Understanding Energy-Based Modeling of Proteins via an Empirically Motivated Minimal Ground Truth Model

Peter W Fields [*1,2] **Vudtiwat Ngampruetikorn** [*3] **Rama Ranganathan** [4,2,5] **David J Schwab** [3] **S E Palmer** [1,2,6]

Abstract

Energy-based models (EBM) of sequences of evolutionarily related families of proteins have the ability to learn the generic constraints necessary to make novel functional sequences, which have been validated by *in vivo* experiments. However, these learned energy functions require re-scaling by a temperature parameter in order to sample novel functional sequences. Here, we generate data from a minimal model motivated by a wide array of empirical evidence for a synergistic cluster of amino acids, or sector, within a sequence. We find our setting captures salient learning behaviors similar to those exhibited by EBMs fitted to real proteins, namely the necessity for temperature tuning to increase generative performance. We discuss how this guides insight into the functional sequence space of proteins.

1. Introduction

Proteins have evolved flexible ways to achieve the same function with significantly different amino acid sequences across many species (Starr & Thornton, 2016; Cocco et al., 2018). Understanding what principles underpin function yet support variation is of both fundamental and technological interest, see, e.g., (Göbel et al., 1994; Neher, 1994; Halabi et al., 2009; Morcos et al., 2011; Fowler & Fields, 2014). Answering this question requires not only a deep biological understanding but also the tools to unlock relevant insights hidden in the data. This has inspired cross-disciplinary efforts among scientists, bioengineers and machine learning practitioners, see, e.g., Rao et al. (2021a;b); Biswas et al. (2021); Marks et al. (2011); Hawkins-Hooker et al. (2021); Trinquier et al. (2021); Rives et al. (2021); Jumper et al. (2021); Lin et al. (2023); Lian et al. (2023); Madani et al. (2023); Ziegler et al. (2023); Sgarbossa et al. (2023); Lipsh-Sokolik et al. (2023).

Here we focus on energy-based modeling (EBM),[1] which is a current method used to build generative models of protein sequence data (Morcos et al., 2011; Cocco et al., 2011; Tubiana et al., 2019; Tagasovska et al., 2022). In particular, we concentrate on EBMs with a pairwise interacting energy function, which underlie direct coupling analysis (DCA), a commonly used technique for fitting a multiple sequence alignment (MSA) of protein sequences from a family of evolutionarily related organisms (Uguzzoni et al., 2017; Russ et al., 2020; Barrat-Charlaix et al., 2021). This model, also known as the Potts model in statistical physics (Nguyen et al., 2017; Cocco et al., 2018), is characterized by the energy function,

$$E(\mathbf{v} \mid \hat{\theta}) = -\sum\nolimits_i \hat{h}_i(v_i) - \sum\nolimits_{i<j} \hat{J}_{ij}(v_i, v_j) \quad (1)$$

where $\mathbf{v} = (v_1, v_2, \ldots, v_N)$ is a sequence with N positions and each position takes q discrete values, $v_i \in \{1, 2, \ldots, q\}$ with $q = 21$ for proteins (20 possible amino acids and one gap state). We let $\hat{\theta} = \{\hat{\mathbf{h}}, \hat{\mathbf{J}}\}$ denote the model parameters. The probability of a sequence is related to its energy via

$$\hat{P}(\mathbf{v} \mid \hat{\theta}) \propto \exp\left(-E(\mathbf{v} \mid \hat{\theta})\right).$$

Importantly, this does not explicitly model physico-chemical potentials governing atomic interactions. Instead, it captures only the information that is encoded in sequence statistics. This class of models has proved surprisingly successful in extracting the relevant constraints of functional proteins, leading to predictions of novel sequences that are verified to be functional in *in vivo* experiments (Russ et al., 2020).

Training EBMs on real MSAs and taking samples from them is a modeling and empirical challenge, however.[2] Although

[*]Equal contribution [1]Department of Physics, University of Chicago, Chicago, IL, USA [2]Center for Physics of Evolving Systems, University of Chicago, Chicago, IL, USA [3]Initiative for the Theoretical Sciences, The Graduate Center, CUNY, New York, NY, USA [4]Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, IL, USA [5]Pritzker School of Molecular Engineering, University of Chicago, Chicago, IL, USA [6]Organismal Biology and Anatomy, University of Chicago, Chicago, IL, USA. Correspondence to: S E Palmer <sepalmer@uchicago.edu>.

Accepted after peer-review at the 1st workshop on Synergy of Scientific and Machine Learning Modeling, SynS & ML ICML, Honolulu, Hawaii, USA. July, 2023.

[1]For a review of EBMs, see, e.g., Lecun et al. (2006).

[2]See Song & Kingma (2021) for a recent review of EBM training in more general settings.

Potts models are surprisingly expressive and have led to new insights about sequence data (Schug et al., 2009; Bravi et al., 2020), they place an unverifiable assumption on the interactions in the system. In fact, higher-order interactions are likely responsible for critical functions of proteins (Starr & Thornton, 2016; Poelwijk et al., 2019). Additionally, the high dimensionality and undersampling of sequence data necessitate regularization, which has a nontrivial effect on inference performance (Kleeorin et al., 2023). Furthermore, EBMs trained on finite samples struggle to imitate the sharp disparity between functional and nonfunctional proteins. The latter do not survive *in vivo* and thus correspond to zero probability or, equivalently, infinite energy, see Fig. 4. In practice, fitted models assign finite energy to all sequences, including nonfunctional ones, although with mostly higher energy; as a result, the synthesis of new sequences often returns many nonfunctional proteins. Increasing the fraction of functional sequences requires an *ad hoc* rescaling of the fitted energy function, $\hat{P}(\mathbf{v} \mid \hat{\theta}) \propto \exp\left(-E(\mathbf{v} \mid \hat{\theta})/T\right)$ with the temperature T set after training to be smaller than one, see Russ et al. (2020).

To explore the learning behaviors and generative performance of EBMs for sequence data, we develop a minimal model of a protein sequence based on the "sector hypothesis" (Lockless & Ranganathan, 1999; Russ et al., 2005; Socolich et al., 2005; Halabi et al., 2009; Reynolds et al., 2011; McLaughlin Jr et al., 2012; Reynolds et al., 2013; Teşileanu et al., 2015; Rivoire et al., 2016; Raman et al., 2016; Salinas & Ranganathan, 2018). Briefly, this hypothesis posits that some highly-correlated subset of amino acids (roughly 10-20%) are responsible for determining the function of a sequence in a given protein family through their collective state. Our model of functional sequences captures this important aspect by explicitly disallowing sequences that exceed the mutation threshold away from archetypal 'functional' patterns. We use the sequences generated from this 'ground-truth' model as the training data for Potts models. This setting allows for a relatively controlled investigation of the generative performance of the fitted Potts models. In particular, we ask how often the learned models produce novel functional sequences and how diverse the generated sequences are.

We show that our setting captures the salient learning behavior of EBMs fitted to real sequence data. We explore the generative performance and its dependence on the interplay between model selection via cross-validation and post-training temperature adjustments. We quantify the trade-off between functionality and novelty in sampled sequences by computing the false positive rate and entropy of fitted models. Finally, we discuss how the lessons from our study guide insight into our understanding of sequence-function relationships.

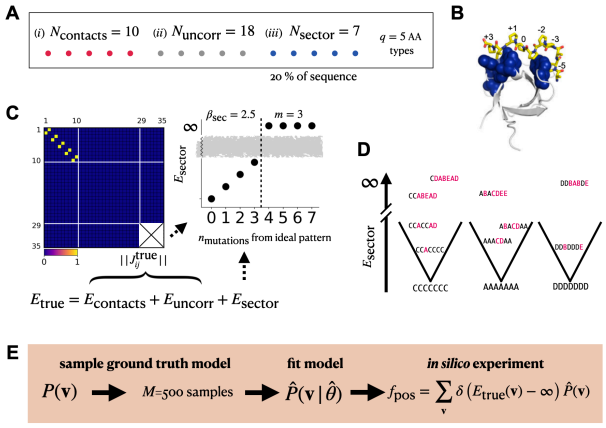

Figure 1. Minimal model for protein sequences (A) Our model captures three salient features of a sequence: (*i*) pairwise structural contacts, (*ii*) uncorrelated, independent positions, and (*iii*) a highly-correlated sector. (B) Example of a sector in the SH3 family of polyproline binding domains (PDB 2ABL). Image adapted from (Halabi et al., 2009). Sector positions (blue) are good predictors of binding sites for ligands (yellow). (C) Energy function of our model. The heatmap depicts the Frobenius norm of the pairwise couplings, $\|J_{ij}\| = (\sum_{a,b} J_{ij}(a,b)^2)^{1/2}$, for each residue pair i and j. The sector (positions 29-35) cannot be described by pairwise interactions and we model this feature separately; the energy of the sector grows linearly with slope β_{sec} with the number of mutations away from ideal patterns up to a threshold m beyond which the energetic cost diverges. (D) Schematic energy landscape of the sector. Exceeding the mutation threshold (here $m = 3$) makes a sequence nonfunctional, resulting in diverging energy and vanishing probability. (E) Fitting workflow. First, we sample M sequences from our minimal model (see A&C). We use these samples to train the pairwise EBM [Eq (1)], from which we generate synthetic sequences. The synthetic sequences that contain more mutations than the mutation threshold contribute to the false positive rate and are *in silico* analogs of nonfunctional proteins in *in vivo* experiments (Russ et al., 2020; Lian et al., 2023).

2. Methods

There is much evidence that mutations with strong deleterious effects on a protein's main biochemical functions are confined to a small subset of sequence positions (McLaughlin Jr et al., 2012; Salinas & Ranganathan, 2018; Poelwijk et al., 2019; Kleeorin et al., 2023). Via statistical analyses of MSAs of various protein families, it has been found that the amino acids in this subset are strongly correlated with each other and weakly correlated with those positions outside of the subset (Lockless & Ranganathan, 1999; Socolich et al., 2005; Russ et al., 2005; Halabi et al., 2009; Reynolds et al., 2011; 2013; Rivoire et al., 2016; Raman et al., 2016). This function-determining, strongly intra-correlated subset of positions, which usually comprises roughly 10-20% of

the sequence, defines the sector. See Fig. 1B for an example. Here we introduce a minimal model that captures this important aspect of protein sequences.

As shown in Fig. 1A, our model consists of a ground truth distribution defined on a sequence of $N = 35$ positions with $q = 5$ amino acid types. The sequence is divided into three parts that reflect the behavior of sequence data from experiments: (*i*) pairwise interactions important for structure but not involved in function (Morcos et al., 2011; Uguzzoni et al., 2017; Kleeorin et al., 2023), (*ii*) uncorrelated positions, and (*iii*) function-determining positions modeled by higher-order correlations as defined above, (Fig. 1A&C). It is possible in principle for a protein to have multiple sectors, sometimes overlapping. This represents an interesting extension to our framework; as a first step, in this work, we consider only one sector and its effects on learning and generative performance. We emphasize that all the above-defined features do not represent physico-chemical potentials of the underlying interactions, but rather a correlational structure that relates directly to fitness. Unfit non-functional sequences therefore correspond to sequences with low probability or high energy.

Of the 35 positions, 7 of them, which is 20 percent of positions, comprise the sector, whose energy landscape has 5 energy basins, each defined by unique, non-overlapping patterns (Fig. 1C & D). A linear function, whose slope is determined by β_{sec}, controls the energy cost of being $n_{\text{mutations}}$ away from an ideal pattern. Importantly, if the number of mutations exceeds a maximum m, the energy goes to infinity. When this occurs, it corresponds to a function-eliminating mutation, which is to say it is a sequence that cannot rescue function in any biological context. For our experiments, we choose $m = 3$ and $\beta_{\text{sec}} = 2.5$. Results are qualitatively insensitive to these choices. Ten positions (out of 35) correspond to structural, non-function determining interactions, modeled via Potts interactions such that $J_{ij}(a, b) = 1$ if $a = b$, and zero otherwise for five pairs of positions; all other positions (18 of 35) are uncorrelated (Fig. 1C). In addition, we set $h_i(a) = 0$ for all i and a.

The parameters m and β_{sec} represent the defining knobs of the minimal ground truth model that determine the shapes of the energy basins and extent of higher-order interactions within the sector. Thus, our empirically motivated model allows for direct control over the extent to which certain sequences will not be in the support of the fitted model. This ability to know which sequences are and are not in the ground truth support allows for a direct measure of how often a fitted model $\hat{P}(\mathbf{v} \mid \hat{\theta})$ will produce non-functional sequences. This false positive rate will serve as an *in silico* biological experiment, see Fig. 1E.

The models are fit via the ratio-matching algorithm (Hyvärinen, 2007), and are 5-fold cross validated to choose

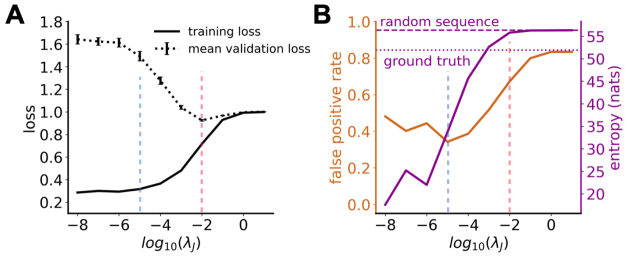

Figure 2. Validation error, false-positive rate, and sequence entropy reach an optimum for different regularization values. (A) Training (solid) and validation (dashed) losses *vs* regularization strength λ_J. (B) False positive rate (left axis) and entropy (right axis) *vs* λ_J. The vertical lines mark the minimum false positive rate (blue, $\lambda_J = 10^{-5}$) and minimum validation loss (red, $\lambda_J = 0.01$).

appropriate hyper-parameters. The objective and loss functions are

$$\text{obj}(\hat{\theta}) = \text{loss}(\hat{\theta}) + \lambda_h \sum_{i,a} \|h_i(a)\|^2 + \lambda_J \sum_{i<j,a,b} \|J_{ij}(a,b)\|^2,$$

$$\text{loss}(\hat{\theta}) = \mathop{\mathbb{E}}_{\mathbf{v}\sim\text{Data}} \sum\nolimits_{\mathbf{v}'} A(\mathbf{v}, \mathbf{v}')\, \sigma\left(E(\mathbf{v} \mid \hat{\theta}) - E(\mathbf{v}' \mid \hat{\theta})\right)^2$$

where $\sigma(x) = 1/(1 + e^{-x})$, $\mathbf{v}'$ are sequences *not* in data, and $A(\mathbf{v}, \mathbf{v}') = 1$ if $\mathbf{v}$ and $\mathbf{v}'$ differ in one position and $A(\mathbf{v}, \mathbf{v}') = 0$ otherwise. In the following, we take $M = 500$ samples with $\beta_{\text{sec}} = 2.5$ and $m = 3$ and set $\lambda_h = 100$ for all learning tasks.

The entropies of the corresponding fitted models are calculated via annealed importance sampling (Neal, 2001). Entropy serves as a measure of the fitted model's estimate for the size of functional sequence space. One may state that the goal of learning, as defined by current experimental work, is to reduce the false positive rate as much as possible while keeping the entropy high. This ensures that functional sequences can be of a wide variety, and not simply a memorization of the training data.

3. Results

Figure 2 illustrates the properties of Potts models [Eq (1)] with parameters fitted to samples from our minimal model for protein sequences (Fig. 1). We see that the validation loss is smallest at an intermediate regularization strength $\lambda_J = 0.01$ (Fig. 2A). But this value does not yield the lowest false positive rate, which occurs at $\lambda_J = 10^{-5}$ (Fig. 2B). The ability to generate unseen, yet functional, sequences requires both low false positive rates and high entropy. These competing objectives make model selection nontrivial. Moreover, the standard loss function does not seem to effectively capture relevant performance measures. Taken together, our results exemplify the well-documented challenges of balancing multiple generative objectives and encoding them in

a loss function.

To facilitate the ability to sample functional sequences from these learned models, the parameters for each are systematically rescaled via temperature, T. As shown in Fig. 3A and B, decreasing T improves the overlap between energy distributions of the training data and sampled data. Fig. 3C displays the decrease in false positive rate associated with lowering T. The accompanying decrease in entropy and false positive rate mirrors empirical results found in real data and experiments, where good overlap with training data energy distributions corresponds to the ability to generate novel functional sequences (Russ et al., 2020).

Why must T be lowered to produce functional sequences? To understand this effect further, we track its impact on sequence entropy and false positive rates for models trained at several regularization strengths. Fig. 3D shows the results of this analysis. Here each curve is parameterized by T. For weak regularization, decreasing temperature lowers the entropy with almost no effect on the false positive rate. At intermediate regularizations, a beneficial trade-off occurs as false positive rates decrease while entropy remains high before dropping off steeply at the lowest temperatures. For strong regularization, the under-fit models lose the ability to generate functional sequences as temperature drops, as not enough structure in the data is found.

4. Discussion

Figure 4 provides an overview of the intuition behind post-training temperature adjustments for improved generative performance. In Fig. 4A, we show a schematic 'true' energy landscape, in which functional sequences (black) occupy low-energy states and non-functional sequences (red) are at infinitely higher energy ($\Delta E \approx \infty$). Ideally, a well-trained model should distinguish functional from non-functional sequences and assign a large energy gap between them. In practice, a tradeoff exists between classification correctness and confidence. Weak regularization results in overfitting, characterized by high misclassification (i.e., many nonfunctional sequences in the energy basins) at high confidence (large energy gap), see Fig. 4B. Strong regularization, on the other hand, yields underfitted models, which encode little relevant information in the data and thus cannot classify any sequences with confidence (no energy gap), see Fig. 4D. A moderately regularized model has high classification correctness but low confidence (small energy gap), see Fig. 4C. Sampling from such a model, despite its low misclassification, can give spurious sequences since their energies, and probabilities, can still be comparable to those of functional ones. In this case, decreasing the temperature during sampling serves as a strategy to increase the confidence of a model and reduce the false positive rate.

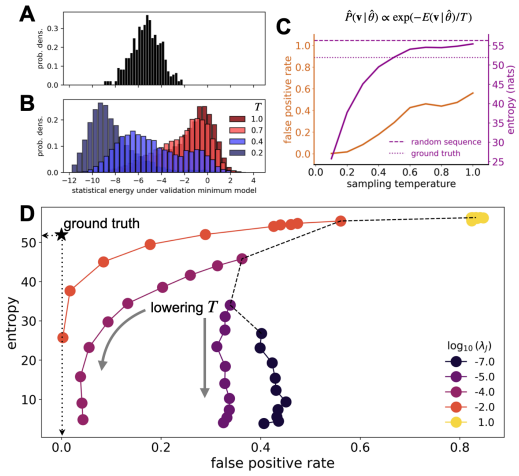

Figure 3. Temperature dependence at the validation minimum. Histograms of statistical energy of training data (A) and sampled sequences at different temperatures (B) under the fitted validation loss-minimum model. (C) Dependence of false positive rate and entropy of validation minimum model on sampling temperature. Lower temperature produces more functional sequences at the cost of lowering entropy, similar to empirical results in real data (Russ et al., 2020). (D) Curves of entropy vs. false positive rate, each parameterized by sampling temperature, for selected models trained at different regularizations. Models at $T = 1$ connected by dotted black line. Note the optimal trade-off between false positive rate and entropy, which corresponds to points in the upper left corner of the graph near the ground truth, occurs as the temperature is lowered on the validation minimum model corresponding to $\lambda_J = 0.01$.

We have developed a minimal model that recapitulates the strong, high-order coupling between amino acids in a sector. Under-sampling from this ground truth model (as one does in real sequence analysis) and fitting the 'wrong' Potts models via standard DCA techniques has reproduced the mysterious need to lower temperature in these pairwise models in order to generate functional sequences. We offer the preliminary insight that even though validation minimum models optimally predict what sequence must be low energy, it does not adequately segregate them from high energy states, and temperature must therefore be lowered. This suggests that our ground truth model could be further exploited in order to understand the optimal training and sampling techniques for real protein sequence data.

Acknowledgments

We would like to thank the reviewers for useful comments and feedback. P.F. would like to thank B. Hoshal for useful comments in editing the manuscript and E. Rouviere and K. Bojanek for useful conversations during conception of the work. This work was supported in part by the National

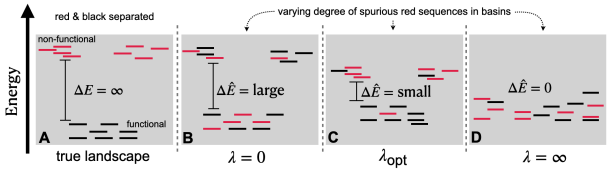

Figure 4. A cartoon of how temperature and regularization interact in EBMs. (A) Under the ground truth model, functional sequences (black) are low energy, and non-functional sequences (red) are high energy, separated by an energy gap $\Delta E \approx \infty$. (B) When under-sampled, an unregularized model $\lambda = 0$, estimates a large energy gap $\Delta\hat{E}$ at the expense of misclassifying functional and non-functional sequences–red and black mixed. (C) For a λ found via cross-validation, red and black sequences are optimally placed on correct sides of the energy gap. This comes at the cost of a smaller value of $\Delta\hat{E}$, which necessitates the lowering of temperature to sample functional sequences, as shown in Fig. 3. (D) For no regularization, $\Delta\hat{E}$ goes to zero, and all functional and non-functional sequences have similar energies.

Science Foundation, through the Center for the Physics of Biological Function (PHY-1734030); and by the National Institutes of Health BRAIN initiative (R01EB026943)

Broader Impact

Characterizing the underpinning rules governing protein structure and function has far-reaching consequences for medical and sustainability goals. Once these rules are understood, the ability to custom-design proteins can allow for better therapeutics and offers natural solutions to issues such as carbon scrubbing the atmosphere or designing renewable energy storage. Insofar as our model shown here represents a step towards this understanding, it can beneficially contribute to meaningful societal impacts. Furthermore, our minimal model represents a mapping from sequence to fitness based on extensive empirical evidence. This is a departure from many approaches in machine learning for proteins that focus on sequence to structure mapping. Benchmarking protein models via this approach could improve the understanding of fitting and sampling beyond energy-based modeling.

References

Barrat-Charlaix, P., Muntoni, A. P., Shimagaki, K., Weigt, M., and Zamponi, F. Sparse generative modeling via parameter reduction of boltzmann machines: application to protein-sequence families. *Physical Review E*, 104(2): 024407, 2021.

Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M., and Church, G. M. Low-n protein engineering with data-efficient deep learning. *Nature Methods*, 18(4):389–396, 2021. doi: 10.1038/s41592-021-01100-y.

Bravi, B., Ravasio, R., Brito, C., and Wyart, M. Direct coupling analysis of epistasis in allosteric materials. *PLoS computational biology*, 16(3):e1007630, 2020.

Cocco, S., Monasson, R., and Sessak, V. High-dimensional inference with the generalized hopfield model: Principal component analysis and corrections. *Physical Review E*, 83(5):051123, 2011.

Cocco, S., Feinauer, C., Figliuzzi, M., Monasson, R., and Weigt, M. Inverse statistical physics of protein sequences: a key issues review. *Reports on Progress in Physics*, 81 (3):032601, 2018.

Fowler, D. M. and Fields, S. Deep mutational scanning: a new style of protein science. *Nature Methods*, 11(8): 801–807, 2014. doi: 10.1038/nmeth.3027.

Göbel, U., Sander, C., Schneider, R., and Valencia, A. Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, 18(4):309–317, 1994. doi: https://doi.org/10.1002/prot.340180402.

Halabi, N., Rivoire, O., Leibler, S., and Ranganathan, R. Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138(4):774–786, 2009.

Hawkins-Hooker, A., Depardieu, F., Baur, S., Couairon, G., Chen, A., and Bikard, D. Generating functional protein variants with variational autoencoders. *PLOS Computational Biology*, 17(2):e1008736, 02 2021. doi: 10.1371/journal.pcbi.1008736.

Hyvärinen, A. Some extensions of score matching. *Computational statistics & data analysis*, 51(5):2499–2512, 2007.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kleeorin, Y., Russ, W. P., Rivoire, O., and Ranganathan, R. Undersampling and the inference of coevolution in proteins. *Cell Systems*, 14(3):210–219, 2023.

Lecun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. *A tutorial on energy-based learning*. MIT Press, 2006.

Lian, X., Praljak, N., Ferguson, A. L., and Ranganathan, R. Deep-learning generative models enable design of synthetic orthologs of a signaling protein. *Biophysical Journal*, 122(3):311a, 2023.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Lipsh-Sokolik, R., Khersonsky, O., Schröder, S. P., de Boer, C., Hoch, S.-Y., Davies, G. J., Overkleeft, H. S., and Fleishman, S. J. Combinatorial assembly and design of enzymes. *Science*, 379(6628):195–201, 2023. doi: 10.1126/science.ade9434.

Lockless, S. W. and Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–299, 1999.

Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos, J. L., Xiong, C., Sun, Z. Z., Socher, R., Fraser, J. S., and Naik, N. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 2023. doi: 10.1038/s41587-022-01618-2.

Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., and Sander, C. Protein 3d structure computed from evolutionary sequence variation. *PLOS ONE*, 6(12):e28766, 12 2011. doi: 10.1371/journal.pone.0028766. URL https://doi.org/10.1371/journal.pone.0028766.

McLaughlin Jr, R. N., Poelwijk, F. J., Raman, A., Gosal, W. S., and Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature*, 491(7422):138–142, 2012.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

Neal, R. M. Annealed importance sampling. *Statistics and computing*, 11:125–139, 2001.

Neher, E. How frequent are correlated changes in families of protein sequences? *Proceedings of the National Academy of Sciences*, 91(1):98–102, 1994. doi: 10.1073/pnas.91.1.98.

Nguyen, H. C., Zecchina, R., and Berg, J. Inverse statistical problems: from the inverse ising problem to data science. *Advances in Physics*, 66(3):197–261, 2017.

Poelwijk, F. J., Socolich, M., and Ranganathan, R. Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nature communications*, 10(1):4213, 2019.

Raman, A. S., White, K. I., and Ranganathan, R. Origins of allostery and evolvability in proteins: a case study. *Cell*, 166(2):468–480, 2016.

Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., and Rives, A. Transformer protein language models are unsupervised structure learners. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=fylclEqgvgd.

Rao, R. M., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., Sercu, T., and Rives, A. Msa transformer. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8844–8856. PMLR, 2021b. URL https://proceedings.mlr.press/v139/rao21a.html.

Reynolds, K. A., McLaughlin, R. N., and Ranganathan, R. Hot spots for allosteric regulation on protein surfaces. *Cell*, 147(7):1564–1575, 2011.

Reynolds, K. A., Russ, W. P., Socolich, M., and Ranganathan, R. Evolution-based design of proteins. In *Methods in enzymology*, volume 523, pp. 213–235. Elsevier, 2013.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118.

Rivoire, O., Reynolds, K. A., and Ranganathan, R. Evolution-based functional decomposition of proteins. *PLoS computational biology*, 12(6):e1004817, 2016.

Russ, W. P., Lowery, D. M., Mishra, P., Yaffe, M. B., and Ranganathan, R. Natural-like function in artificial ww domains. *Nature*, 437(7058):579–583, 2005.

Russ, W. P., Figliuzzi, M., Stocker, C., Barrat-Charlaix, P., Socolich, M., Kast, P., Hilvert, D., Monasson, R., Cocco, S., Weigt, M., et al. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369(6502):440–445, 2020.

Salinas, V. H. and Ranganathan, R. Coevolution-based inference of amino acid interactions underlying protein function. *elife*, 7:e34300, 2018.

Schug, A., Weigt, M., Onuchic, J. N., Hwa, T., and Szurmant, H. High-resolution protein complexes from integrating genomic information with molecular simulation. *Proceedings of the National Academy of Sciences*, 106(52):22124–22129, 2009.

Sgarbossa, D., Lupo, U., and Bitbol, A.-F. Generative power of a protein language model trained on multiple sequence alignments. *eLife*, 12:e79854, 2023. doi: 10.7554/eLife.79854.

Socolich, M., Lockless, S. W., Russ, W. P., Lee, H., Gardner, K. H., and Ranganathan, R. Evolutionary information for specifying a protein fold. *Nature*, 437(7058):512–518, 2005.

Song, Y. and Kingma, D. P. How to train your energy-based models, 2021.

Starr, T. N. and Thornton, J. W. Epistasis in protein evolution. *Protein science*, 25(7):1204–1218, 2016.

Tagasovska, N., Frey, N. C., Loukas, A., Hotzel, I., Lafrance-Vanasse, J., Kelly, R. L., Wu, Y., Rajpal, A., Bonneau, R., Cho, K., Ra, S., and Gligorijevic, V. A pareto-optimal compositional energy-based model for sampling and optimization of protein sequences. In *NeurIPS 2022 AI for Science: Progress and Promises*, 2022. URL https://openreview.net/forum?id=U2rNXaTTXPQ.

Teşileanu, T., Colwell, L. J., and Leibler, S. Protein sectors: Statistical coupling analysis versus conservation. *PLOS Computational Biology*, 11(2):e1004091, 02 2015. doi: 10.1371/journal.pcbi.1004091.

Trinquier, J., Uguzzoni, G., Pagnani, A., Zamponi, F., and Weigt, M. Efficient generative modeling of protein sequences using simple autoregressive models. *Nature Communications*, 12(1):5800, 2021. doi: 10.1038/s41467-021-25756-4.

Tubiana, J., Cocco, S., and Monasson, R. Learning compositional representations of interacting systems with restricted boltzmann machines: Comparative study of lattice proteins. *Neural computation*, 31(8):1671–1717, 2019.

Uguzzoni, G., John Lovis, S., Oteri, F., Schug, A., Szurmant, H., and Weigt, M. Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis. *Proceedings of the National Academy of Sciences*, 114(13):E2662–E2671, 2017.

Ziegler, C., Martin, J., Sinner, C., and Morcos, F. Latent generative landscapes as maps of functional diversity in protein sequence space. *Nature Communications*, 14(1):2222, 2023. doi: 10.1038/s41467-023-37958-z.