definition and sampling of mutational paths.—We assume the sequence landscape is modeled through a probability distribution $P_{\text{model}}(\mathbf{v})$ over amino-acid sequences $\mathbf{v}$ of length N. Informally speaking, P_{model} quantifies the probability that $\mathbf{v}$ is a member of the protein family of interest, i.e., shares its common structural and functional properties, and can be learned from homologous sequence data [15,16]. For natural protein families, exact expressions for P_{model} are not available, but approximate distributions can be inferred from multisequence alignments (MSA) using unsupervised learning techniques. Previous works have shown that the inferred P_{model} can serve as a proxy for the protein fitness [17–20].

Hereafter, we use restricted Boltzmann machines (RBM) [21], a class of generative models based on two-layer graphs [22]. RBMs define a joint probability distribution of the protein sequence $\mathbf{v}$ (carried by the visible layer) and of its M-dimensional latent representation $\mathbf{h}$ (present on the hidden layer) as

$$P_{\text{RBM}} \propto \exp \left(\sum_{i=1}^{N} g_i(v_i) + \sum_{\mu=1}^{M} h_\mu I_\mu(\mathbf{v}) - \sum_{\mu=1}^{M} \mathcal{U}_\mu(h_\mu) \right), \quad (1)$$

where $I_\mu(\mathbf{v}) = \sum_i w_{i,\mu}(v_i)$ is the input to hidden unit μ. The g_i's and $\mathcal{U}_\mu$'s are local potentials acting on, respectively, visible and hidden units, and the $w_{i\mu}$'s are the interactions between the two layers. They are learned by maximizing the marginal probabilities $P_{\text{model}}(\mathbf{v}) = \int d\mathbf{h}\, P_{\text{RBM}}(\mathbf{v}, \mathbf{h})$ over the sequences $\mathbf{v}$ in a multisequence alignment of the family. While other unsupervised

procedures providing approximate P_{model} can be used, such as direct coupling analysis [15,16], RBMs offer a convenient way to monitor the changes in sequences along mutational paths, as we will see below.

We consider mutational paths of T steps, $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{T-1}\}$, anchored at their extremities defined by the sequences $\mathbf{v}_{\text{start}}$ and $\mathbf{v}_{\text{end}}$. The probability of a path reads

$$\mathcal{P}[\mathcal{V}|\mathbf{v}_{\text{start}}, \mathbf{v}_{\text{end}}] \propto \prod_{t=1}^{T-1} P_{\text{model}}(\mathbf{v}_t) \times \pi(\mathbf{v}_{\text{start}}, \mathbf{v}_1)$$

$$\times \prod_{t=1}^{T-2} \pi(\mathbf{v}_t, \mathbf{v}_{t+1}) \times \pi(\mathbf{v}_{T-1}, \mathbf{v}_{\text{end}}), \qquad (2)$$

where the "transition" factor $\pi(\mathbf{v}, \mathbf{v}')$ increases with the similarity between the sequences $\mathbf{v}, \mathbf{v}'$. In practice we choose $\pi = 1$ if the two sequences are identical, $e^{-\Lambda}$ if they differ by one mutation (with $\Lambda > 0$), and 0 if they are two or more mutations apart. This choice generates "continuous" paths, along which successive sequences differ by one mutation at most. Other choices for π, more plausible from an evolutionary point of view will be introduced below.

The probability $\mathcal{P}(\mathcal{V})$ can be sampled as follows. Starting from a path $\mathcal{V}^0$, we randomly pick up an intermediate sequence $\mathbf{v}_t$ and attempt at mutating one amino acid, under the constraint that the Hamming distance of the trial sequence $\mathbf{v}'$ with $\mathbf{v}_{t-1}$ and $\mathbf{v}_{t+1}$ is at most 1. The mutation is then rejected or accepted, i.e., $\mathbf{v}_t \leftarrow \mathbf{v}'$ according to detailed balance. To improve the quality of the sampled mutational paths we introduce a fictitious inverse temperature β and resort to simulated annealing. We then sample paths from $\mathcal{P}[\mathcal{V}]^\beta$, where β is initially very small and is progressively ramped up to some target value. The complete procedure and the proof of detailed balance are given in Supplemental Material, Sec. 1 [23].

Benchmarking mutational path sampling on in silico proteins.—We benchmark the performances of our MC procedure on a model of lattice proteins (LP) [10,33]. In LP, sequences of 27 amino acids may fold into $\simeq 10^5$ different self-avoiding conformations going through the nodes of a $3 \times 3 \times 3$ cubic lattice. The sequence landscape associated to a structure $\mathbf{S}$ [Fig. 1(a)] is defined by the probability $p_{\text{nat}}(\mathbf{v}|\mathbf{S})$ that a sequence $\mathbf{v}$ has $\mathbf{S}$ as its native fold; p_{nat} can be exactly computed from the energies of interactions between adjacent amino acids, see Supplemental Material, Sec. 2 [23] for details.

We first generate many sequences $\mathbf{v}$ with high p_{nat} values for the fold $\mathbf{S}$ of Figs. 1(b) and 1(c) following the procedure of [18]. We next compute the top two principal components (PC) of these sequence data using one-hot encoding: PC1 corresponds to an extended electrostatic mode, and PC2 identifies possible Cys-Cys bridges [Figs. 1(d) and 1(e)]. Projecting the sequences onto these two PCs reveals two subfamilies separated along PC1 [Fig. 1(a)], associated
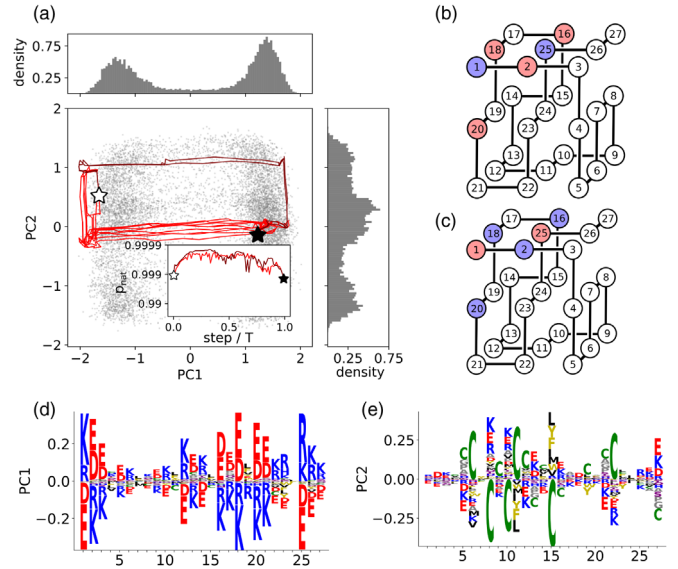

FIG. 1. Mutational paths for lattice proteins, joining sequences white star = DRGIQCLAQMFEKEMRKKKRRKCYLECD and black star = RECCAVCHQRFKDKIDEDYEDAWLKCN belonging to the family with structure shown in (b) and (c). Red and blue colors, respectively, correspond to negatively and positively charged amino acids. Cysteine is denoted by a green C. (a) Projections of 10^4 LP sequences in the family (gray dots) along the top two PC of their correlation matrix. Red and maroon lines show some representative paths sampled from Eq. (2). The relative numbers of maroon (2) and red (10) paths respect the statistics over all sampled paths. Parameter values: $\beta = 3$, $\Lambda = 2$, $T = 82$. Sides: histograms of projections along PC1 (top) and PC2 (right). Inset: folding probabilities p_{nat} along each path. (b),(c) The fold of the LP family is stabilized by alternating configurations of charges. (d),(e) Sequence logos of PC1 and PC2.

with opposite chains of alternating charges along the electrostatic mode [Figs. 1(b) and 1(c)]. We will use our path sampling procedure to interpolate between the two subfamilies, see start (white star) and end (black star) sequences in Fig. 1(a).

To mimic the approach followed for natural proteins we train a RBM on the LP sequence data generated above, to infer an approximate expression for p_{nat} from the data; see Supplemental Material, Sec. 3 [23] for the inference of the RBM model. We then use our sampling algorithm to produce mutational paths, see Fig. 1(a). The algorithm is able to find excellent mutational paths in terms of the ground-truth folding probabilities p_{nat} of intermediate sequences, even higher than the ones of $\mathbf{v}_{\text{start}}, \mathbf{v}_{\text{end}}$ when imposing high β [inset of Fig. 1(a)]. Repeated runs of the sampling procedure give different paths that cluster into two classes, shown in red and maroon in Fig. 1(a). While few paths exploit a transient introduction of Cys-Cys interaction (on sites 6, 11, and 22) to stabilize the structure while flipping the electrostatic residues (maroon cluster); most introduce additional stabilizing electrostatic contacts

along the path (red cluster). See Supplemental Material, Sec. 4 [23] for details.

Mutational path sampling from data-driven models of natural proteins.—We next show that our path sampling procedure can be applied to natural proteins. We train a RBM from MSA data of the WW family, a protein domain binding specifically proline-rich peptides [11,34] and sample mutational paths between the Human YAP1 domain and three natural sequences known to have different binding specificities [35]. Figure 2(a) shows some sampled paths in the plane spanned by the inputs $I(\mathbf{v})$ [Eq. (1)] to two RBM hidden units chosen to cluster natural WW sequences depending on their binding specificities [22]. Intermediate sequences have high probabilities according to the RBM model, see Fig. 2(b). We then use AlphaFold [36] to assess the quality of the intermediate sequences; AlphaFold is able to predict the phenotypic effects of few mutations [37], and to compare the resulting structures to natural folds through template modeling scores (TM score) [24], ranging from 0—unrelated proteins—up to 1—perfect match. We obtain the TM score > 0.5, indicating a high similarity between the folds of sequences sampled along the path and of natural WW, see Supplemental Material, Sec. 5.3 [23] for details.

We next estimate binding affinities for each class using ProteinMPNN [13], an autoregressive structural-based probabilistic model that takes as input a backbone structure of a protein-ligand complex and predicts the affinity score of a putative protein sequence. Here, we use available complexes of natural WW domains of binding classes I, II/III, IV with their cognate peptides, see Fig. 2(c) and Supplemental Material, Sec. 5.4 [23]. As expected, along the I → II/III path the affinities to class I (respectively, II/III)–cognate peptides decrease (increase), see Fig. 2(d). Interestingly, Fig. 2(d) shows the existence of a region on the I → IV path in which the predicted affinities with respect to both complexes are high. It has been experimentally shown that some natural WW domains belonging to class I have also class IV activity [38]. This promiscuity may be favored by the fact that class I and IV cognate peptides bind two distinct loops of the WW domain [Fig. 2(c)]. In Supplemental Material, Sec. 5.2 [23] we corroborate these results by sampling more paths between class I and IV. To further assess the specificity of sequences on the sampled path, we train approximate class-specific RBM models from sequences in the quadrants of Fig. 2(a), see Supplemental Material, Sec. 3 [23]. The crossovers between the log-likelihoods of the class-specific RBMs in Figs. 2(e) suggests the presence of specificity switches along the I → II/III and I → IV paths. The scores provided by class-specific RBMs and ProteinMPNN are correlated along the paths, see Supplemental Material Fig. S6 [23].

Mean-field theory of mutational paths.—To better characterize the typical properties of mutational paths we resort to mean-field theory, by formally sending $N \to \infty$, while
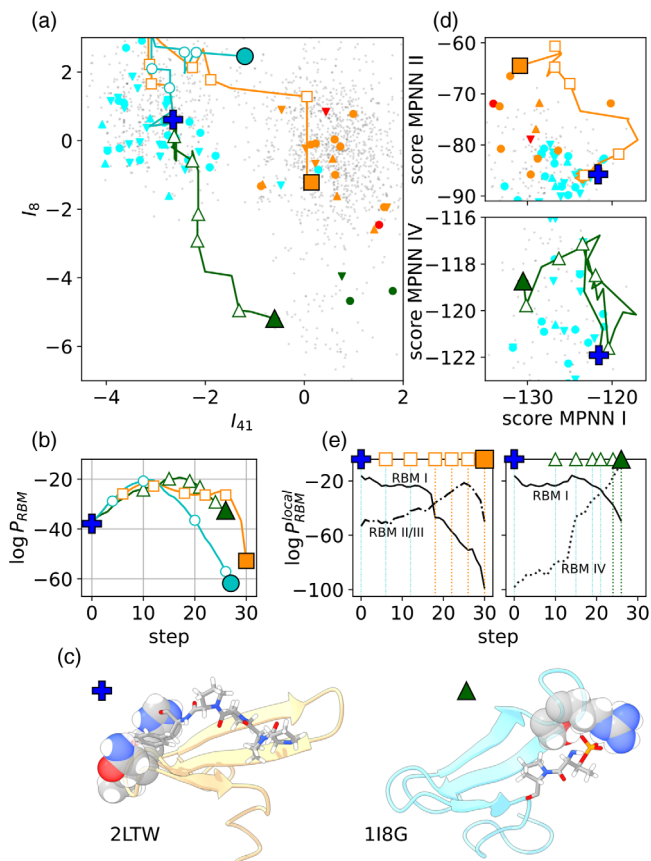

FIG. 2. Mutational paths of the WW domain using RBM trained on the PFAM PF00397 family, see Supplemental Material, Sec. 3 [23] for details about implementation. (a) Natural sequences $\mathbf{v}$ (gray dots) projected onto the plane of inputs I_μ (here $M = 50$) of two hidden units clustering sequences according to the types of ligands they bind [34]. Colored sequences: Experimentally tested: I (cyan), II (red), III (orange), IV (green). Upper (Lower) triangles: natural (artificial), from [38]. Circles: natural, from [39]. Blue cross represents the YAP1 domain. Lines shows the projection of three representative paths connecting YAP1 to a sequence in classes I (circle), II (square), and IV (triangle). Intermediate sequences (empty symbols) are listed in Supplemental Material, Sec. 5.1 [23]. Parameters: $\beta = 3$, $\Lambda = 0.1$.(b) Log P_{RBM} for sequences along the paths. (c) Complexes (WW domain and cognate peptides) for classes I (blue cross) and IV (green triangle) [40]. Atoms corresponding to the two binding pockets are highlighted. (d) ProteinMPNN scores for binding affinity, see Supplemental Material, Sec. 5.4 [23]. x axis measures the affinity to class I reference structure while y axes show affinity to classes II/III (Top) and IV (Bottom) reference structure respectively. (e) Log-likelihood along the paths from I to II (Left) and from I to IV (Right) according to class-specific RBMs trained on sequences in the three quadrants (Solid: I, Dot-dashed: II/III, Dotted: IV).

keeping the number T of steps finite. To allow for $\mathcal{O}(N)$ mutations between contiguous sequences we write the transition factor in Eq. (2) as $\pi(\mathbf{v}, \mathbf{v}') = e^{-N\Phi(q)}$, where the potential Φ is a decreasing function of the overlap (fraction of conserved amino acids between successive

sequences) $q = (1/N) \sum_i \delta_{v_i, v_i'}$. Φ controls the evolutionary dynamics along the path, and will be made precise below.

Mean-field theory exploits the bipartite nature of the RBM architecture and allows us to monitor two sets of order parameters characterizing the paths $\mathcal{V}$: the mean values of the hidden-unit inputs, $m_t^\mu = (1/N)\langle I_\mu(\mathbf{v}_t)\rangle$, and of the overlaps, $q_t = (1/N) \sum_i \langle \delta_{v_{i,t}, v_{i,t+1}}\rangle$; here, $\langle \cdot \rangle$ denotes the average over $\mathcal{P}(\mathcal{V})^\beta$.

The $T \times (M+1)$ order parameters m_t^μ and q_t are determined through minimization of the path free-energy density f_{path}, see Supplemental Material, Sec. 6 [23], with

$$f_{\text{path}}(\{m_t^\mu\}, \{q_t\}) = -\sum_{t,\mu} [\Gamma_\mu(m_t^\mu) - m_t^\mu \Gamma_\mu'(m_t^\mu)]$$
$$+ \sum_t [\Phi(q_t) - q_t \Phi'(q_t)]$$
$$- \frac{1}{\beta N} \sum_i \ln Z_i(\{m_t^\mu\}, \{q_t\}). \quad (3)$$

Here, $\Gamma_\mu(m) = (1/N) \ln \int dh\, e^{N\,mh - \mathcal{U}_\mu(h)}$ and Z_i is the following site-dependent partition function,

$$Z_i(\{m_t^\mu\}, \{q_t\})$$
$$= \sum_{\{v_t\}} \exp\left(\beta \sum_t g_i(v_t) + \beta \sum_{t,\mu} \Gamma_\mu'(m_t^\mu) w_{i\mu}(v_t)\right.$$
$$\left. - \beta \sum_t \Phi'(q_t) \delta_{v_t, v_{t+1}}\right). \quad (4)$$

Z_i can be efficiently estimated through products of transfer matrices, of sizes 21×21. While the expression of f_{path} is exact for sequence length $N \to \infty$, we show below it is accurate even in the cases of LP ($N = 27$) and WW ($N = 31$).

Choice of the elastic potential.—The potential Φ can enforce continuity (Cont) requirements, e.g., successive sequences along the path differ by, say, K mutations at most, or mimic the evolutionary (Evo) dynamics of natural sequences through stochastic mutations.

In the Cont scenario the potential Φ should forbid large jumps along the paths. We thus consider a hard-wall repulsive potential [Fig. 3(a)],

$$\Phi_{\text{Cont}}(q) = \frac{\phi(T)}{q - q_c(T)} \quad \text{if } q_c(T) < q \le 1, +\infty \text{ otherwise.} \quad (5)$$

The location of the hard wall, $q_c(T) = 1 - \gamma/T$, allows the path to explore at most $K \equiv T \times N(1 - q_c) = \gamma N$ mutations in T steps. Choosing $\gamma \ge D/N$ (D being the Hamming distance between $\mathbf{v}_{\text{start}}$ and $\mathbf{v}_{\text{end}}$) is therefore sufficient to interpolate between the two edge sequences, with larger
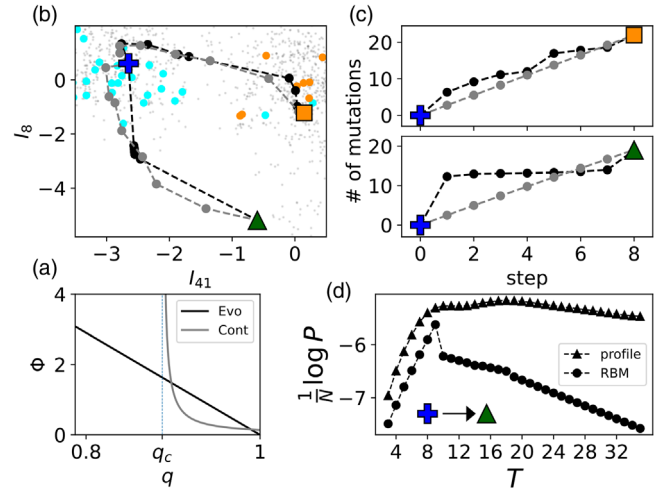

FIG. 3. Mean-field theory of mutational paths for the RBM model trained on WW domain. (a) Sketches of the potentials Φ_{Evo} (black) and Φ_{Cont} (gray) vs q. (b) Same two-dimensional representation as in Fig. 2(a) for the mean-field paths with Evo (black lines) and Cont (gray lines) potentials. (c) Cumulative numbers of mutations vs t. Here, $\mu = 10^{-5}$ and $\gamma = 0.9$, so that the cumulative numbers match for $t = T$. (d) Log-probability of joining class I and class IV natural WW domains in T steps with the profile (triangles) and RBM (circles) models. Jumps signal the onset of several new mutations, e.g., 4 in the mean-field *free* path at $T = 10$.

values of γ authorizing more flexible paths. The proportionality constant $\phi(T) = 1/T^2$ is set to guarantee the existence of a well-defined limit for large T.

In the Evo scenario, the potential should emulate Kimura's model of neutral evolution [25], while the P_{model} factors in Eq. (2) correspond to selection. Denoting the mutation rate (over a time interval corresponding to one step of the path) by μ, the potential is given by [41]

$$\Phi_{\text{Evo}}(q) = (1-q) \ln \left(1 + \frac{A}{e^{\mu A/(A-1)} - 1}\right), \quad (6)$$

where $A = 21$ is the number of amino acids plus the gap state; a derivation of Φ_{Evo} can be found in Supplemental Material, Sec. 8 [23]. This potential is linearly decreasing with q, see Fig. 3(a).

Cont and Evo mean-field paths between class-specific WW domains are shown in Fig. 3(b); both follow similar traces in the specificity plane, in agreement with the paths in Fig. 2(a). However, mutations are homogeneously spread along the Cont path, with $\simeq N\gamma/T$ mutations at each step [Fig. 3(c)]. Conversely, the Evo path is highly heterogeneous, with some steps accumulating many mutations and others barely any [Fig. 3(c)]; see Supplemental Material, Sec. 6.1 [23] for the list of consensus sequences computed with mean-field theory. Interestingly, most steps along the Evo path I → IV are concentrated in the region characterized by promiscuous sequences binding both

ligand classes as mentioned above. The linearity of Φ_{Evo} makes the transition probabilities π in $\mathcal{P}$ in Eq. (2) independent of the location of mutations, concentrating intermediate sequences in the region of highest fitness.

Mean-field based estimation of evolutionary distance.— As an application of our mean-field approach we show how it can be used to estimate evolutionary distances between sequences with complex data-driven models, including epistatic interactions between residues. The probability that sequence $\mathbf{v}_{\text{end}}$ is reached after T steps of stochastic mutations with rate μ starting from $\mathbf{v}_{\text{start}}$ is given by

$$P(\mathbf{v}_{\text{start}} \to \mathbf{v}_{\text{end}}|T) \sim \exp\left[-N(f_{\text{path}}^{\text{constrained}} - f_{\text{path}}^{\text{free}})\right], \quad (7)$$

where $f_{\text{path}}^{\text{constrained}}$ is the free energy in Eq. (4) (with potential Φ_{Evo}) minimized under boundary conditions matching both $\mathbf{v}_{\text{start}}$ and $\mathbf{v}_{\text{end}}$, while $f_{\text{path}}^{\text{free}}$ is obtained by releasing the boundary condition at the end extremity of the path. Details on the numerical optimization are given in Supplemental Material, Sec. 6.2 [23].

This probability can be computed as a function of T to determine the optimal time (evolutionary distance) T^* at which it is maximal. For purely neutral evolution, $f_{\text{path}}^{\text{free}} = 0$ and the probability $P(\mathbf{v}_{\text{start}} \to \mathbf{v}_{\text{end}}|T)$ can be exactly computed; T^* then coincides with the predictions of Kimura's theory of neutral evolution [25], see Supplemental Material, Sec. 8 [23]. T^* can also be easily computed for profile models [14], where selection acts independently from site to site, see Fig. 3(d) for an illustration of WW. Our mean-field theory allows us to go well beyond profile models, and to compute the probability P in the presence of epistatic effects in the RBM model inferred from WW sequence data. Figure 3(d) shows that the evolutionary distance T^* may then substantially differ from its profile counterpart, showing the effectiveness of our mean-field approach to deal with complex sequence models.

Conclusion.—Proteins with known (annotated) functional specificity form a tiny subset of available sequences. Learning accurate, generative class-specific models from these limited data is generally not possible [42]. Our path-based approach, inspired by evolutionary dynamics, circumvents this issue and offers an effective way to design proteins interpolating between different functional subclasses without annotated sequences (apart the anchors of the paths).

In addition, we have introduced a mean-field analysis of paths generated by RBM, characterizing the trajectories of the inputs to the hidden units and of the overlaps between successive sequences. Mean field is a powerful computational scheme in the presence of strong interactions between residues, e.g., to estimate evolutionary distances. This result opens the way to ancestral reconstruction and to the prediction of phylogenetic trees [14] with data-driven, epistatic models.

A potentially interesting biological finding in our study of the WW domain is that paths interpolating between classes I and IV go through a region apparently deprived of natural sequences, albeit corresponding to high RBM likelihood [43] and high AlphaFold/ProteinMPNN scores for both ligands (Fig. 2). While experimental investigations are needed to check our finding, these intermediate sequences are putatively unspecialized, and possibly similar to ancestral proteins [6].

*Corresponding author.
eugenio.mauri@phys.ens.fr
†Corresponding author.
remi.monasson@phys.ens.fr

[1] O. Kuchner and F. H. Arnold, Trends Biotechnol. **15**, 523 (1997).

[2] W. P. Russ, M. Figliuzzi, C. Stocker, P. Barrat-Charlaix, M. Socolich, P. Kast, D. Hilvert, R. Monasson, S. Cocco, M. Weigt, and R. Ranganathan, Science **369**, 440 (2020).

[3] A. Hawkins-Hooker, F. Depardieu, S. Baur, G. Couairon, A. Chen, and D. Bikard, PLoS Comput. Biol. **17**, 1 (2021).

[4] P. Tian and R. B. Best, PLoS Comput. Biol. **16**, 1 (2020).

[5] S. F. Greenbury, A. A. Louis, and S. E. Ahnert, Nat. Ecol. Evol. **6**, 1742 (2022).

[6] O. Khersonsky and D. S. Tawfik, Annu. Rev. Biochem. **79**, 471 (2010).

[7] E. Weinan and E. Vanden-Eijnden *et al.*, Annu. Rev. Phys. Chem. **61**, 391 (2010).

[8] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, Annu. Rev. Phys. Chem. **53**, 291 (2002).

[9] T. Mora, A. M. Walczak, and F. Zamponi, Phys. Rev. E **85**, 036710 (2012).

[10] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).

[11] M. Sudol, Prog. Biophys. Molec. Biol. **65**, 113 (1996).

[12] M. Socolich, S. Lockless, W. Russ, H. Lee, K. Gardner, and R. Ranganathan, Nature (London) **437**, 512 (2005).

[13] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. Wicky, A. Courbet, R. J. de Haas, N. Bethel *et al.*, Science **378**, eadd2187 (2022).

[14] J. Felsenstein, *Inferring Phylogenies* (Sinauer Associates Sunderland, MA, 2004), Vol. 2.

[15] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Proc. Natl. Acad. Sci. U.S.A. **108**, E1293 (2011).

[16] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt, Rep. Prog. Phys. **81**, 032601 (2018).

[17] K. Shekhar, C. F. Ruberman, A. L. Ferguson, J. P. Barton, M. Kardar, and A. K. Chakraborty, Phys. Rev. E **88**, 062705 (2013).

[18] H. Jacquin, A. Gilson, E. Shakhnovich, S. Cocco, and R. Monasson, PLoS Comput. Biol. **12,** 1 (2016).

[19] A. Di Gioacchino, J. Procyk, M. Molari, J. S. Schreck, Y. Zhou, Y. Liu, R. Monasson, S. Cocco, and P. Šulc, PLoS Comput. Biol. **18,** 1 (2022).

[20] M. Bisardi, J. Rodriguez-Rivas, F. Zamponi, and M. Weigt, Mol. Biol. Evol. **39,** msab321 (2022).

[21] A. Fischer and C. Igel, in *Iberoamerican Congress on Pattern Recognition* (Springer, New York, 2012), pp. 14–36.

[22] J. Tubiana, S. Cocco, and R. Monasson, eLife **8,** e39397 (2019).

[23] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.130.158402 for additional discussion, which includes Refs. [2,10,13,18,21,22,24–32].

[24] Y. Zhang and J. Skolnick, Proteins **57,** 702 (2004).

[25] M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, England, 1983).

[26] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, JAX: Composable transformations of PYTHON+NumPy programs (2018), available online at: http://github.com/google/jax.

[27] X. Espanel and M. Sudol, J. Biol. Chem. **274,** 17284 (1999).

[28] M. Levitt and M. Gerstein, Proc. Natl. Acad. Sci. U.S.A. **95,** 5913 (1998).

[29] S. Miyazawa and R. L. Jernigan, J. Mol. Biol. **256,** 623 (1996).

[30] T. Tieleman, in *ICML '08: Proceedings of the 25th International Conference on Machine learning* (Association for Computing Machinery, New York, NY, USA, 2008), pp. 1064–1071.

[31] J. Tubiana, Restricted Boltzmann machines: From compositional representations to protein sequence analysis, Ph.D. thesis, Université Paris sciences et lettres, Paris, France, 2018.

[32] J. Tubiana, Probabilistic graphical models (pgm) (2018), available online at: https://github.com/jertubiana/PGM.

[33] E. I. Shakhnovich and A. M. Gutin, Proc. Natl. Acad. Sci. U.S.A. **90,** 7195 (1993).

[34] A. Zarrinpar and W. A. Lim, Nat. Struct. Biol. **7,** 611 (2000).

[35] M. Sudol and T. Hunter, Cell **103,** 1001 (2000).

[36] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, Nature (London) **596,** 583 (2021).

[37] J. M. McBride, K. Polev, V. Reinharz, B. A. Grzybowski, and T. Tlusty, arXiv:2204.06860.

[38] W. P. Russ, D. M. Lowery, P. Mishra, M. B. Yaffe, and R. Ranganathan, Nature (London) **437,** 579 (2005).

[39] L. Otte, U. Wiedemann, B. Schlegel, J. R. Pires, M. Beyermann, P. Schmieder, G. Krause, R. Volkmer-Engert, J. Schneider-Mergener, and H. Oschkinat, Protein Sci. **12,** 491 (2003).

[40] E. F. Pettersen, T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. I. Croll, J. H. Morris, and T. E. Ferrin, Protein Sci. **30,** 70 (2021).

[41] I. Leuthäusser, J. Stat. Phys. **48,** 343 (1987).

[42] Our class-specific RBMs in Fig. 2 were used to assess membership to a class, a much easier task than design. Membership to families in PFAM are decided by Hidden Markov Models, which neglect correlations between residues.

[43] We stress that RBM trained on all sequence data, mixing several classes, cannot detect a change of specificity, contrary to RBM models restricted to each class (Fig. 2(e)).