Mutational paths in protein-sequence landscapes: from sampling to low-dimensional characterization

Eugenio Mauri, Simona Cocco, Rémi Monasson

▶ **To cite this version:**

Eugenio Mauri, Simona Cocco, Rémi Monasson. Mutational paths in protein-sequence landscapes: from sampling to low-dimensional characterization. 2022. hal-03645394v1

HAL Id: hal-03645394

https://hal.science/hal-03645394v1

Preprint submitted on 21 Apr 2022 (v1), last revised 6 Feb 2023 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Mutational paths in protein-sequence landscapes: from sampling to low-dimensional characterization

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: April 21, 2022)

Understanding how protein functionalities vary along mutational paths is an important issue in evolutionary biology and in bioengineering. We here propose an algorithm to sample mutational paths in the sequence space, realizing a trade-off between protein optimality and path stiffness. The algorithm is benchmarked on exactly solvable models of proteins in silico, and applied to data-driven models of natural proteins learned from sequence data. Using mean-field theory, we monitor the projections of the sequence on relevant modes along the path, allowing for an interpretation of the protein sequence trajectory. Qualitative changes observed in paths as their lengths are varied can be explained by the existence of a phase transition in infinitely-long strings of strongly coupled Hopfield models.

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 from an evolutionary point of view, and 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 mutations interpolating between two fixed sequences at the edges, such that the intermediate proteins maintain good functionality. 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 much better *global* paths, involving mutations and their reversions and reaching more favorable regions in the sequence space (Fig. 1).

While various methods exist for building transition paths between the minima of a multi-dimensional continuous landscape [8, 9] they cannot be easily adapted to the case of discrete configurations. We hereafter propose a Monte Carlo algorithm to sample mutational paths in sequence space. 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 sub-

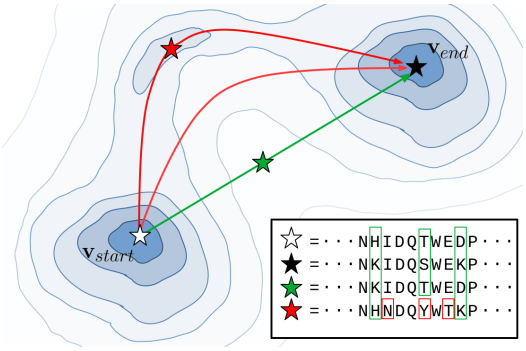

FIG. 1. **Mutational paths between two subfamilies in the sequence landscape associated to a protein family.** Darker blue levels correspond to increasing values of the protein fitness. Paths are either direct (green: each site carries the amino acid present at the same position in the initial or in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

families. We then apply our algorithm to the WW domains, a small binding module involved in the regulation of protein complexes [11] and studied in early works on sequence-based design [12]. The paths obtained between proteins with different specificities have high likelihoods and folding scores according to AlphaFold [13]. Furthermore, we show how mean-field theory can be applied to track informative projections of the multi-dimensional trajectories of sequences along the paths. Last of all, we observe that global paths, if long enough, significantly outperform direct paths. This crossover is related to the existence of a thermodynamic phase transition in infinitely-long strings of coupled Hopfield-like models that we analytically solve.

Algorithm for mutational path sampling. We assume the sequence landscape is modeled through a probability distribution $P_{model}(\mathbf{v})$ over amino-acid sequences $\mathbf{v}$ of length L. 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 [14, 15]. 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) [16], a class of generative models based on two-layer graphs [17]. 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 [14, 15], 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 sequences, $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_T\}$ through

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

where $\pi(\mathbf{v}, \mathbf{v}') = 1$ if the sequences $\mathbf{v}$ and $\mathbf{v}'$ are identical, $e^{-\Lambda}$ if they differ by one mutation (with $\Lambda > 0$), and 0 if they are two or more mutations apart. 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 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.

Benchmarking mutational path sampling on in silico proteins. We benchmark the performances of our MC procedure on a model of Lattice Proteins (LP) [10, 18]. In LP a sequences of 27 amino acids may fold into $\simeq 10^5$ different self-avoiding conformations going through the nodes of a $3\times3\times3$ cubic lattice. The sequence landscape associated to a conformation $\mathbf{S}$ (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.

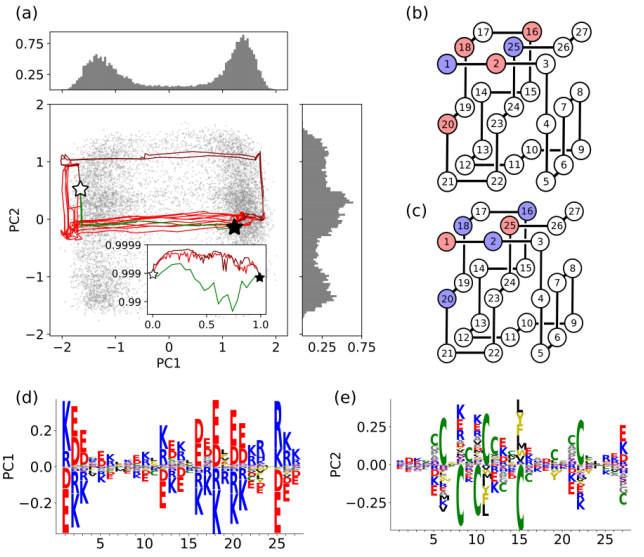

FIG. 2. **Mutational paths for lattice proteins**, joining sequences ☆=DRGIQCLAQMFEKEMRKKRRKCYLECD and ★=RECCAVCHQRFKDKIDEDYEDAWLKCN. Red and blue colors respectively correspond to negatively and positively charged amino acids. Cysteine is denoted by a green C. (a) Projections of 10^4 LP sequences (grey 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, target $\beta = 3$, $\Lambda = 2$, $T_{direct} = 24$, $T_{global} = 82$. 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) Native folds of the sequences in the family, corresponding to opposite alternating configurations of charges along PC1. (d,e) Logos of the top two PCs.

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 [19]. We next compute the top two Principal Components (PC) of these sequence data (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 reveals 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 procedure 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,

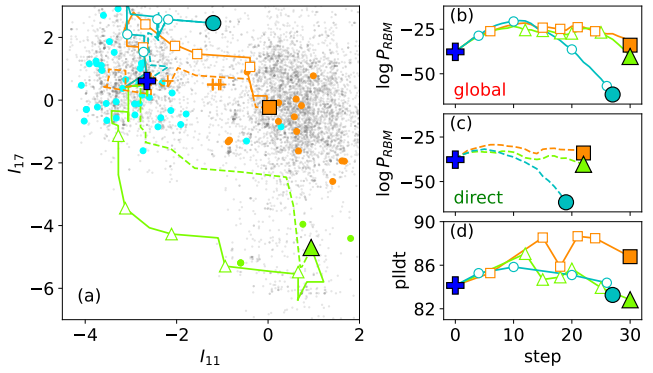

FIG. 3. **Mutational paths of the WW domain** using RBM trained on the PFAM PF00397 family, see Supplemental Material, Sec. 3 for details about implementation. (a) Natural sequences $\mathbf{v}$ (grey dots) projected in the plane of inputs I of two hidden units selected to cluster sequences according to the types of ligands they bind: I (cyan), II/III (orange), IV (green), see classification in [20]. 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 [21]) and IV (triangle). Empty symbols show intermediate sequences tested with AlphaFold in (d). Parameters: $\beta = 3$, $\Lambda = 0.1$. (b)-(c) Log P_{RBM} for sequences along global and direct paths. (d) AlphaFold confidence scores of predicted folds for intermediate sequences vs. nb. of mutations along the path.

see Fig. 2(a). The algorithm is able to find excellent global mutational paths in terms of the ground truth folding probability p_{nat}. 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 extremity sequences. 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 boost p_{nat} by transiently introducing Cys-Cys interactions (maroon cluster), most (red cluster) stabilize the structure by realizing more contacts between positively and negatively-charged amino acids than direct paths (Supplemental Material, Sec 4).

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 [11, 20] and sample mutational paths, either global or direct, between the Human YAP1 domain and three natural sequences known to have different binding specificities [22]. The quality of the sequences along the path is assessed from their probabilities P_{RBM} within the RBM model, and from 3D structure predictions obtained using AlphaFold [13]. Figure 3(a) shows some sampled paths in the 2-dimensional space spanned by the inputs $I(\mathbf{v})$ to two RBM hidden units chosen to cluster natural WW sequences depending on their binding specificities [17]. 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 report AlphaFold's confidence scores of intermediate sequences along global paths in Fig. 3(d), indicating that these sequences have well defined folds. Furthermore, we compare these predicted folds to that of natural WW using Template Modelling scores (TM-score) [23], which measure structure similarity from 0 -unrelated proteins- up to 1 -perfect match. We obtain TM-score > 0.5, indicating a high similarity between the folds of sequences along the path and of natural WW.

Mean-field characterization of mutational paths. To understand how mutational paths explore the sequence space we introduce a mean-field theory exploiting the bipartite nature of the RBM architecture. Mean field 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 $P(\mathcal{V})^\beta$.

In the mean-field framework a step along the path can involve multiple mutations. The transition factor π in Eq. (2) is defined through

$$\pi(\mathbf{v},\mathbf{v}') = e^{-N\,\Phi(q)}\ ,\ \text{where}\ q = \frac{1}{N}\sum_i \delta_{v_i,v'_i} \qquad (3)$$

the potential Φ forbids small overlaps $q \ll 1$, *i.e.* discontinuous jumps along the paths. We impose $q > q_c = 1 - \gamma/T$, allowing for the path to explore at most $T \times N(1-q_c) = \gamma N$ mutations in T steps. Choosing $\gamma \geq D/N$ is therefore sufficient to interpolate between the two edge sequences, with larger values of γ authorizing more flexible paths. In practice we set $\Phi(q) = 1/(T^2|q-q_c|)$; Other choices of potentials with hard wall constraints give similar results.

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

$$f_{\rm path}(\{m_t^\mu\},\{q_t\}) = -\sum_{t,\mu}\left(\Gamma_\mu(m_t^\mu) - m_t^\mu\,\Gamma'_\mu(m_t^\mu)\right) \qquad (4)$$
$$+\sum_t\left(\Phi(q_t) - q_t\,\Phi'(q_t)\right) - \frac{1}{\beta N}\sum_i \ln Z_i\big(\{m_t^\mu\},\{q_t\}\big)\ .$$

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,

$$Z_i(\{m_t^\mu\},\{q_t\}) = \sum_{\{v_t\}}\exp\left(\beta\sum_i g_i(v_t) + \right.$$
$$\left. +\beta\sum_{t,\mu}\Gamma'_\mu(m_t^\mu)\,w_{i\mu}(v_t) - \beta\sum_t \Phi'(q_t)\,\delta_{v_t,v_{t+1}}\right)\ . \qquad (5)$$

Z_i can be efficiently estimated through products of $A\times A$-dimensional transfer matrices, where A is the number of Potts states. For global paths, $A = 21$ (20 amino acids plus the gap symbol), while $A = 2$ for direct paths. The

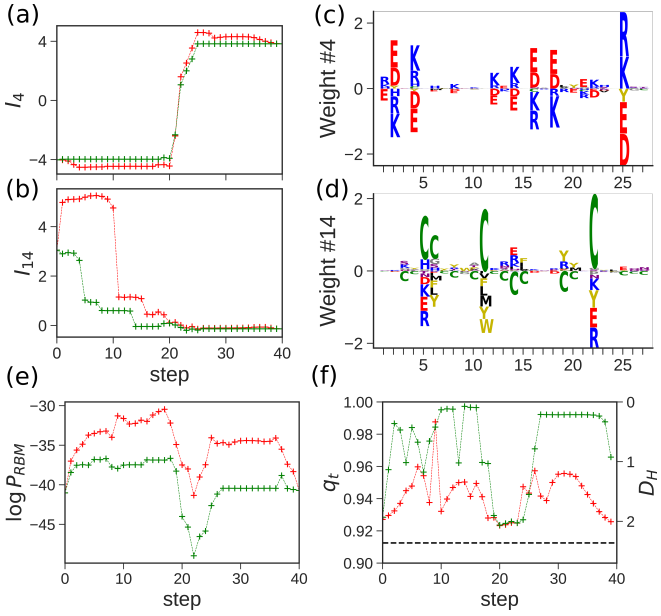

FIG. 4. **Mean-field description of mutational paths** in lattice proteins. (a)-(b) Values of two inputs similar to PC in Fig. 2 vs. number t of mutations along paths of length $T = 40$. Red and green lines correspond to, respectively, global and direct paths. Parameters: $\beta = 3$, $\gamma = 3.5$ chosen so that the average distance to the direct space of the mean-field solutions is the same as in Fig. 2. (c)-(d). Logos of the attached weights $w_{i,\mu}(v)$. (e) Log-likelihoods of sequences along the same paths as in panels (a),(b). (f) Overlap q_t (left scale) and average number of mutations $D_H = N(1-q_t)$ (right scale) between sequences at steps t and $t+1$ vs. t; The dark line shows q_c.

derivation of f_{path} is exact when the 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 ($N = 31$), as shown below.

The trajectories of the inputs m_t^μ and of the overlaps q_t reveal which and when latent factors of RBM enter into play throughout the interpolation between the initial and final sequences. Figures 4(a,b) show the trajectories of inputs associated to the weights in Figs. 4(c,d) for the lattice protein studied in Fig. 2. The dynamics explains how optimal paths exploit Cysteine-Cysteine interactions (not present in the initial and final sequences) in order to maintain the structure of the protein when the signs of the charge along the electrostatic chain are reversed (Fig. 4(a,b)), and agrees with the average behaviour of the paths sampled in Fig. 2(a), see Supplemental Material, Sec. 4.

Sequences along global paths have substantially higher probabilities than along direct paths (Fig. 4(e)). The exploration of favourable regions in the landscape is made possible by the slightly higher number of mutations between successive sequences in the former case than in the latter, see Fig. 4(f). Along global paths, most of the intermediate mutational steps do not abruptly affect the inputs nor the probability, with the exception of the bump in the overlap q at step $\sim$10, possibly related to the presence of preparatory mutations for the Cys-related transition in Fig. 4(b,d).

Crossover between global and direct paths. The results reported in Figs. 2, 3, 4 indicate that sequences along global mutational paths have larger scores than along direct paths, see Fig. 3(b,c). To better understand the differences between global and direct paths we introduce and analyze in details a toy-model capturing the effects of extra dimensions with respect to the direct subspace of sequences. This Hopfield-Potts (HP) model includes $M = 2$ patterns, and $A \geq 3$ Potts symbols. Denoting the first three symbols by a, b, c, the patterns are set to $w_{1,i}(v) = \delta_{v,a} + \omega\,\delta_{v,c}$ and $w_{2,i}(v) = \delta_{v,b} + \omega\,\delta_{v,c}$, uniformly over sites i, and define the sequence distribution $P_{HP}(\mathbf{v}) \propto \exp\left[\frac{1}{2}\sum_{i,j}\sum_{\mu=1,2} w_{\mu,i}(v_i)w_{\mu,j}(v_j)\right]$. The initial and final sequences are chosen, respectively, as $v_i^{start} = a$ and $v_i^{end} = b$; hence, ω quantifies the attractiveness of the global direction c (orthogonal to the direct subspace spanned by a, b). We then couple T such HP models to form a 1D-string with controlled stiffness (through the transition factors π in Eq. (3)), and anchored in $\mathbf{v}^{start}$ and $\mathbf{v}^{end}$.

As HP models are a special case of RBM with quadratic potentials $\mathcal{U}(h) \propto h^2$ [24] the path free-energy for trajectories over the inputs and the overlaps in Eq. (5) is exact when $N \to \infty$. The optimal trajectories can be analytically studied in great details, see Supplemental Material, Sec. 7, with the following results. For $\omega < \omega_c = \frac{1}{2}$, mutational paths typically *lie within the direct subspace* (Fig. 1): the attraction along the c direction is too weak to counterbalance the stiffness of the path imposed by the potential Φ. For $\omega > \omega_c$ optimal paths leave the direct subspace and *explore the global space* if their lengths exceed

$$T_{c.o.} = \max\left(\frac{1}{\gamma^2\Theta\left(\omega - \frac{1}{2}\right)}, \frac{1}{(\gamma-1)^2\left(\omega - \frac{1}{3}\right)}\right), \quad (6)$$

where $1/\gamma < \Theta \leq 1$ is the fraction of the T steps in which the sequences $\mathbf{v}^t$ along direct paths are distinct from $\mathbf{v}^{start}$ and $\mathbf{v}^{end}$, see Supplemental Material, Sec. 7.3.

The resulting phase diagram is shown in Fig. 5(a). As expected, for lower values of γ, paths become stiffer, and $T_{c.o.}$ increases. The average distance to the direct subspace, $d_{DS} = \frac{1}{N}\sum_i \langle(1-\delta_{v_i,v_i^{start}})(1-\delta_{v_i,v_i^{end}})\rangle$, and the average log-probability of intermediate sequences, $\langle\frac{1}{N}\log P_{HP}\rangle$ are shown in Fig. 5(b); their behaviours confirm the existence of the crossover at $T_{c.o.}$. This crossover is also observed for natural proteins, such as WW, when keeping the length T fixed and varying the flexibility γ of the path, see Fig. 5(c). As in the HP model case the non-orthogonality of the weight vectors learned by the RBM provides multiple opportunities for mutational paths to escape the direct subspace.

Conclusion. In this work we have shown how data-driven, in particular, RBM models of protein sequence data can be used to sample mutational paths. Though our sampling procedure was illustrated on short proteins, it can be easily applied to longer enzymes, with $>$ 100

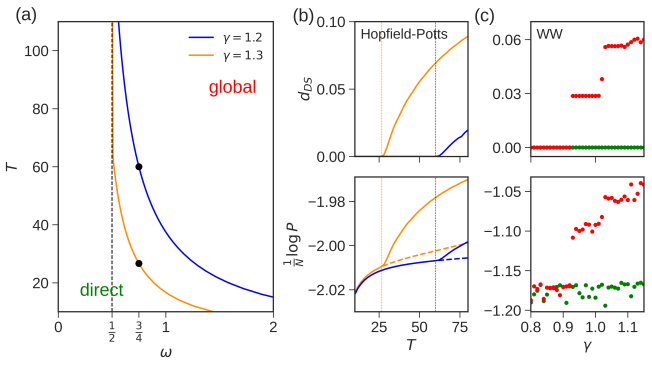

FIG. 5. **Crossover between direct and global mutational paths.** (a). Behavior of $T_{c.o.}$ vs. ω for the HP model and two values of γ, see Eq. (6). The black dots show the crossovers for $\omega = \frac{3}{4}$. (b) d_{DS} (Top) and $(\log P_{HP})/N$ averaged over intermediate sequences (Bottom; solid line: global, dashed: direct) vs. path length T; same parameters as in (a). (c) Mean-field estimates of d_{DS} (Top) and of $(\log P_{RBM})/N$ (Bottom; red: global paths, green: direct) vs. γ for mutational paths of the WW domain of length $T = 10$. Initial sequence: YAP 1, final sequence: green triangle in Fig. 3. In all panels $\beta = 3$.

amino acids, whose functionalities could be experimentally tested.

The analytical study of a toy Hopfield-Potts model reveals the existence of a qualitative change in mutational paths with their length. Long paths can explore favorable detours in the global sequence landscape, which are not accessible to shorter paths. An illustration, in the WW case, is the high-quality global green path going through a region with few natural sequences in Fig. 3(a). It would be very interesting to test this striking prediction experimentally.

In addition, the use of mean-field theory allows us to follow the dynamics of relevant latent factors along the paths, and understand how the transition from one functionality to another is implemented through sequential changes of few residues at a time. Extending our mean-field analysis to the case of an extensive number of RBM weight vectors (finite M/N) would allow for better monitoring the dynamics of the few inputs of interest along paths interpolating between subfamilies.

Acknowledgements. We are grateful to M. Bisardi, A. Di Gioacchino, A. Murugan, R. Ranganathan, J. Tubiana 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] E. Vanden-Eijnden *et al.*, Annual review of physical chemistry **61**, 391 (2010).
[9] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, Annual review of physical chemistry **53**, 291 (2002).
[10] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[11] M. Sudol, Progress in biophysics and molecular biology **65**, 113 (1996).
[12] M. Socolich, S. Lockless, W. Russ, H. Lee, K. Gardner, and R. Ranganathan, Nature **437**, 512 (2005).
[13] 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).
[14] 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).
[15] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt, Reports on Progress in Physics **81**, 032601 (2018).
[16] A. Fischer and C. Igel, in *Iberoamerican congress on pattern recognition* (Springer, 2012) pp. 14–36.
[17] J. Tubiana, S. Cocco, and R. Monasson, eLife **8**, e39397 (2019).
[18] 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.
[19] H. Jacquin, A. Gilson, E. Shakhnovich, S. Cocco, and R. Monasson, PLOS Computational Biology **12**, 1 (2016).
[20] A. Zarrinpar and W. A. Lim, Nature structural biology **7**, 611 (2000).
[21] X. Espanel and M. Sudol, Journal of Biological Chemistry **274**, 17284 (1999).
[22] M. Sudol and T. Hunter, Cell **103**, 1001 (2000).
[23] Y. Zhang and J. Skolnick, Proteins: Structure, Function, and Bioinformatics **57**, 702 (2004).
[24] A. Barra, A. Bernacchia, E. Santucci, and P. Contucci, Neural Networks **34**, 1 (2012).