Out-of-equilibrium noise facilitates inference from protein sequence data

Nicola Dietler, Cyril Malbranke, and Anne-Florence Bitbol*
Institute of Bioengineering, School of Life Sciences,
École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland and
SIB Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland
(Dated: May 14, 2025)

Homologous proteins have similar three-dimensional structures and biological functions that shape their sequences. The constrained sequence variability among homologs is exploited by statistical inference methods, such as Potts models, and by deep learning methods, such as AlphaFold, to infer protein structure and function. In a minimal model, we show that both fluctuating selection strength and the emergence of function facilitate coevolution-based inference of structural contacts. Our conclusions extend to realistic synthetic data. Thus, out-of-equilibrium noise arising from ubiquitous variations in natural selection promotes the success of inference from protein sequences.

Introduction. During evolution, proteins evolve through random mutations of their sequences and natural selection for their function. This coupling between sequence and function leads to coevolution between amino acid sites in multiple sequence alignments (MSAs) of homologous proteins, which share a common ancestry and a similar three-dimensional structure and function. Amino-acid usage is correlated between columns that are in contact in the three-dimensional structure of a protein, due to the need to preserve physico-chemical complementarity. This has allowed to predict structural contacts from sequences, using mutual information [1] and then Potts models [2–5], which are pairwise maximum entropy models inferred on MSAs. Coevolution is efficiently captured in the attention coefficients of protein language models trained on MSAs [6, 7], and constitutes an important ingredient of the major success of AlphaFold for protein structure prediction [7]. Beyond structure prediction, coevolution can be exploited to infer interaction partners [8–11] and functional sectors of collectively correlated amino acids [12–15], as well as to investigate protein evolution [16–18].

Natural environments often change over time [19–24], which yields variations of evolutionary pressures on proteins [19]. How do variable selection pressures impact inference from protein sequences? To address this question, we consider a minimal model of protein sequences under selection to preserve structural contacts, and we evolve these synthetic sequences under fluctuating selection strength. We find that the resulting out-of-equilibrium noise can facilitate the inference of structural contacts via Potts models. We demonstrate that this conclusion extends to realistic synthetic data. Next, we address implications for the evolution of function. In our minimal model, we consider an ancestral non-functional protein that evolves along a phylogenetic tree under natural selection to fold into a given structure. Again, out-of-equilibrium noise strongly improves contact prediction.

* anne-florence.bitbol@epfl.ch

Model and methods. We consider a minimal model, where protein sequences are represented by sequences of Ising spins $(\sigma_1, \ldots, \sigma_L)$, where L is the length of the sequence, and where each spin sits on one node of a fixed Erdős-Rényi random graph. Structural contacts are modelled by pairwise ferromagnetic couplings on the edges of this graph, all taken equal to 1 for simplicity, leading to the Hamiltonian

$$H = \sum_{(i,j)\in\mathcal{E}} \sigma_i \sigma_j \,, \tag{1}$$

where $\mathcal{E}$ denotes the set of edges of the graph, see also [25]. We sample independent equilibrium sequences using a Metropolis–Hastings algorithm by proposing spin flips at random sites and accepting them with probability

$$p = \min\left[1, \exp(-\Delta H/T)\right] \,, \tag{2}$$

where ΔH represents the energy variation associated to the flip, and T the Monte Carlo sampling temperature, see Fig. 1(a). This yields sequences that are only subject to the structural constraints defined by the graph. Selection strength is set by T, with higher T corresponding to weaker selection [25]. We model varying selection strength by switches between two temperatures T_1 and $T_2 > T_1$, under a telegraph process with timescale τ. We start from independent equilibrium sequences generated at temperature T_1, see Fig. 1(b).

To go beyond this minimal model, we generate realistic synthetic data using a Potts model inferred from natural sequences via bmDCA [26, 27], an approach whose generative power was experimentally validated [27]. Again, we start from sequences generated at T_1, and then switch between T_1 and T_2 via a telegraph process.

Finally, we consider a minimal model for the emergence of structure and function, taking into account phylogeny. We start from a random sequence, representing a non-functional ancestor, and we apply selection for contacts using the minimal model described above, and the Hamiltonian in Eq. 1, while evolving sequences on a binary branching tree, see Fig. 1(c).

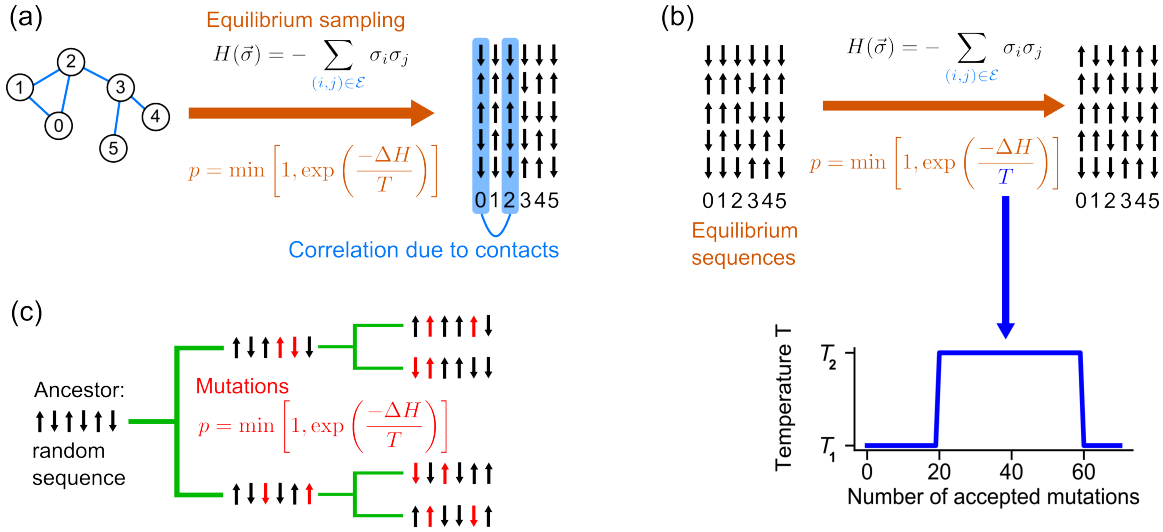

FIG. 1. **Minimal model**. (a) We model structural contacts as couplings set to 1 on the edges of an Erdős-Rényi random graph. Couplings between other nodes are set to 0. Equilibrium sampling of independent sequences is performed using a Metropolis–Hastings algorithm under the Hamiltonian in Eq. 1, with spin flip acceptance probability in Eq. 2, starting from random initial sequences. The MSA formed from the resulting sequences features correlations arising from couplings, shown between columns 0 and 2 (blue). (b) To generate sequences with fluctuating selection strength, we start from sequences generated at equilibrium with temperature T_1. We then evolve each of these sequences using the same Metropolis–Hastings algorithm as before, but with a sampling temperature T that switches between T_1 and T_2. (c) To minimally model the emergence of structure and function, we take a random sequence representing a non-functional ancestor, and we evolve it along a binary branching tree (green) where mutations (red) are accepted with the probability in Eq. 2, yielding selection for structural contacts. On each branch of the tree, μ mutations are accepted, with $\mu = 2$ in our schematic.

In all cases, we assess the impact of variable selection on inference from our synthetic data by considering the problem of structural contact inference using inferred Potts models. We compare results to those obtained from sets of independent equilibrium sequences.

Fluctuating selection facilitates inference. In our model, selection strength, tuned via the sampling temperature T, shapes sequences. For very low T, essentially, only mutations that decrease the energy given by the Hamiltonian in Eq. 1 are accepted. There is then little variability among sequences, which makes inference difficult. Conversely, for very high T, almost all mutations are accepted, generated sequences are extremely noisy, and inference is again difficult. Due to this tradeoff between variability and noise, inference is most successful at intermediate temperatures. In practice, for the present minimal model, it was shown that inference is most successful around and moderately above the ferromagnetic-paramagnetic phase transition [28], which occurs around $T = 4.2$ with our parameter values [25] (see also [29]).

How much does the time variability of selection pressures impact inference from sequence data? To answer this question within our minimal model, we start from a set of independent sequences equilibrated at $T_1 = 1$. We evolve them with a sampling temperature that switches between $T_1 = 1$ and $T_2 = 15$, following a telegraph process with timescale τ. Sequences are saved at regular (Monte Carlo) time points in the simulation, and contacts are inferred from them via Potts models.

Fig. 2 shows the True Positive (TP) fraction, i.e. the fraction of true contacts among the top N inferred Potts model couplings, where N is the actual number of contacts, versus the number of accepted mutations (Monte Carlo time), for different values of τ. Panel (a) shows results averaged over many realizations of the telegraph process and of the sequence evolution. We observe that the average TP fraction increases with time before plateauing. Hence, fluctuating selection strength is beneficial for inference performance. Moreover, higher TP fractions are obtained for faster temperature switches. Importantly, inference performance can become much better than when holding temperature fixed at either T_1 or T_2. This further confirms the beneficial role of variable selection strength on inference.

Fig. 2(b) shows that, in single realizations of the process, switching to strong selection (temperature T_1; white shade) after a weak selection phase (temperature T_2; gray shade) yields a transient improvement in inference performance. Indeed, increasing selection on sequences where noise partly obscures structural correlations decreases this noise, without immediately resulting in sequences lacking diversity. Thus, switching between weak and strong selection effectively enhances structural correlations in sequences. However, if more time is spent at a given temperature, inference performance reverts to its equilibrium performance at that temperature. The ob-

served transient improvement of inference performance results from out-of-equilibrium noise. Faster alternations enhance these out-of-equilibrium effects, which improves inference performance, see Fig. 2(a).

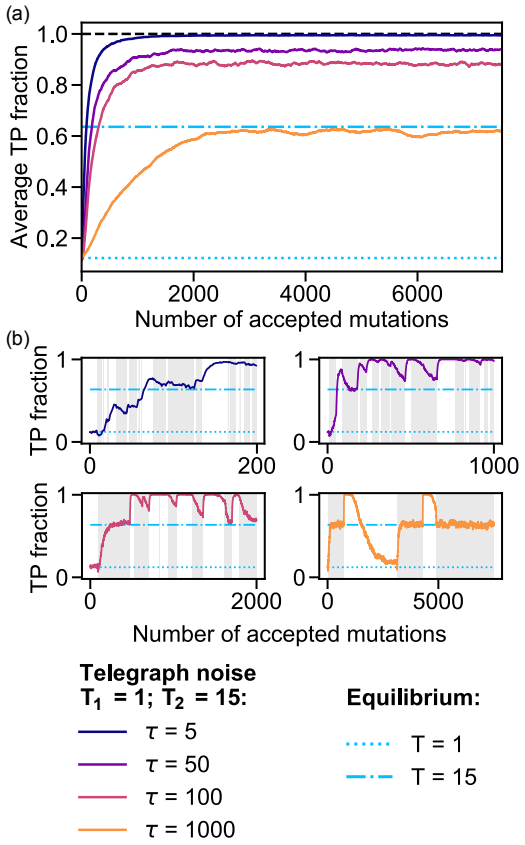

FIG. 2. **Impact of fluctuating selection strength on contact prediction.** TP fraction versus number of accepted mutations, while switching selection strength (i.e. the sampling temperature) via a telegraph process with timescale τ. (a) For each value of τ, the TP fraction is averaged over 1000 realizations of the telegraph process and data generation. (b) The TP fraction is shown for single realizations of the telegraph process and data generation. Gray background: T_2; white background: T_1. TP fractions for equilibrium sequences generated at T_1 and T_2 are shown for reference. Sequences are generated using our minimal model, see Fig. 1(a-b), with $T_1 = 1$ and $T_2 = 15$. For each realization, we generate an MSA of 2048 sequences of length 200 using an Erdős-Rényi random graph with probability 0.02 to represent contacts, and infer contacts via mean-field DCA (mfDCA) [3, 4, 25], with pseudocount 0.01.

Extension to realistic synthetic data. How do the observations made with our minimal model generalize to more realistic data? To address this question, we infer a Potts model from a natural MSA, corresponding to protein family PF0004 (AAA ATPase) from Interpro/Pfam [30] via bmDCA [26, 27]. Next, we generate sequences from this Potts model, at equilibrium and under fluctuating selection strength (see Model and Methods).

Fig. 3 shows the TP fraction versus the number of accepted mutations for this realistic synthetic data. As with our minimal model, we find that alternating between strong and weak selection yields better inference performance than keeping selection strength constant. Besides, quick alternations of phases at low and high temperature improve inference performance. Inference performance under variable selection then gets almost as good as that obtained at temperature $T = 1$, while two temperatures leading to lower inference performances were employed. Thus, our conclusions from the minimal model generalize to realistic synthetic data.

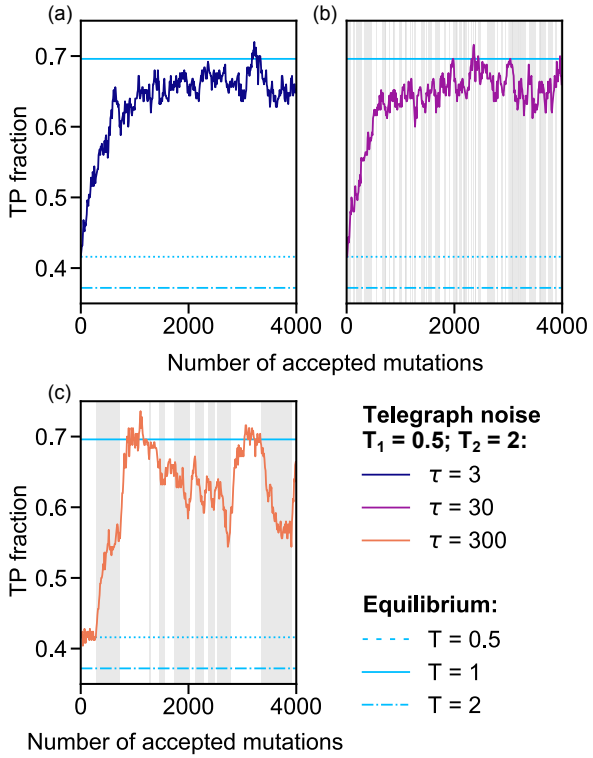

FIG. 3. **Impact of fluctuating selection strength on contact prediction: PF0004 family.** As in Fig. 2(b), the TP fraction is shown versus the number of accepted mutations under switching selection strength via a telegraph process, but for realistic data composed of MSAs of 70,000 sequences generated from a Potts model inferred on a natural MSA of 39,277 sequences from the PF0004 family (with length 132), following Refs. [26, 31]. Gray background: T_2; white background: T_1 (alternations not shown in the first panel for readability). Inference performances for equilibrium sequences generated at T_1, T_2, and $T = 1$, are shown for reference. Contact inference is performed using plmDCA [32, 33], with pseudocount 0.01 and no phylogenetic reweighting.

Minimal model for the emergence of function. Beyond fluctuating environments, out-of-equilibrium noise may be present when function and structure emerge from a non-functional ancestor. To address this case, we generate sequences starting from a random ancestral sequence,

with selection for contacts (at fixed T) in our minimal model and phylogeny modeled by a binary branching tree (see Model and Methods).

Fig. 4 shows the TP fraction versus the number μ of accepted mutations per branch in the phylogeny. For sampling temperatures $T < 4$, below the ferromagnetic-paramagnetic transition in our minimal model, we observe a maximum in the TP fraction for an intermediate value of μ. Inference then performs much better than the equilibrium TP fraction that it tends to for large μ. Hence, under strong selection, the out-of-equilibrium noise originating from the random ancestor helps contact inference. The beneficial effect of out-of-equilibrium noise for inference that we evidenced under fluctuating selection thus also exists during function emergence.

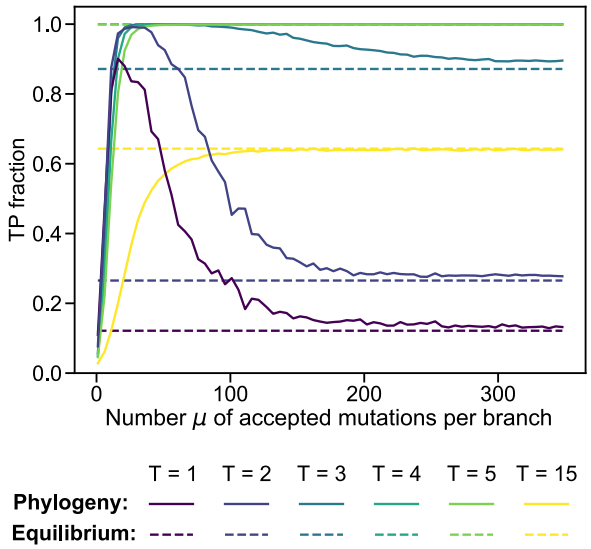

FIG. 4. **Impact of a non-functional ancestor on contact inference with phylogeny.** The TP fraction is shown versus the number μ of accepted mutations per branch of the phylogenetic tree. Sequences are generated as explained in Fig. 1(a,c). We generate sequences of length 200, using the same Erdős-Rényi graph as in Fig. 2. The phylogenetic tree comprises 11 generations, resulting in 2048 sequences at the leaves, which constitute our MSA, and μ mutations are accepted on each branch. We infer contacts via mfDCA [3, 4, 25] with pseudocount 0.01.

Discussion. Using a minimal model, we showed that alternating between weak and strong selection facilitates structural contact inference from sequences, compared to constant weak or strong selection. We demonstrated that this extends to realistic synthetic sequences. Next, we considered a minimal model of function and structure emergence in proteins, starting from a non-functional ancestor, and accounting for phylogeny. We showed that under strong selection, the presence of a non-functional ancestor leads to improved contact inference.

Under strong selection, the lack of variability in sequences makes coevolution-based inference challenging. Indeed, measuring correlations require some sequence diversity. Alternating with weak selection, or starting from a non-functional ancestor, both transiently increase diversity in sequence data, which facilitates inference. This demonstrates the beneficial impact of out-of-equilibrium noise on inference from protein sequence data.

Our results were obtained using synthetic data, allowing us to tweak selection pressures and to choose ancestral sequences. However, our extension from the minimal model to realistic data suggests that our results are generic and hold for natural data, where variable selection pressures are challenging to infer. Besides, while we considered the problem of structural contact inference in proteins, our findings fundamentally regard the amount of variability in data, which impacts coevolution-based inference. Several other problems can be addressed by such methods, including the inference of interaction partners [8–10, 34] and of functional sectors [12–15] in proteins, as well as RNA structure inference [35]. We expect out-of-equilibrium noise to be helpful for these inference problems too. Finally, while we focused on inference based on Potts models, coevolution is also at the heart of recent machine learning models, including in particular AlphaFold [7]. We thus expect out-of-equilibrium noise to be helpful for inference by these approaches too.

Here, we made the simplifying hypothesis that the Hamiltonian used for selection, and encoding structural contacts, remained fixed through evolution, and we varied the sampling temperature or took an out-of-equilibrium ancestral sequence. However, the objective function of selection, and in particular the target optimal structure, can also potentially change in time. Addressing this case would be an interesting perspective. Such changes of selection pressure should also provide out-of-equilibrium noise.

During natural evolution, environments change, due both to external conditions and to living organisms modifying their environments [19, 20]. Therefore, our findings are relevant for natural data, and provide insight into the causes of the practical success of coevolution-based inference. Beyond natural evolution, our results have implications for directed evolution in the laboratory [36]. In recent experiments [37, 38], structural prediction was performed from sequences evolved *in vitro* at constant selection strength. It was later shown [39] that the selection strength used in these experiments could be optimized to facilitate contact inference. Our present results suggest that varying selection strength during such experiments, e.g. by varying antibiotic concentration, would further facilitate inference.

Code accessibility. Our code is available at `https://github.com/Bitbol-Lab/OutOfEquilibrium-ProtSeq`.

This research was partly funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 851173, to A.-F. B.).

[1] S. D. Dunn, L. M. Wahl, and G. B. Gloor, Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction, Bioinformatics **24**, 333 (2008).

[2] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Identification of direct residue contacts in protein-protein interaction by message passing, Proc. Natl. Acad. Sci. U.S.A. **106**, 67 (2009).

[3] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Direct-coupling analysis of residue coevolution captures native contacts across many protein families, Proc. Natl. Acad. Sci. U.S.A. **108**, E1293 (2011).

[4] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, Protein 3D structure computed from evolutionary sequence variation, PLoS ONE **6**, e28766 (2011).

[5] J. P. Barton, E. De Leonardis, A. Coucke, and S. Cocco, ACE: adaptive cluster expansion for maximum entropy graphical model inference, Bioinformatics **32**, 3089 (2016).

[6] R. M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, and A. Rives, MSA Transformer, Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research, **139**, 8844 (2021).

[7] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, Highly accurate protein structure prediction with AlphaFold, Nature **596**, 583 (2021).

[8] L. Burger and E. van Nimwegen, Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method, Mol. Syst. Biol. **4**, 165 (2008).

[9] A.-F. Bitbol, R. S. Dwyer, L. J. Colwell, and N. S. Wingreen, Inferring interaction partners from protein sequences, Proc. Natl. Acad. Sci. U.S.A. **113**, 12180 (2016).

[10] A. F. Bitbol, Inferring interaction partners from protein sequences using mutual information, PLoS Comput. Biol. **14**, e1006401 (2018).

[11] U. Lupo, D. Sgarbossa, and A.-F. Bitbol, Pairing interacting protein sequences using masked language modeling, Proc. Natl. Acad. Sci. USA **121**, e2311887121 (2024).

[12] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, Protein sectors: evolutionary units of three-dimensional structure, Cell **138**, 774 (2009).

[13] O. Rivoire, K. A. Reynolds, and R. Ranganathan, Evolution-based functional decomposition of proteins, PLoS Comput. Biol. **12**, e1004817 (2016).

[14] S.-W. Wang, A.-F. Bitbol, and N. Wingreen, Revealing evolutionary constraints on proteins through sequence analysis, PLOS Comput. Biol. **15**, e1007010 (2019).

[15] N. Dietler, A. Abbara, S. Choudhury, and A.-F. Bitbol, Impact of phylogeny on the inference of functional sectors from protein sequence data, PLOS Comput. Biol. **20**, 1 (2024).

[16] J. A. de la Paz, C. M. Nartey, M. Yuvaraj, and F. Morcos, Epistatic contributions promote the unification of incompatible models of neutral molecular evolution, Proc. Natl. Acad. Sci. USA **117**, 5873 (2020).

[17] L. D. Bari, M. Bisardi, S. Cotogno, M. Weigt, and F. Zamponi, Emergent time scales of epistasis in protein evolution, Proc. Natl. Acad. Sci. USA **121**, e2406807121 (2024).

[18] M. De Leonardis, A. Pagnani, and P. Barrat-Charlaix, Reconstruction of ancestral protein sequences using autoregressive generative models, Mol. Biol. Evol. **42**, msaf070 (2025).

[19] V. Mustonen and M. Lässig, Molecular evolution under fitness fluctuations, Phys. Rev. Lett. **100**, 108101 (2008).

[20] G. Bell, Fluctuating selection: the perpetual renewal of adaptation in variable environments, Phil. Trans. R. Soc. B **365**, 87 (2010).

[21] A. Melbinger and M. Vergassola, The impact of environmental fluctuations on evolutionary fitness functions, Sci. Rep. **5**, 15211 (2015).

[22] K. Wienand, E. Frey, and M. Mobilia, Evolution of a fluctuating population in a randomly switching environment, Phys. Rev. Lett. **119**, 158301 (2017).

[23] A. Taitelbaum, R. West, M. Assaf, and M. Mobilia, Population dynamics in a changing environment: Random versus periodic switching, Phys. Rev. Lett. **125**, 048105 (2020).

[24] I. Meyer and N. M. Shnerb, Evolutionary dynamics in fluctuating environment, Phys. Rev. Res. **2**, 023308 (2020).

[25] N. Dietler, U. Lupo, and A.-F. Bitbol, Impact of phylogeny on structural contact inference from protein sequence data, J. R. Soc. Interface **20**, 20220707 (2023).

[26] M. Figliuzzi, P. Barrat-Charlaix, and M. Weigt, How Pairwise Coevolutionary Models Capture the Collective Residue Variability in Proteins?, Mol. Biol. Evol. **35**, 1018 (2018).

[27] 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, An evolution-based model for designing chorismate mutase enzymes, Science **369**, 440 (2020).

[28] V. Ngampruetikorn, V. Sachdeva, J. Torrence, J. Humplik, D. J. Schwab, and S. E. Palmer, Inferring couplings in networks across order-disorder phase transitions, Phys. Rev. Research **4**, 023240 (2022).

[29] C. A. Gandarilla-Pérez, P. Mergny, M. Weigt, and A.-F. Bitbol, Statistical physics of interacting proteins: Impact of dataset size and quality assessed in synthetic sequences, Phys. Rev. E **101**, 032413 (2020).

[30] J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn, and A. Bateman, Pfam: The protein families database in 2021, Nucleic Acids Res. **49**, D412 (2020).

[31] U. Lupo, D. Sgarbossa, and A.-F. Bitbol, Protein language models trained on multiple sequence alignments learn phylogenetic relationships, Nat. Commun. **13**, 6298 (2022).

[32] M. Ekeberg, C. Lovkvist, Y. Lan, M. Weigt, and E. Aurell, Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models, Phys. Rev. E **87**, 012707 (2013).

[33] M. Ekeberg, T. Hartonen, and E. Aurell, Fast pseudo-likelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences, J. Comput. Phys. **276**, 341 (2014).

[34] A. Gerardos, N. Dietler, and A. F. Bitbol, Correlations from structure and phylogeny combine constructively in the inference of protein partners from sequences, PLOS Comput. Biol. **18**, e1010147 (2022).

[35] M. B. Zerihun, F. Pucci, E. K. Peter, and A. Schug, pydca v1.0: a comprehensive software for direct coupling analysis of RNA and protein sequences, Bioinformatics **36**, 2264 (2019).

[36] F. H. Arnold, Directed evolution: Bringing new chemistry to life, Angew. Chem. Int. Ed. **57**, 4106 (2018).

[37] M. Fantini, S. Lisi, P. De Los Rios, A. Cattaneo, and A. Pastore, Protein Structural Information and Evolutionary Landscape by In Vitro Evolution, Mol. Biol. Evol. **37**, 1179 (2019).

[38] M. A. Stiffler, F. J. Poelwijk, K. P. Brock, R. R. Stein, A. Riesselman, J. Teyra, S. S. Sidhu, D. S. Marks, N. P. Gauthier, and C. Sander, Protein structure from experimental evolution, Cell Syst. **10**, 15 (2020).

[39] M. Bisardi, J. Rodriguez-Rivas, F. Zamponi, and M. Weigt, Modeling Sequence-Space Exploration and Emergence of Epistatic Signals in Protein Evolution, Mol. Biol. Evol. **39**, msab321 (2021).