Research Statement for Alan N. Amin

Motivation

Probabilistic models of biological sequences are classical tools used to design drugs [SJCK21, SRK+21], make predictions about human health [FND+21], and learn basic biology [RIM18, TGN+22]. Sequence data is high dimensional so a probabilistic model must make biological assumptions to predict and infer. However, these assumptions can come at the cost of the flexibility of the model, fundamentally limiting its ability to make accurate predictions and learn new biology. Modern sequencing efforts and high-throughput experimentation are generating an ever-increasing amount of sequence data, in principle providing increasing information to learn the complexity of real sequence data. To leverage this wealth of data I plan to build nonparametric models and tests of sequences that incorporate prior biological knowledge while remaining flexible using modern machine learning methods. I will build methods to perform efficient, flexible, and reliable prediction and inference from DNA and protein data, at large and small scale, and in supervised and unsupervised settings.

Current work

* denotes equal contribution.

Building flexible models for sequence data

Amin A N*, Weinstein E N*, Marks D S. A generative nonparametric Bayesian model for whole genomes, NeurIPS, 2021

Huge sequencing experiments have generated a wealth of genomics data in human populations, microbiomes, and across life [HEM+21, TLH+07, PTM07]. To compare, test, and interpret these datasets, and use these datasets to design and predict the properties of new sequences, we wish to fit them with generative probabilistic models. Parametric autoregressive models are commonly used to fit large genomic datasets as they can incorporate biological priors and scale to large sequence datasets [CSIT04]. However these models are strongly misspecified – that are not nearly flexible enough to accurately fit genomic data. In this paper we developed a scalable nonparametric Bayesian model for sequence data that is flexible in theory and in practice. The model is based on a conjugate prior on the space of all autoregressive models. This prior has two hyperparameters: first, a particular parametric autoregressive model is "embedded" as the centre of the prior, so a practitioner can build in their biological assumptions; the second hyperparameter determines how "concentrated" the prior is around its centre, allowing the posterior to relax away from the embedded model if it is misspecified. We proved that our model is flexible in theory – it converges to data distribution under very weak assumptions. To demonstrate the practical utility of this method, we scaled our model to giga-bases of genomic data using large-scale k-mer counters and showed substantially higher accuracy than parametric autoregressive models. Lastly, we used the model to interpret complex variation and test for changes in large genomic datasets.

Amin A N, Weinstein E N*, Marks D S*. Biological Sequence Kernels with Guaranteed Flexibility. preprint, 2023

Sequence kernels are applied extensively in biology for a number of statistical procedures including regression for drug design [SMG+22] and hypothesis testing to learn biology [BGR+06]. Their appeal is that they 1) make biological assumptions and 2) are in principle flexible and can therefore reliably leverage large modern datasets to learn complex patterns. Given this appeal, there are a variety of sequence kernels that make diverse biological assumptions. However, the purported flexibility of these kernel procedures has not been demonstrated theoretically. Surprisingly we find that using most popular off-the-shelf kernels can result in inconsistent statistical procedures in theory and in practice. The issue is that these kernels can fail to have certain desirable theoretical properties, such as "universality"; there is however a lack of theoretical tools to prove these properties for kernels in infinite discrete spaces. We proved that one property previously described in the literature - "having discrete masses" - implies many desirable kernel properties and can be proven using symmetries possessed by many popular sequence kernels. We proved if and when four popular types of sequence kernels

have discrete masses: kernels that compare sequences position-by-position [SRR07], alignment kernels [Hau99], k-mer spectrum kernels [LEN02], and deep kernels [YWBA18]. For kernels that do not have discrete masses, we developed alternative kernels that capture the same biological assumptions but have discrete masses. Finally, we showed that the alternative kernels fix severe pathologies in previous tests and models.

Evaluating density estimation models of sequence data

Weinstein E N*, Amin A N*, Frazer J, Marks D S. Non-identifiability and the blessings of misspecification in models of molecular fitness and phylogeny. NeurIPS (Oral), 2022

Unsupervised models of protein sequences learn the distribution of sequences seen in nature to accurately predict how mutations affect the fitness of a protein [FND+21, RIM18]. Model predictions are based on an assumed relationship between density estimation and fitness – that sequences are observed in nature in proportion to their fitness. However, this assumption is problematic in theory as nuisance effects – such as the phylogenetic relatedness of sequences and biased sequencing efforts – may perturb the observed distribution of sequences in nature. In this paper we ask when fitness is identifiable in theory and what allows its accurate prediction in practice. We showed that, in theory, fitness is almost never identifiable from phylogenetically correlated data; however, in practice, models make accurate fitness effect predictions. One hypothesis that explains this paradoxical observation is that phylogenetic and other nuisance effects are small. If this were the case then better density estimation would lead to better fitness prediction. To address this, we compared a panel of popular sequence models and observed the opposite – models that performed worse density estimation performed substantially better fitness effect prediction. We argue that better density estimators fit nuisance effects which harms their fitness estimation; some models however are misspecified in such a way that they cannot fit nuisance effects such as phylogeny. We demonstrated this effect in simulation and found that models that perform good fitness estimation do so because they are misspecified in a helpful way. This suggests that to build better fitness prediction models one should think about misspecification or devise alternative identification strategies.

Amin A N, Weinstein E N*, Marks D S*. A Kernelized Stein Discrepancy for Biological Sequences. ICML, 2023

Once a generative model has been fitted to sequence data, it is necessary to assess the quality of the fit. The standard way is to draw sequences from the model and test whether they match training sequences using a set of statistics such as the number of hydrophobic residues in a protein sequence [SRK+21, ZFM+22]. However, a generative model that passes this test may still poorly fit the data by not capturing biology outside of these statistics. As well, for some models it is difficult to sample sequences to perform the test. To evaluate generative models in a flexible way, we built a nonparameteric goodness-of-fit test for models of biological sequences. This test is consistent and does not require sequences from the model. The test is based on 1) building a stochastic process for sequences that is stationary for the model distribution, 2) applying it to the data points, and 3) using a kernel to evaluate "how stationary" the datapoints are. Biological assumptions about how the model fails to fit the data can be incorporated in the choice of kernel. We carefully built the stochastic process and choose kernels such that the test is also flexible in theory – it is consistent – given assumptions that are satisfied by popular models of biological sequences. Lastly, we tested the fit of state-of-the-art deep learning models on protein sequence families and showed that our test is very sensitive, often requiring fewer than 100 sequences to detect that these models do not fit the data they were trained on.

Future work

Flexible models with biological priors make accurate predictions and learn new biology. Designing general methods to build such models is a machine learning challenge whose solution would greatly impact how biologists learn from sequences. Two future practical projects addressing this problem involve building flexible models to learn the biology of the most poorly understood proteins across life; and building neural network architectures that incorporate biological assumptions to learn the biology of the 98% of the human genome that does not code for proteins. Two future theoretical projects

addressing this problem involve finding realistic assumptions that establish practical convergence rates for optimization and sampling algorithms, and kernel approximations of sequences.

Causal, flexible models of protein function and evolution

To learn how a protein acts or evolves, biologists experimentally measure the effects of mutations on the function of the protein. However, experiments can be too time-consuming when measurements of viral proteins are needed quickly, like early in a pandemic [TGN+22]. Experiments can also be impossible to design for poorly understood "disordered" proteins which are connected to neurodegeneration [Uve15]. To quickly and automatically predict the impact of mutations on a protein's fitness, biologists have long turned to learning from sequence data: they use generative probabilistic models to infer patterns that are conserved across sequences observed in nature – patterns that are conserved are likely important for the fitness of the protein. In practice, these models struggle to make accurate predictions in the settings where they can have the largest impact: viral and disordered proteins [RIM18]. The poor performance of these models is due to inductive biases that do not hold for viral and disordered proteins. One may build models with biases appropriate for viral and disordered proteins to get better predictions. However, these biases may be too restrictive to make accurate predictions, especially as there is little domain knowledge; they also fundamentally restrict the biology one can infer to known phenomena built into the parameters. Ideally one could use modern machine learning methods to *learn* the correct biases with more flexible models. However, we have shown that this approach is fundamentally limited in practice due to a lack of *identifiability* [WAFM22]: in theory, it is impossible to distinguish patterns that are conserved across sequences in nature because they determine protein fitness from patterns that are conserved due to phylogentic correlations – that some proteins happen to come from related organisms.

To accurately infer fitness and learn new biology with flexible models, I propose to build new methods of learning from sequences in nature that are not limited by non-identifiability. To do so, I will build causal identification strategies for fitness. I will devise identification strategies based on observing many proteins from the same organisms; I will use some proteins as proxies for confounding non-fitness effects [WB21, PPR20]. Once I have built an identifications strategy, I will infer fitness with modern machine learning methods that fit into this identification strategy, and are also *flexible*. Identifying harmful mutations of a sequence is a task in identifying *out-of-distribution* (OOD) data. A central challenge in building accurate OOD detection methods is incorporating accurate inductive biases [KIW20]. Thus I will build methods that incorporate accurate inductive biases. I will test the ability of models to predict experimental measurements of these proteins; models that predict experiments well can replace experiments for biologists. Finally, I will use these models to infer the biology governing viral protein evolution and disordered protein function.

Biologically inspired neural network architectures for sequence data

To understand the mechanisms of disease and make diagnoses in the clinic, biologists build models and perform experiments to predict whether mutations in human proteins cause disease. However, only 2% of the human genome codes for proteins; the remaining 98% is not tractable to the same models and experimentation [AWM21, KAI+20] but is responsible for most heritable common disease [FBSG+15]. To predict the effects of mutations in the "non-coding" 98%, biologists now use high-dimensional, high-resolution experimental measurements of "epigenetic features": biologists first fit a deep neural network model to predict epigenetic features from sequence; then they use the model to predict the effect of a mutation on the features – if a mutation is predicted to strongly disrupt the epigenetic features then it is likely to cause disease [AWS+21, ZPT+19, ZTY+18, CWTZ22]. While these models make accurate predictions for human genomic sequences they were trained on, in practice, they generalize poorly to predict the effects of new mutations [LSP+19]. They are thus strongly limited in their ability to predict whether new mutations cause disease. For models to generalize well, they require accurate inductive biases so that they may fit simple explanations of the data. Deep learning models in other domains incorporate inductive biases such as equivariance [KT18] and Gaussian process priors [SZSG19]. However models for predicting epigenetic features lack even the most basic inductive biases for biological sequences. For example, insertions and deletions are common mutations in the human genome [KFT+20] that often do not cause disease but result in large changes to the input of neural networks. To build models that learn the biology of the non-coding genome and generalize to accurately predict disease, I propose to build neural network architectures with biases for biological

sequences.

A promising way to build in such biases is to take inspiration from sequence kernels which encode diverse biological assumptions in their architectures; for example, the parameters of the alignment kernel control the assumed effect of an insertion or deletion [Hau99]. Using the theory of sequence kernels we have developed [AWM23a], I will combine the two types of models by building "neuralized" sequence kernels [WHSX16] as well as neural network architectures that limit to sequence kernels at infinite width [JGH18]. Once I have built these biased models, I will test their ability to predict experimental measurements of non-coding mutations. These models will allow for a unique and powerful way for predicting disease and inferring the biology governing the non-coding genome.

Theoretical guarantees for optimization, sampling, and kernels for sequences

A crucial step in designing biological sequences is often optimizing a function or sampling from a distribution. However, optimizing and sampling are notoriously challenging in discrete space due to high-dimensionality, lack of derivatives, and multimodality. To address each of these challenges, a number of diverse sampling algorithms have been built for discrete data [Zan20, GSH+21, SDXR22]. To pick the correct algorithm for each situation, and develop new algorithms, it is important that we have theoretical results that describe how quickly each algorithm converges under different realistic assumptions. For algorithms for Euclidean data, there are convergence results that come from assumptions such as convexity [MTB22, WWJ16, MCC+21]. Algorithms for sequence data on the other hand do not come with convergence rate guarantees as it is not yet clear what assumptions are appropriate to make. In [AWM23c], on the way to proving the consistency of a hypothesis test, we found assumptions that 1) guarantee fast convergence for an MCMC algorithm and 2) are obeyed by popular sequence distributions. We proved our results by using a Lyapunov function approach that was previously used to study random walks on infinite graphs and groups [Woe00]. To analyze the properties of existing algorithms and develop new practical algorithms for sequences under different realistic assumptions, I will extend my theoretical methods to prove convergence rates for different sequence optimization and sampling methods.

Kernels are useful tools for practical machine learning, and for theoretically studying or designing neural network architectures. Different kernels learn better under different assumptions. Theoretical results help guide us to pick the correct kernel for each situation. However, beyond recently proving consistency under very general conditions [AWM23b], we have no such results for commonly used sequence kernels. For kernels on Euclidean space, there are results describing which functions each kernel can approximate more quickly [GKKW02]: by calculating the eigenvalues in Mercer's theorem, we can obtain convergence rates. To evaluate which kernels should be used under which assumptions, and design new kernels and kernel-inspired neural networks appropriate under different assumptions, I will prove approximation rates for sequence kernels.

During my PhD I developed theoretical tools to build methods that reliably learn from sequences under very general assumptions. These future theoretical directions will find the specific assumptions that describe the setting of biological sequences and use them to build methods that learn efficiently on real data.

References

[AWM21] Alan N Amin, Eli N Weinstein, and Debora S Marks. A generative nonparametric bayesian model for whole genomes. *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, December 2021.

[AWM23a] Alan N Amin, Eli N Weinstein, and Debbie S Marks. Kernels with guaranteed flexibility for reliable machine learning on biological sequences. February 2023.

[AWM23b] Alan Nawzad Amin, Eli Nathan Weinstein, and Debora Susan Marks. Biological sequence kernels with guaranteed flexibility. April 2023.

[AWM23c] Alan Nawzad Amin, Eli Nathan Weinstein, and Debora Susan Marks. A kernelized stein discrepancy for biological sequences. *40th International Conference on Machine Learning, ICML 2023*, page to appear, 2023.

[AWS+21] Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, and Julia Zeitlinger. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.*, 53(3):354–366, March 2021.

[BGR+06] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–57, July 2006.

[CSIT04] Niranjan Chakravarthy, A Spanias, L D Iasemidis, and K Tsakalis. Autoregressive modeling and feature analysis of DNA sequences. *EURASIP J. Appl. Signal Processing*, 2004(1):13–28, 2004.

[CWTZ22] Kathleen M Chen, Aaron K Wong, Olga G Troyanskaya, and Jian Zhou. A sequence-based global map of regulatory activity for deciphering human genetics. *Nat. Genet.*, 54(7):940–949, July 2022.

[FBSG+15] Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, Stephan Ripke, Felix R Day, Shaun Purcell, Eli Stahl, Sara Lindstrom, John R B Perry, Yukinori Okada, Soumya Raychaudhuri, Mark J Daly, Nick Patterson, Benjamin M Neale, and Alkes L Price. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.*, 47(11):1228–1235, 2015.

[FND+21] Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly Brock, Yarin Gal, and Debora S Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, November 2021.

[GKKW02] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer New York, 2002.

[GSH+21] Will Grathwohl, Kevin Swersky, Milad Hashemi, David Duvenaud, and Chris J Maddison. Oops I took a gradient: Scalable sampling for discrete distributions. 2021.

[Hau99] David Haussler. Convolution kernels on discrete structures UCSC CRL. 1999.

[HEM+21] Bjarni V Halldorsson, Hannes P Eggertsson, Kristjan H S Moore, Hannes Hauswedell, Ogmundur Eiriksson, Magnus O Ulfarsson, Gunnar Palsson, Marteinn T Hardarson, Asmundur Oddsson, Brynjar O Jensson, Snaedis Kristmundsdottir, Brynja D Sigurpalsdottir, Olafur A Stefansson, Doruk Beyter, Guillaume Holley, Vinicius Tragante, Arnaldur Gylfason, Pall I Olason, Florian Zink, Margret Asgeirsdottir, Sverrir T Sverrisson, Brynjar Sigurdsson, Sigurjon A Gudjonsson, Gunnar T Sigurdsson, Gisli H Halldorsson, Gardar Sveinbjornsson, Kristjan Norland, Unnur Styrkarsdottir, Droplaug N Magnusdottir, Steinunn Snorradottir, Kari Kristinsson, Emilia Sobech, Gudmar Thorleifsson, Frosti Jonsson, Pall Melsted, Ingileif Jonsdottir, Thorunn Rafnar, Hilma Holm, Hreinn Stefansson, Jona Saemundsdottir, Daniel F Gudbjartsson, Olafur T Magnusson, Gisli Masson, Unnur Thorsteinsdottir, Agnar Helgason, Hakon Jonsson, Patrick Sulem, and Kari Stefansson. The sequences of 150,119 genomes in the UK biobank. November 2021.

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. June 2018.

[KAI+20] Jason C Klein, Vikram Agarwal, Fumitaka Inoue, Aidan Keith, Beth Martin, Martin Kircher, Nadav Ahituv, and Jay Shendure. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat. Methods*, 17(11):1083–1091, 2020.

[KFT+20] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, Laura D Gauthier, Harrison Brand, Matthew Solomonson, Nicholas A Watts, Daniel Rhodes, Moriel Singer-Berk, Eleina M England, Eleanor G Seaby, Jack A Kosmicki, Raymond K Walters, Katherine Tashman, Yossi Farjoun, Eric Banks, Timothy Poterba, Arcturus Wang, Cotton Seed, Nicola Whiffin, Jessica X Chong, Kaitlin E Samocha, Emma Pierce-Hoffman, Zachary Zappala, Anne H O'Donnell-Luria, Eric Vallabh Minikel, Ben Weisburd, Monkol Lek, James S Ware, Christopher Vittal, Irina M Armean, Louis Bergelson, Kristian Cibulskis, Kristen M Connolly, Miguel Covarrubias, Stacey Donnelly, Steven Ferriera, Stacey Gabriel, Jeff Gentry, Namrata Gupta, Thibault Jeandet, Diane Kaplan, Christopher Llanwarne, Ruchi Munshi, Sam Novod, Nikelle Petrillo, David Roazen, Valentin Ruano-Rubio, Andrea Saltzman, Molly Schleicher, Jose Soto, Kathleen Tibbetts, Charlotte Tolonen, Gordon Wade, Michael E Talkowski, Carlos A Aguilar Salinas, Tariq Ahmad, Christine M Albert, Diego Ardissino, Gil Atzmon, John Barnard, Laurent Beaugerie, Emelia J Benjamin, Michael Boehnke, Lori L Bonnycastle, Erwin P Bottinger, Donald W Bowden, Matthew J Bown, John C Chambers, Juliana C Chan, Daniel Chasman, Judy Cho, Mina K Chung, Bruce Cohen, Adolfo Correa, Dana Dabelea, Mark J Daly, Dawood Darbar, Ravindranath Duggirala, Josée Dupuis, Patrick T Ellinor, Roberto Elosua, Jeanette Erdmann, Tõnu Esko, Martti Färkkilä, Jose Florez, Andre Franke, Gad Getz, Benjamin Glaser, Stephen J Glatt, David Goldstein, Clicerio Gonzalez, Leif Groop, Christopher Haiman, Craig Hanis, Matthew Harms, Mikko Hiltunen, Matti M Holi, Christina M Hultman, Mikko Kallela, Jaakko Kaprio, Sekar Kathiresan, Bong Jo Kim, Young Jin Kim, George Kirov, Jaspal Kooner, Seppo Koskinen, Harlan M Krumholz, Subra Kugathasan, Soo Heon Kwak, Markku Laakso, Terho Lehtimäki, Ruth J F Loos, Steven A Lubitz, Ronald C W Ma, Daniel G MacArthur, Jaume Marrugat, Kari M Mattila, Steven McCarroll, Mark I McCarthy, Dermot McGovern, Ruth McPherson, James B Meigs, Olle Melander, Andres Metspalu, Benjamin M Neale, Peter M Nilsson, Michael C O'Donovan, Dost Ongur, Lorena Orozco, Michael J Owen, Colin N A Palmer, Aarno Palotie, Kyong Soo Park, Carlos Pato, Ann E Pulver, Nazneen Rahman, Anne M Remes, John D Rioux, Samuli Ripatti, Dan M Roden, Danish Saleheen, Veikko Salomaa, Nilesh J Samani, Jeremiah Scharf, Heribert Schunkert, Moore B Shoemaker, Pamela Sklar, Hilkka Soininen, Harry Sokol, Tim Spector, Patrick F Sullivan, Jaana Suvisaari, E Shyong Tai, Yik Ying Teo, Tuomi Tiinamaija, Ming Tsuang, Dan Turner, Teresa Tusie-Luna, Erkki Vartiainen, James S Ware, Hugh Watkins, Rinse K Weersma, Maija Wessman, James G Wilson, Ramnik J Xavier, Benjamin M Neale, Mark J Daly, and Daniel G MacArthur. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.

[KIW20] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Why normalizing flows fail to detect out-of-distribution data. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, number Article 1728 in NIPS'20, pages 20578–20589, Red Hook, NY, USA, December 2020. Curran Associates Inc.

[KT18] Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2747–2755. PMLR, 2018.

[LEN02] Christina Leslie, Eleazar Eskin, and William Stafford Noble. The spectrum kernel: a string kernel for SVM protein classification. *Pac. Symp. Biocomput.*, pages 564–575, 2002.

[LSP+19] Li Liu, Maxwell D Sanderford, Ravi Patel, Pramod Chandrashekar, Greg Gibson, and Sudhir Kumar. Biological relevance of computationally predicted pathogenicity of noncoding variants. *Nat. Commun.*, 10(1):330, January 2019.

[MCC+21] Yi-An Ma, Niladri S Chatterji, Xiang Cheng, Nicolas Flammarion, Peter L Bartlett, and Michael I Jordan. Is there an analog of nesterov acceleration for gradient-based MCMC? *Bernoulli*, 27(3), May 2021.

[MTB22] Céline Moucer, Adrien Taylor, and Francis Bach. A systematic approach to lyapunov analyses of continuous-time models in convex optimization. May 2022.

[PPR20] Aahlad Puli, Adler J Perotte, and Rajesh Ranganath. Causal estimation with functional confounders. *Adv. Neural Inf. Process. Syst.*, 33:5115–5125, December 2020.

[PTM07] Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, 35(SUPPL. 1):61–65, 2007.

[RIM18] Adam J Riesselman, John B Ingraham, and Debora S Marks. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods*, 15(10):816–822, 2018.

[SDXR22] Haoran Sun, Hanjun Dai, Wei Xia, and Arun Ramamurthy. Path auxiliary proposal for MCMC in discrete space. *ICLR 2022*, May 2022.

[SJCK21] Sam Sinai, Nina Jain, George M Church, and Eric D Kelsic. Generative AAV capsid diversification by latent interpolation. *bioRxiv*, page 2021.04.16.440236, 2021.

[SMG+22] Samuel Stanton, Wesley Maddox, Nate Gruver, Phillip Maffettone, Emily Delaney, Peyton Greenside, and Andrew Gordon Wilson. Accelerating bayesian optimization for biological sequence design with denoising autoencoders. March 2022.

[SRK+21] Jung-Eun Shin, Adam J Riesselman, Aaron W Kollasch, Conor McMahon, Elana Simon, Chris Sander, Aashish Manglik, Andrew C Kruse, and Debora S Marks. Protein design and variant prediction using autoregressive generative models. *Nat. Commun.*, 12(1):2403, April 2021.

[SRR07] Sören Sonnenburg, Gunnar Rätsch, and Konrad Rieck. Large-scale learning with string kernels. In *Large-Scale Kernel Machines*. The MIT Press, 2007.

[SZSG19] Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. Functional variational bayesian neural networks. March 2019.

[TGN+22] Nicole N Thadani, Sarah Gurev, Pascal Notin, Noor Youssef, Nathan J Rollins, Chris Sander, Yarin Gal, and Debora S Marks. Learning from pre-pandemic data to forecast viral antibody escape. July 2022.

[TLH+07] Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project. *Nature*, 449(7164):804–810, 2007.

[Uve15] Vladimir N Uversky. Intrinsically disordered proteins and their (disordered) proteomes in neurodegenerative disorders. *Front. Aging Neurosci.*, 7:18, March 2015.

[WAFM22] Eli N Weinstein, Alan N Amin, Jonathan Frazer, and Debora S Marks. Non-identifiability and the blessings of misspecification in models of molecular fitness and phylogeny. *Adv. Neural Inf. Process. Syst.*, December 2022.

[WB21] Yixin Wang and David Blei. A proxy variable view of shared confounding. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10697–10707. PMLR, 2021.

[WHSX16] Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In Arthur Gretton and Christian C Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 370–378, Cadiz, Spain, 2016. PMLR.

[Woe00] Wolfgang Woess. *Random Walks on Infinite Graphs and Groups*. Cambridge University Press, February 2000.

[WWJ16] Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *Proc. Natl. Acad. Sci. U. S. A.*, 113(47):E7351–E7358, 2016.

[YWBA18] Kevin K Yang, Zachary Wu, Claire N Bedbrook, and Frances H Arnold. Learned protein embeddings for machine learning. *Bioinformatics*, 34(15):2642–2648, August 2018.

[Zan20] Giacomo Zanella. Informed proposals for local MCMC in discrete spaces. *J. Am. Stat. Assoc.*, 115(530):852–865, April 2020.

[ZFM+22] Jan Zrimec, Xiaozhi Fu, Azam Sheikh Muhammad, Christos Skrekas, Vykintas Jauniskis, Nora K Speicher, Christoph S Börlin, Vilhelm Verendel, Morteza Haghir Chehreghani, Devdatt Dubhashi, Verena Siewers, Florian David, Jens Nielsen, and Aleksej Zelezniak. Controlling gene expression with deep generative design of regulatory DNA. *Nat. Commun.*, 13(1):5099, August 2022.

[ZPT+19] Jian Zhou, Christopher Y Park, Chandra L Theesfeld, Aaron K Wong, Yuan Yuan, Claudia Scheckel, John J Fak, Julien Funk, Kevin Yao, Yoko Tajima, Alan Packer, Robert B Darnell, and Olga G Troyanskaya. Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat. Genet.*, 51(6):973–980, 2019.

[ZTY+18] Jian Zhou, Chandra L Theesfeld, Kevin Yao, Kathleen M Chen, Aaron K Wong, and Olga G Troyanskaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.*, 50(8):1171–1179, 2018.