Mutational paths with sequence-based models of proteins: from sampling to mean-field characterisation

Eugenio Mauri, Simona Cocco, Rémi Monasson

Mutational paths with sequence-based models of proteins: from sampling to mean-field characterisation

Eugenio Mauri, Simona Cocco, Rémi Monasson[1]

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

(Dated: October 21, 2022)

Identifying and characterizing mutational paths is an important issue in evolutionary biology and in bioengineering. We here introduce a generic description of mutational paths in terms of the goodness of sequences and of the mutational dynamics (how sequences change) along the path. We first 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 the properties of mutational paths for different mutational dynamics of interest, and show how it can be used to extend Kimura's estimate of evolutionary distances to sequence-based epistatic models of selection.

Introduction. Designing proteins with controlled properties, such as stability, binding affinity and specificity is a central goal in bioengineering. Directed evolution setups result in the discovery of new proteins with enhanced activities or affinities to a specific substrate [1]. Over the past years, much progress 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 shown to be robust generative models, 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 (Fig. 1), see however [4]. Yet solving this problem would be important in bio-engineering, *e.g.* to help design proteins with gradually changing functionalities. In addition, it would shed light on the navigability of the sequence landscape [5], and on how specificity emerged from ancestral, promiscuous proteins [6]. Informally speaking, a path is a succession of mutants interpolating between two fixed sequences at the edges, such that intermediate proteins maintain good functionality and contiguous sequences along the path differ by few mutations. The latter constraint depends on the objective, e.g. producing mutational paths plausible from an evolutionary point of view, or optimized for experimental validations. In particular, due to the huge number of possible paths mutagenesis experiments generally restrict to direct paths going through the 2^D mutants containing the amino acids appearing in the two edge sequences (differing on D sites), see Fig. 1 [7]. However, constraining paths to be direct may preclude the discovery of better global paths, involving mutations and their reversions and reaching more favorable regions in the sequence space [8] (Fig. 1).

While various methods exist for building transition paths between the minima of a multi-dimensional continuous landscape [9, 10] dealing with discrete configurations requires the development of specific procedures [11]. We hereafter propose a Monte Carlo algorithm

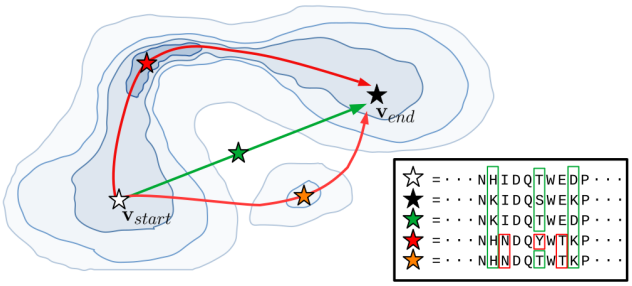

FIG. 1. **Mutational paths between two sequences in the landscape associated to a protein family.** Darker blue levels correspond to increasing values of the protein fitness. Paths joining the start and end sequences are either direct (green: each site carries the amino acid present at the same position in the initial or in the final sequence) or global (red: no restriction on amino acids), making possible the exploration of high fitness regions.

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 [12], 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 [13, 14]. The functionality of the sequences along the paths is validated with structure (ligand+protein)-informed software [15]. 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 [16].

Definition and sampling of mutational paths. We assume the sequence landscape is modeled through a probability distribution $P_{model}(\mathbf{v})$ 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.* share its common structural and functional properties, and can be learned from homologous

sequence data [17, 18]. For natural protein families, exact expressions for P_{model} are not available, but approximate distributions can be inferred from multi-sequence alignments (MSA) using unsupervised learning techniques.

Hereafter, we use Restricted Boltzmann Machines (RBM) [19], a class of generative models based on two-layer graphs [20]. RBM 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

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

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 $P_{model}(\mathbf{v}) = \int d\mathbf{h}\, P_{RBM}(\mathbf{v}, \mathbf{h})$ over the sequences $\mathbf{v}$ in a multi-sequence alignment of the family. While other unsupervised procedures providing approximate P_{model} can be used, such as Direct Coupling Analysis [17, 18], RBM offer a convenient way to interpret and to visualize the changes in sequences along mutational paths, as we will see below.

We define the probability of a mutational path of T steps, $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_{T-1}\}$ through

$$\mathcal{P}[\mathcal{V}|\mathbf{v}_{start}, \mathbf{v}_{end}] \propto \prod_{t=1}^{T-1} P_{model}(\mathbf{v}_t) \times$$
$$\pi(\mathbf{v}_{start}, \mathbf{v}_1) \times \prod_{t=1}^{T-2} \pi(\mathbf{v}_t, \mathbf{v}_{t+1}) \times \pi(\mathbf{v}_{T-1}, \mathbf{v}_{end}) \quad (2)$$

where the 'transition' factor $\pi(\mathbf{v}, \mathbf{v}')$ increases with the similarity between the sequences $\mathbf{v}, \mathbf{v}'$. As an illustration we may 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 sequences $\mathbf{v}_t$ and attempt at mutating one amino acid, under the constraint that the Hamming distances of the trial sequence $\mathbf{v}'$ with $\mathbf{v}_{t-1}$ and $\mathbf{v}_{t+1}$ be at most 1. The mutation is then rejected or accepted, *i.e.* $\mathbf{v}_t \leftarrow \mathbf{v}'$ according to detailed balance. Note that for global paths amino acids can take any values. For direct paths each amino acid has to coincide with the one either in $\mathbf{v}_{start}$ or in $\mathbf{v}_{end}$ on the same site, and the length T of the path matches the Hamming distance $D = N\big(1 - q(\mathbf{v}_{start}, \mathbf{v}_{end})\big)$ between the two edge sequences. 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 the value of β is initially very small and progressively ramped up to some target value. The complete procedure and the proof of detailed balance are given in Supplemental Material, Sec. 1.

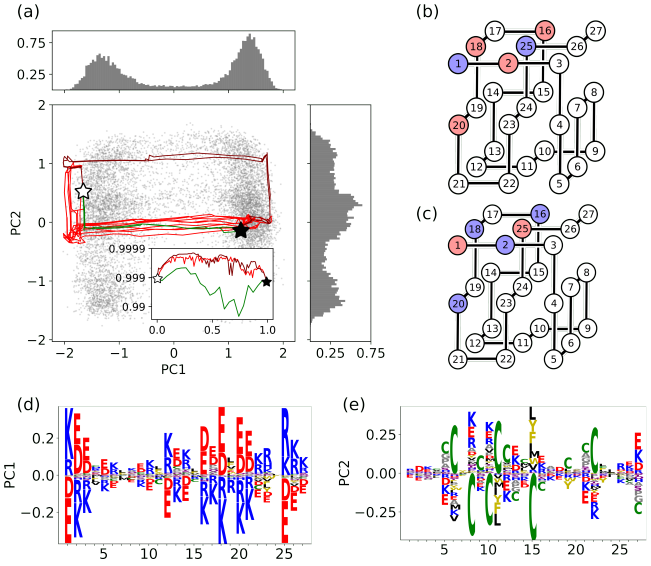

FIG. 2. **Mutational paths for lattice proteins**, joining sequences ☆=DRGIQCLAQMFEKEMRKKRRKCYLECD and ★=RECCAVCHQRFKDKIDEDYEDAWLKCN belonging to the family with structure shown in b) and c). Red and blue colors respectively correspond to negatively and positively charged amino acids. Cysteine is denoted by a green C. (a) Projections of 10^4 LP sequences in the family (grey dots) along the top two PC of their correlation matrix. Green lines represent direct paths, while red and maroon lines show some global paths sampled from Eq. (2); here, we set inverse temperature $\beta = 3$ and the mutation penalty $\Lambda = 2$, while the length of the direct and global paths are respectively $T_{direct} = 24$, $T_{global} = 82$. The relative numbers of maroon (2) and red (10) paths respect the statistics over all sampled paths. Sides: histograms of projections along PC1 (top) and PC2 (right). Inset: folding probability p_{nat} along each path vs. number of mutations/T.(b,c) Structure of the family. Sequences having opposite alternating configurations of charges along PC1 fold equally well. (d,e) Sequence logos of the top two PCs.

Benchmarking mutational path sampling on in silico proteins. We benchmark the performances of our MC procedure on a model of Lattice Proteins (LP) [12, 21]. 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. 2(a)) is defined by the probability $p_{nat}(\mathbf{v}|\mathbf{S})$ 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 for details.

We first generate many sequences $\mathbf{v}$ with high p_{nat} values for the fold $\mathbf{S}$ of Figs. 2(b,c) following the procedure of [22]. We next compute the top two Principal Components (PC) of these sequence data using one-hot encoding (Figs. 2(d,e)): PC1 corresponds to an extended electrostatic mode, and PC2 identifies possible Cys-Cys bridges. Projecting the sequences onto these two PCs re-

veals two sub-families separated along PC1 (Fig. 2(a)), associated to opposite chains of alternating charges along the electrostatic mode (Figs. 2(b,c)). We will use our path sampling procedure to interpolate between the two sub-families, see start (white star) and end (black star) sequences in Fig. 2(a).

To mimick 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 for details about the inference of the RBM model. We then use our sampling algorithm to produce global mutational paths, see Fig. 2(a). The algorithm is able to find excellent global mutational paths in terms of the ground truth folding probability p_{nat} (insert Fig. 2(a)). By fixing the target inverse temperature β to a value larger than one, we are able to obtain p_{nat} values along the path higher than those of the sequences at the extremities. Repeated runs of the sampling procedure give different paths that cluster into two classes, shown in red and maroon in Fig. 2(a). While few global paths exploit a transient introduction of Cys-Cys interaction to stabilize the structure while flipping the electrostatic residues (marron cluster); most (red cluster) introduce additional stabilizing electrostatic contacts along the path (red cluster). See Supplemental Material, Sec 4 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. To do so we train a RBM from MSA data of the WW family, a protein domain binding specifically proline-rich peptides [13, 23] and sample mutational paths, either global or direct, between the Human YAP1 domain and three natural sequences known to have different binding specificities [26]. Figure 3(a) shows some sampled paths in the 2-dimensional space spanned by the inputs $I(\mathbf{v})$ (see Eq. 1) to two RBM hidden units chosen to cluster natural WW sequences depending on their binding specificities [20]. Figures 3(b,c) show the probabilities of sequences along global and direct paths are comparable to the ones of natural proteins, with significantly higher values for global paths. We then use AlphaFold [27] to assess the quality of the intermediates sequences; AlphaFold is able to predict the phenotypic effects of mutations [28], and to compare the resulting structures to natural folds through Template Modelling scores (TM-score) [29], ranging from from 0 -unrelated proteins- up to 1 -perfect match. We obtain 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. 5.2 for details.

We next estimate binding affinity for each class using ProteinMPNN [15], 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 complexes of known natural WW domains of classes I, II/III, IV with their cognate peptides, see Fig. 3(h) and Supplemental Material, Sec. 5.3 for details. As expected, along the I $\rightarrow$ II/III path the affinities of sequences to class I (II/III) –cognate peptides decrease (increase), see Fig. 3(d). Interestingly, Fig. 3(e) shows the existence of a region on the I $\rightarrow$ IV path in which the predicted

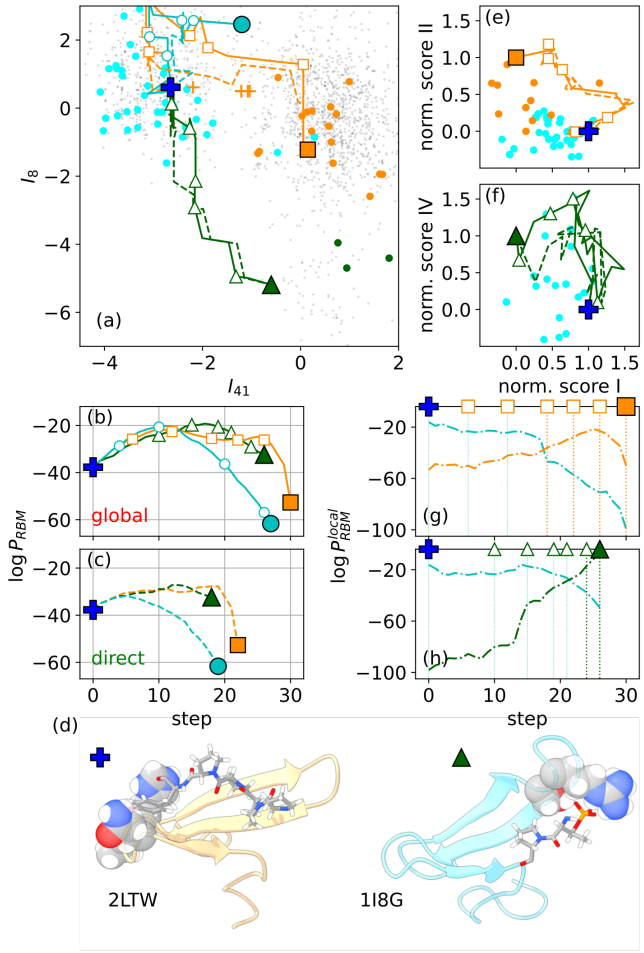

FIG. 3. **Mutational paths of the WW domain** using RBM trained on the PFAM PF00397 family, see Supplemental Material, Sec. 3 for details about implementation. (a) Natural sequences $\mathbf{v}$ (grey dots) projected in the plane of inputs I of two hidden units selected to cluster sequences according to the types of ligands they bind: I (cyan), II/III (orange), IV (green), see classification in [23]. Blue cross represents the YAP1 domain. Lines shows the projection of six representative paths (dashed: direct, solid: global) connecting YAP1 to sequences in classes I (circle), II/III (square; note the vicinity of the direct path with variants of YAP1 -orange crosses- tested in [24]) and IV (triangle). Empty symbols show intermediate sequences tested also with AlphaFold. All the sequences standing for these symbols are shown in Supplemental Material, Sec. 5.1. Parameters: $\beta = 3$, $\Lambda = 0.1$.(b)-(c) Log P_{RBM} for sequences along global and direct paths. (d) Complexes (WW domain and cognate peptides) for classes I (blue cross) and IV (green triangle) [25]. The atoms corresponding to the two binding pockets are highlighted in the structures. (e)-(f) Normalized ProteinMPNN scores for binding affinity, see Supplemental Material, Sec. 5.3. x-axis measures the affinity to class I reference structure while y-axes show affinity to classes II/III (e) and IV (f) reference structure respectively. (g)-(h) Log-likelihood along the global paths from I to II (g) and from I to IV (h) according to RBM trained on class-specific sequence data (Cyan: I, Orange: II/III, Green: IV).

affinities with respect to both complexes are high. This promiscuity may be favored by the fact that class I and IV cognate peptides bind two distinct loops of the WW domain (Fig. 3(h)). To further assess the specificity of sequences on the sampled path, we train class-specific RBM models from sequences in the associated quadrants in Fig. 3(a). The cross-overs between the log-likelihoods of the class-specific RBMs in Fig. 3(g,h) locate the specificity switches along the I $\to$ II/III and I $\to$ IV paths.

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 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 $q = \frac{1}{N}\sum_i \delta_{v_i,v'_i}$. Φ controls the elastic properties of 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, $m_t^\mu = \frac{1}{N}\langle I_\mu(\mathbf{v}_t)\rangle$, and of the overlaps (fraction of conserved amino acids between successive sequences), $q_t = \frac{1}{N}\sum_i \langle \delta_{v_{i,t},v_{i,t+1}}\rangle$; 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, with

$$f_{\rm path}(\{m_t^\mu\},\{q_t\}) = -\sum_{t,\mu}\left(\Gamma_\mu(m_t^\mu) - m_t^\mu\,\Gamma'_\mu(m_t^\mu)\right) \quad (3)$$
$$+\sum_t\left(\Phi(q_t) - q_t\,\Phi'(q_t)\right) - \frac{1}{\beta N}\sum_i \ln Z_i\big(\{m_t^\mu\},\{q_t\}\big)\ .$$

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

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

Z_i can be efficiently estimated through products of transfer matrices, of sizes 21×21 for global paths, and 2×2 for direct paths. The expression of f_{path} is exact for sequence length $N\to\infty$ and the numbers of hidden units, M, and of steps, T remain finite, and is an accurate approximation 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. 4(a)),

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

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 $\mathbf{v}_{start}$ and $\mathbf{v}_{end}$), is therefore sufficient to interpolate between the two edge sequences, with larger 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 [30], 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 we show in Supplemental Material, Sec. 7, that the potential is given by [31]

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

where $A=21$ is the number of amino acids plus the gap state. It is linearly decreasing with q, see Fig. 4(a).

Cont and Evo mean-field paths between class-specific WW domains are shown in Fig. 4(b); both follow similar traces in the specificity plane, in agreement with the sampled paths in Fig. 3(a). However, the distribution of mutations along the paths largely differ between the two scenarios, compare Fig. 4(c-d). Mutations are homogeneously spread along the Cont path, with $\simeq N\gamma/T$ mutations at each step. Conversely, the Evo path is highly

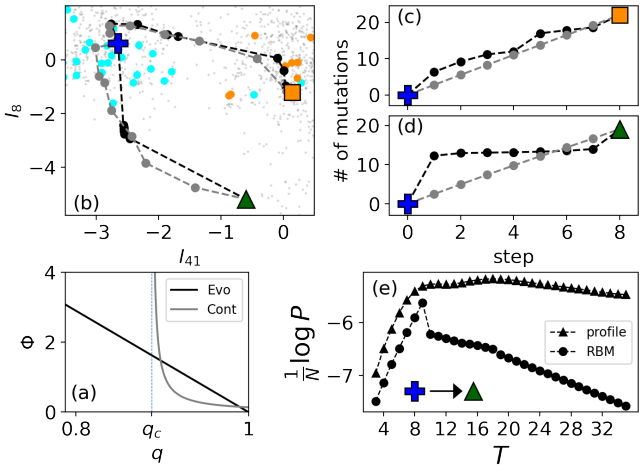

FIG. 4. **Mean-field theory of mutational paths** for the RBM model trained on WW domain. (a) Sketches of the potentials $\Phi_{\rm Evo}$ (black) and $\Phi_{\rm Cont}$ (gray) vs. q. (b) Same two-dimensional representation as in Fig. 3(a) for the mean-field paths with Evo (black lines) and Cont (grey lines) potentials. (c)-(d) Cumulative numbers of mutations vs. t. Here, $\mu = 10^{-5}$ and $\gamma = 0.9$, so that the cumulative numbers match for $t = T$. (e) 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 $T = 10$.

heterogeneous, with some steps accumulating many mutations and others barely any, see Supplemental Material, Sec. 6.1 for the list of consensus sequences computed with the 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, which could then correspond to ancestral, not yet specialized sequences [6].

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 $\mathbf{v}_{end}$ be reached after T steps of stochastic mutations with rate μ starting from $\mathbf{v}_{start}$ is given by

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

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

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 neural evolution, $f_{path}^{free} = 0$ and the probability $P(\mathbf{v}_{start} \to \mathbf{v}_{end}|T)$ can be exactly computed; T^* then coincides with the predictions of Kimura's theory of neutral evolution [30], see Supplemental Material, Sec. 7. Kimura's result can be easily extended to the case of profile models [16], where selection acts independently from site to site, see Fig. 4(e) for an illustration of WW. Our mean-field theory theory allows us to go well beyond profile models, and to compute the probability P in the presence of epistatis effects in the RBM model inferred from WW sequence data. Figure 4(e) 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. In this work we have introduced numerical and analytical tools to sample and characterize mutational paths in data-driven models (RBM) of protein sequences. Our sampling algorithm was illustrated on the WW domain, but can be applied to longer enzymes, with > 100 amino acids. We have validated the structural and functional properties of intermediate sequences along the paths with AlphaFold and ProteinMPNN, two deep-learning-based computational methods. A potentially interesting biological finding is that the path interpolating between specificity classes I and IV go through a region apparently deprived of natural sequences, albeit corresponding to high RBM likelihood[32] and high ProteinMPNN scores for both ligands (Fig. 3). These intermediate sequences are putatively unspecialized, and possibly similar to ancestral proteins. Investigating experimentally their properties would be very interesting.

In addition, we have shown how RBM models are amenable to mean-field analysis, through the determination of the time-trajectory of the mean inputs to the hidden units and of the overlaps between successive sequences along the path. 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 [16] with data-driven, epistatic models.

Acknowledgements. We are particularly grateful to J. Tubiana for his help on the use of ProteinMPNN. We also thank M. Bisardi, A. Di Gioacchino, A. Murugan, and F. Zamponi for discussions. This work was supported by the ANR-17 RBMPro and ANR-19 Decrypted CE30-0021-01 projects. E.M. is funded by a ICFP Labex fellowship of the Physics Department at ENS.

[1] O. Kuchner and F. H. Arnold, Trends in Biotechnology **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 Computational Biology **17**, 1 (2021).
[4] P. Tian and R. B. Best, PLOS Computational Biology **16**, 1 (2020).
[5] S. F. Greenbury, A. A. Louis, and S. E. Ahnert, bioRxiv (2021), 10.1101/2021.10.11.463990.
[6] O. Khersonsky and D. S. Tawfik, Annual Review of Biochemistry **79**, 471 (2010).
[7] F. J. Poelwijk, M. Socolich, and R. Ranganathan, Nat. Commun. **10**, 1 (2019).
[8] N. C. Wu, L. Dai, C. A. Olson, J. O. Lloyd-Smith, and R. Sun, eLife **5**, e16965 (2016).
[9] E. Vanden-Eijnden *et al.*, Annual review of physical chemistry **61**, 391 (2010).
[10] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, Annual review of physical chemistry **53**, 291 (2002).
[11] T. Mora, A. M. Walczak, and F. Zamponi, Physical Review E **85**, 036710 (2012).
[12] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[13] M. Sudol, Progress in biophysics and molecular biology **65**, 113 (1996).
[14] M. Socolich, S. Lockless, W. Russ, H. Lee, K. Gardner, and R. Ranganathan, Nature **437**, 512 (2005).
[15] 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 , eadd2187 (2022).
[16] J. Felsenstein and J. Felenstein, *Inferring phylogenies*, Vol. 2 (Sinauer associates Sunderland, MA, 2004).
[17] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Proceedings of the National Academy of Sciences **108**, E1293 (2011).
[18] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt, Reports on Progress in Physics **81**, 032601 (2018).

[19] A. Fischer and C. Igel, in *Iberoamerican congress on pattern recognition* (Springer, 2012) pp. 14–36.
[20] J. Tubiana, S. Cocco, and R. Monasson, eLife **8**, e39397 (2019).
[21] E. I. Shakhnovich and A. M. Gutin, Proceedings of the National Academy of Sciences **90**, 7195 (1993), https://www.pnas.org/doi/pdf/10.1073/pnas.90.15.7195.
[22] H. Jacquin, A. Gilson, E. Shakhnovich, S. Cocco, and R. Monasson, PLOS Computational Biology **12**, 1 (2016).
[23] A. Zarrinpar and W. A. Lim, Nature structural biology **7**, 611 (2000).
[24] X. Espanel and M. Sudol, Journal of Biological Chemistry **274**, 17284 (1999).
[25] 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 Science **30**, 70 (2021).
[26] M. Sudol and T. Hunter, Cell **103**, 1001 (2000).
[27] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, Nature **596**, 583 (2021).
[28] J. M. McBride, K. Polev, V. Reinharz, B. A. Grzybowski, and T. Tlusty, arXiv preprint arXiv:2204.06860 (2022).
[29] Y. Zhang and J. Skolnick, Proteins: Structure, Function, and Bioinformatics **57**, 702 (2004).
[30] M. Kimura, *The neutral theory of molecular evolution* (Cambridge University Press, 1983).
[31] I. Leuthäusser, Journal of statistical physics **48**, 343 (1987).
[32] 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. 3(g,h)).