Accelerating Protein Design Using Autoregressive Generative Models

Adam Riesselman [*1 2 3] **Jung-Eun Shin** [*2] **Aaron Kollasch** [*2] **Conor McMahon** [4] **Elana Simon** [5 6]
Chris Sander [7] **Aashish Manglik** [8 9] **Andrew Kruse** [4] **Debora Marks** [1 2 10]

Abstract

A major biomedical challenge is the interpretation of genetic variation and the ability to design functional novel sequences. Since the space of all possible genetic variation is enormous, there is a concerted effort to develop reliable methods that can capture genotype to phenotype maps. State-of-art computational methods rely on models that leverage evolutionary information and capture complex interactions between residues. However, current methods are not suitable for a large number of important applications because they depend on robust protein or RNA alignments. Such applications include genetic variants with insertions and deletions, disordered proteins, and functional antibodies. Ideally, we need models that do not rely on assumptions made by multiple sequence alignments. Here we borrow from recent advances in natural language processing and speech synthesis to develop a generative deep neural network-powered autoregressive model for biological sequences that captures functional constraints without relying on an explicit alignment structure. Application to unseen experimental measurements of 42 deep mutational scans predicts the effect of insertions and deletions while matching state-of-art missense mutation prediction accuracies. We then test the model on single domain antibodies, or nanobodies, a complex target for alignment-based models due to the highly variable complementarity determining regions. We fit the model to a naïve llama immune repertoire and generate a diverse, optimized library of 10^5 nanobody sequences for experimental validation. Our results demonstrate the power of the 'alignment-free' autoregressive model in mutation effect prediction and design of traditionally challenging sequence families.

[*]Equal contribution [1]Department of Biomedical Informatics, Harvard Medical School (HMS) [2]Department of Systems Biology, HMS [3]Currently at insitro [4]Department of Biological Chemistry and Molecular Pharmacology, HMS [5]Harvard College [6]Currently at Reverie Labs [7]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute [8]Department of Pharmaceutical Chemistry, University of California San Francisco (UCSF) [9]Department of Anesthesia and Perioperative Care, UCSF [10]Broad Institute of Harvard and MIT. Correspondence to: Debora Marks <debora_marks@hms.harvard.edu>.

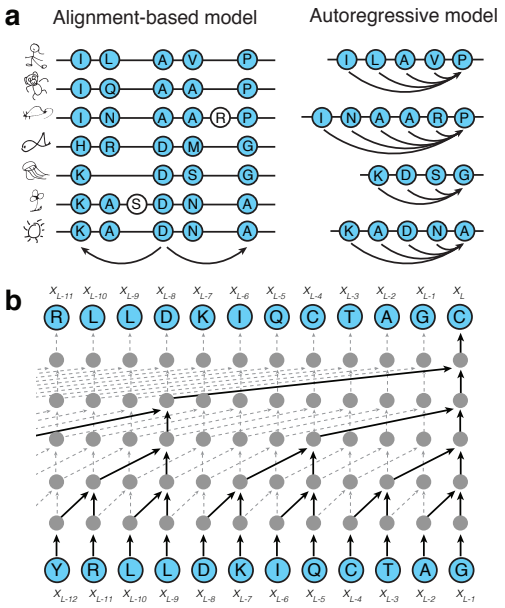

Figure 1. Autoregressive models of biological sequences. a. Instead of finding correlations between columns in a multiple sequence alignment (left), the autoregressive model predicts a residue given all the preceding positions (right). b. Causal dilated convolutions are used to model the autoregressive likelihood.

1. Introduction

Designing and generating biomolecules with known function is a major goal of biotechnology and biomedicine. Experimental methods to characterize sequence variants are expensive, labor-intensive, and time consuming because the sequence space of potentially functional proteins is astronomically large. High-throughput techniques have opened up opportunities to phenotype thousands of variants in a single experiment (Kosuri & Church, 2014; Fowler & Fields, 2014), but the sampled diversity of sequences pales in comparison to that found in nature. However, the information contained in natural sequences can be used to intelligently narrow the search space to bias library design for functional biomolecules.

Across evolution, there are large amounts of genetic variation available for many protein families. These sequences

contain information about what contributes to a stable, functional protein. Generative models of protein sequences are powerful tools in learning these constraints to predict structure, mutation effects, and residue-residue interactions (Hopf et al., 2017; Mann et al., 2014; Riesselman et al., 2018). Yet these methods to date are dependent on alignments of homologous sequences, which impose a forced structure on the data that can introduce artifacts and exclude important information, particularly by alignment gaps caused by insertions and deletions (indels). This is especially a challenge for sequences with regions of high variability or low complexity, including antibodies and disordered proteins.

Antibodies are valuable tools for molecular biology and therapeutics because they can detect low concentrations of target antigens with high sensitivity and specificity (Muyldermans, 2013). Single-domain antibodies, or nanobodies, are composed solely of the variable domain of the canonical antibody heavy chain. Because of their many applications, there is demand for the rational design of novel antibodies. Recently, phage and yeast display methods have allowed for high-throughput screening of libraries of antibodies and nanobodies (Säll et al., 2016; McMahon et al., 2018). Although these synthetic libraries can be used to isolate a high-affinity sequence, they often contain a substantial fraction of non-functional proteins because library construction methods lack higher-order sequence constraints. Generative models can be applied to learn such constraints from large repertoires of natural antibody and nanobody sequences. These sequences contain four highly conserved framework regions interspersed with three variable length, hypermutated complementarity determining regions (CDRs), which makes high-confidence alignments difficult to construct (Muyldermans, 2013). Here, we apply an autoregressive generative model to this family to predict and generate probable protein sequences and design an optimized nanobody library by leveraging a model trained on a natural immune repertoire.

2. Related Work

Many applications in text-to-speech (Graves et al., 2013; Wang et al., 2017) and translation (Bahdanau et al., 2014; Sutskever et al., 2014) tackle context dependent word and audio prediction and generation using generative models with an autoregressive likelihood (Sutskever et al., 2011; LeCun et al., 2015). In this way, an audio clip is decomposed into discrete time steps, a sentence into words, and a protein sequence into amino acid residues.

Recently, autoregressive models have been applied across protein families to learn general features and predict mutation effects in unaligned protein sequences (Alley et al., 2019; Rives et al., 2019). However, accurate mutation effect prediction with these models require semi-supervised training using a held-out fraction of labeled experimental mutagenesis data. In contrast, our method is fully unsupervised and does not require such data to make predictions or generate sequences. Though the conditional generation of protein sequences has been demonstrated on a small number of proteins (Costello & Martin, 2019), we present how to leverage autoregressive models to design hundreds of thousands of diverse, fit protein sequences using computationally tractable techniques.

3. Methods

3.1. Model

Sequences are represented by a 21 letter alphabet for proteins or 5 letter alphabet for RNAs, one for each residue type and a 'start/stop' character. The log-likelihood for a sequence is the sum of the cross-entropy between the true residue at each position and the predicted distribution over possible residues, conditioned on the previous characters. We adopt a residual causal dilated convolutional neural network architecture with 6 blocks of 9 dilated convolutional layers and both weight normalization (Salimans & Kingma, 2016) and layer normalization (Ba et al., 2016). To help prevent overfitting, we use L2 regularization on the weights and place Dropout layers ($p = 0.5$) immediately after each of the 6 residual blocks (Srivastava et al., 2014). We use a batch size of 30 and a channel size of 48 for all sequence families tested. Six models are built for each family: three replicates in both the N-to-C and C-to-N directions, respectively. Each model is trained for 250,000 updates using Adam with default parameters (Kingma & Ba, 2014), and the gradient norm is clipped (Pascanu et al., 2013) to 100.

3.2. Data collection

Mutation effects, sequence families, and previous effect predictions for validation were curated from published work (Riesselman et al., 2018). The naïve llama immune repertoire was acquired from McCoy et al. (2014). Due to the large number of sequences in the llama immune repertoire, sequence weights were approximated using Linclust (Steinegger & Söding, 2018) by clustering sequences at both 80% and 90% sequence identity thresholds.

3.3. Nanobody library generation

Using the N-to-C terminus model trained on llama nanobody sequences, we generated 37,744,800 CDR3 sequences by ancestral sampling (Sutskever et al., 2011), conditioned on the germline framework-CDR1-CDR2 sequence and continued until generation of the stop character. Duplicates of the training set or generated sequences and those not matching the reference final beta

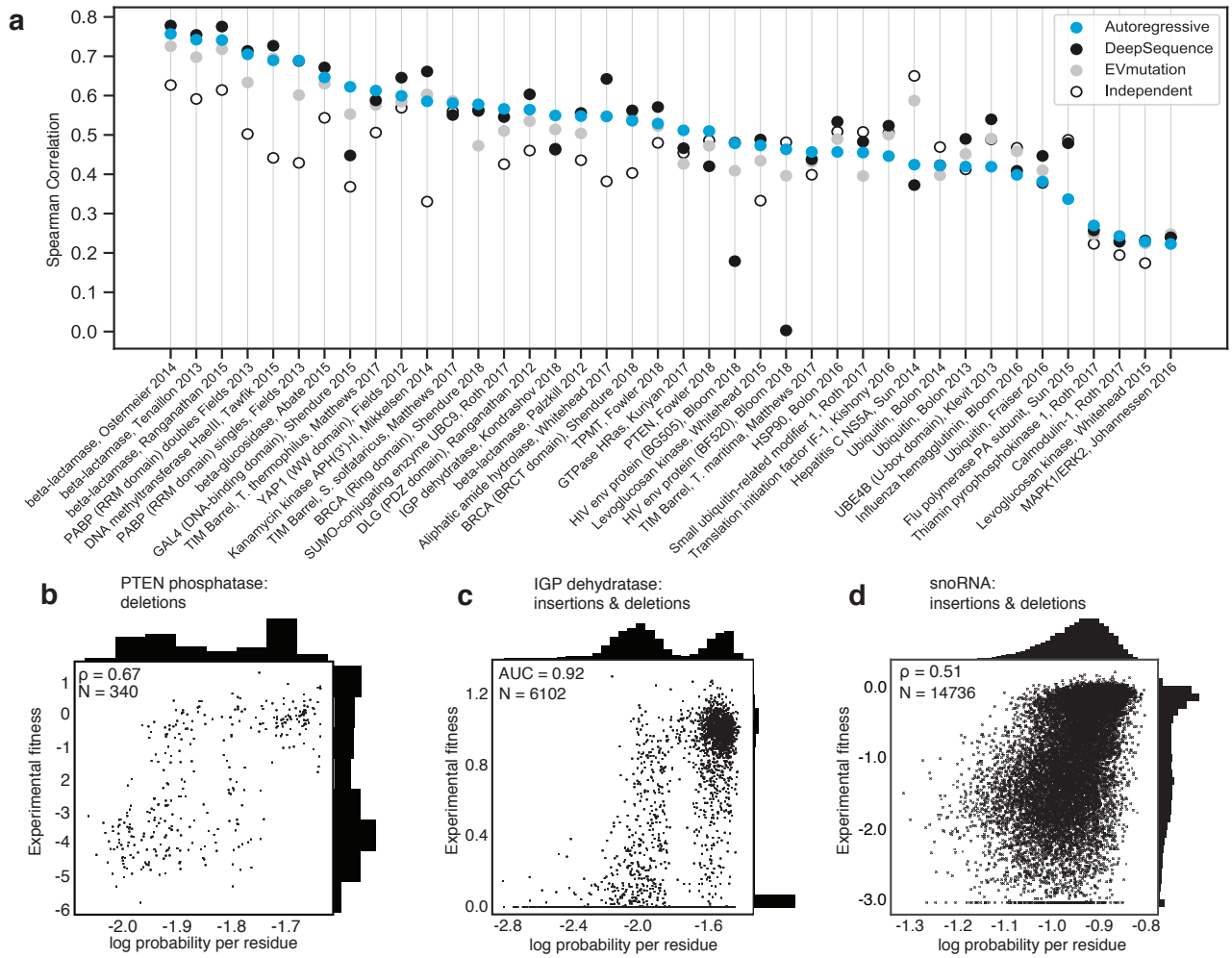

Figure 2. The autoregressive model predicts the effects of insertions, deletions, and missense mutations. a. The autoregressive model matches state-of-art missense mutation effect prediction. Length normalized log probabilities predict the effects of insertions and deletions in (b) PTEN phosphatase, (c) IGP dehydratase, and (d) snoRNA.

strand were excluded. CDR3 sequences were also removed if they contained glycosylation (NxS and NxT) sites, asparagine deamination (NG) motifs, or sulfur-containing amino acids (cysteine and methionine), resulting in 3,690,554 sequences.

From this large number of sequences, we then sought to choose roughly 200,000 CDR3 sequences that are both deemed fit by the model and as diverse from one another as possible to cover the largest amount of sequence space. First, we featurized these sequences into fixed length, L2 normalized kmer vectors with kmers of size 1, 2, and 3. We then used BIRCH clustering (Zhang et al., 1996) to find diverse members of the dataset in O(n) time. We used a diameter threshold of 0.575, resulting in 382,675 clusters. From the cluster centroids, we chose the 185,836 most probable sequences for final library construction.

4. Results

4.1. An autoregressive generative model of biological sequences

Protein sequences observed in organisms today result from mutation and selection for functional, folded proteins over time scales of a few days to a billion years. Generative models can be used to parameterize this view of evolution. Namely, they express the probability that a sequence $\boldsymbol{x}$ would be generated by evolution as $p(\boldsymbol{x} \mid \boldsymbol{\theta})$, where parameters $\boldsymbol{\theta}$ capture the constraints essential to functional sequences.

We use an autoregressive model to obviate the need for structured alignments. The probability distribution $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ can be decomposed into the product of conditional probabilities on previous characters along a sequence of length L

(Figure 1a) via an autoregressive likelihood:

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = p(x_1 \mid \boldsymbol{\theta}) \prod_{i=2}^{L} p(x_i \mid x_1, \ldots, x_{i-1}; \boldsymbol{\theta})$$

Many different neural network architectures can model an autoregressive likelihood, including attention-based models (Vaswani et al., 2017) and recurrent neural networks (Sutskever et al., 2011). However, we encountered exploding gradients (Pascanu et al., 2013) during training on long sequence families with LSTM (Hochreiter & Schmidhuber, 1997) or GRU (Cho et al., 2014) architectures. Instead, we parameterize this process with dilated convolutional neural networks (Figure 1b), which are feed-forward deep neural networks that aggregate long-range dependencies in sequences over an exponentially-large receptive field (van den Oord et al., 2016; Kalchbrenner et al., 2016; Gupta & Rush, 2017). The causal structure of the model allows for efficient training to a set of sequences, inference of mutation effects, and sampling of new sequences.

4.2. The generative model predicts experimental mutation effects

Once the model is fit to a family of protein sequences, the log-ratio of likelihoods of individual sequences:

$$\log \frac{p\left(\boldsymbol{x}^{(\text{Mutant})} \mid \boldsymbol{\theta}\right)}{p\left(\boldsymbol{x}^{(\text{Wild-type})} \mid \boldsymbol{\theta}\right)}$$

can be used to estimate the plausibility of mutant sequence $\boldsymbol{x}^{(\text{Mutant})}$ relative to its wild-type, un-mutated counterpart, $\boldsymbol{x}^{(\text{Wild-type})}$. This log-ratio has been shown to be predictive of mutation effects (Hopf et al., 2017; Riesselman et al., 2018). Importantly, this approach is fully unsupervised: rather than learning from experimental mutation effects, we can learn evolutionary constraints using only the space of natural sequences.

We compare log-ratio predictions from our model to 40 deep mutational scans across 33 different proteins, totaling 690,257 individual sequences. The autoregressive model is consistently able to predict the effects of mutations across a wide array of proteins and experimental assays (Figure 2a). When compared to other generative models trained on alignments of the same sequences, the autoregressive model is able to consistently match or outperform a model with only site-independent terms (30/40 datasets) and the EVmutation model with pairwise dependencies (30/40 datasets; Hopf et al., 2017); and it competitively matches the state-of-the-art results of DeepSequence (19/40 datasets; Riesselman et al., 2018).

Previous alignment-based generative models are constrained to predicting the effects of missense mutations.

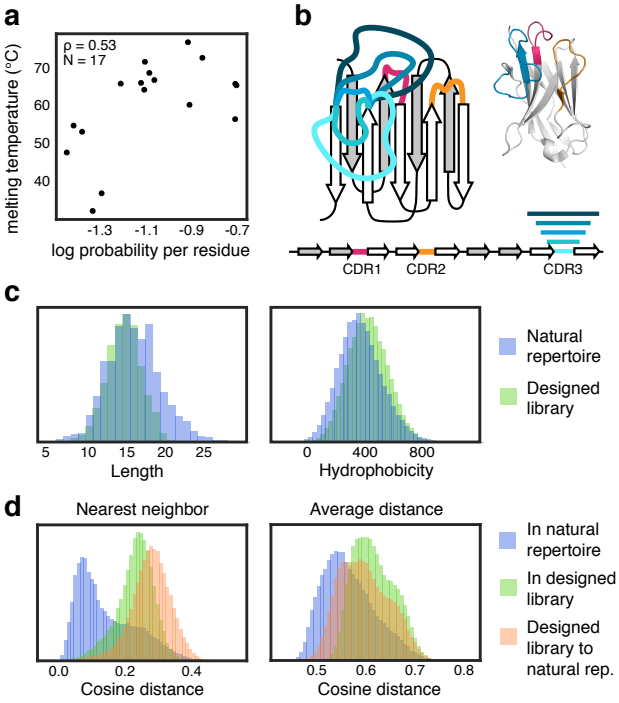

Figure 3. The model predicts nanobody stability and is capable of generating an optimized library. a. The model, trained on naïve sequences, predicts nanobody stability as measured by melting temperature. b. Conditioned on the framework-CDR1-CDR2 sequence, a diverse set of CDR3 sequences are generated and selected. c. Generated sequences have similar biochemical properties to naïve nanobody sequences. d. Cosine distances to nearest neighbors within the designed library and from the designed library to the naïve repertoire are greater than distances within the naïve repertoire (left). Average cosine distances within the designed library, from the designed library to the naïve repertoire, and within the naïve repertoire fall in a similar range (right).

However, in-frame insertions and deletions can also have large phenotypic consequences for protein function, yet these changes have proved difficult to model. We compare the length-normalized probabilities (bits-per-residue) calculated by the autoregressive model to experimental assays for the fitness of mutated biomolecules, using rank correlation (ρ) and area under the receiver-operator curve (AUC), identifying the two groups with a two-component Gaussian mixture model. The model is able to capture the effects of single amino acid deletions on PTEN phosphatase (ρ=0.67, AUC=0.83, N=340; Figure 2b; Mighell et al., 2018), multiple amino acid insertions and deletions in imidazoleglycerol-phosphate (IGP) dehydratase (ρ=0.68, AUC=0.92, N=6102; Figure 2c; Pokusaeva et al., 2018), and insertions and deletions in yeast snoRNA (ρ=0.51, AUC=0.76, N=14736; Puchta et al., 2016).

4.3. Generating an efficient library of functional nanobodies

Natural nanobody sequences are selected against unfavorable biochemical properties such as instability, polyspecificity, and aggregation during affinity maturation (Muyldermans, 2013). We sought to learn the constraints that characterize functional nanobodies by fitting the autoregressive model to a set of 1.2 million nanobody sequences from the llama immune repertoire (McCoy et al., 2014). Using this model, we find that length-normalized log-probability calculations predict the thermostability of new llama nanobody sequences (ρ=0.53, N=17; Figure 3a; Kunz et al., 2017).

Conditioning on the germline framework-CDR1-CDR2 nanobody sequence, we then generate a library of 185,836 highly diverse, biochemically well-behaved, "fit" CDR3 sequences that have a similar distribution of properties as the naïve llama immune repertoire (Methods; Figure 3c). The generated sequences are diverse from the naïve repertoire and are more distant from one another than sequences in the naïve repertoire (Figure 3d). Globally, the generated sequences are similarly distant to the naïve llama immune repertoire as the naïve sequences themselves, indicating that we have not entered an uncharacterized region of sequence space. Initial experiments to characterize this library confirm the functional diversity of even this small set of sequences.

5. Discussion

Here we show how neural network-powered autoregressive models can predict the effects of mutations in an unsupervised manner from evolutionary sequences alone. These models are capable of predicting the effects of insertions and deletions that would otherwise be ignored by previous methods. We then use this model to generate a diverse synthetic nanobody library with improved biochemical properties. Our streamlined library enables rapid, efficient discovery of candidate antibodies, quickly providing a starting point for affinity maturation to enhance binding affinity.

We anticipate advances in machine learning will greatly enhance both sequence modeling and rational library design. The addition of latent variables will allow controlled generation of desirable sequences (Bowman et al., 2015; Costello & Martin, 2019). Though others have reported difficulty training these models (Killoran et al., 2017), new machine learning techniques may make these models practical in the future (Kim et al., 2018; Yang et al., 2017; van den Oord & Vinyals, 2017). We also anticipate better strategies to explore diverse yet fit regions of sequence space during generation, either by exploiting variance explained by latent variables (Greener et al., 2018) or diverse beam search strategies (Vijayakumar et al., 2016). Finally, incorporating protein structural information may further improve the power of autoregressive models (Ingraham et al., 2019).

Due to their flexibility, deep autoregressive models will also open the door to new opportunities in probabilistic biological sequence analysis. Unlike alignment-based techniques, since no homology between sequences is explicitly required, generative models with autoregressive likelihoods can be applied to disordered proteins, multiple protein families, promoters and enhancers, or even entire genomes. For example, classical tumor suppressor genes, such as p53, BRCA1, and VHL, include long disordered regions where these models may prove particularly useful. With the increased number of available sequences and growth in both computing power and new machine learning algorithms, autoregressive sequence models will be indispensable for biological sequence analysis into the future.

6. Acknowledgements

We would like to thank John Ingraham, members of the Marks and Sander labs, and Harvard Research Computing.

References

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G. M. Unified rational protein engineering with sequence-only deep representation learning. *bioRxiv*, 2019. doi: 10.1101/589333.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., and Bengio, S. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

Cho, K., Van Merrinboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014.

Costello, Z. and Martin, H. G. How to hallucinate functional proteins. *arXiv preprint arXiv:1903.00458*, 2019.

Fowler, D. M. and Fields, S. Deep mutational scanning: a new style of protein science. *Nature Methods*, 11(8): 801807, 2014. doi: 10.1038/nmeth.3027.

Graves, A., Mohamed, A., and Hinton, G. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6645–6649, May 2013.

Greener, J. G., Moffat, L., and Jones, D. T. Design of metalloproteins and novel protein folds using variational autoencoders. *Scientific Reports*, 8(1):16189, 2018. ISSN 2045-2322.

Gupta, A. and Rush, A. M. Dilated convolutions for modeling long-distance genomic dependencies. *arXiv preprint arXiv:1710.01278*, 2017.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. ISSN 0899-7667.

Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P., Springer, M., Sander, C., and Marks, D. S. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35(2):128, 2017. ISSN 1546-1696.

Ingraham, J., Garg, V. K., Barzilay, R., and Jaakkola, T. Generative models for graph-based protein design. *ICLR workshop on Deep Generative Models for Highly Structured Data*, 2019.

Kalchbrenner, N., Espeholt, L., Simonyan, K., van den Oord, A., Graves, A., and Kavukcuoglu, K. Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099*, 2016.

Killoran, N., Lee, L. J., Delong, A., Duvenaud, D., and Frey, B. J. Generating and designing DNA with deep generative models. *arXiv preprint arXiv:1712.06148*, 2017.

Kim, Y., Wiseman, S., Miller, A. C., Sontag, D., and Rush, A. M. Semi-amortized variational autoencoders. *arXiv preprint arXiv:1802.02550*, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kosuri, S. and Church, G. M. Large-scale de novo DNA synthesis: technologies and applications. *Nature Methods*, 11(5):499507, 2014. doi: 10.1038/nmeth.2918.

Kunz, P., Flock, T., Soler, N., Zaiss, M., Vincke, C., Sterckx, Y., Kastelic, D., Muyldermans, S., and Hoheisel, J. D. Exploiting sequence and stability information for directing nanobody stability engineering. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1861(9):2196–2205, 2017. ISSN 0304-4165.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521(7553):436, 2015. ISSN 1476-4687.

Mann, J. K., Barton, J. P., Ferguson, A. L., Omarjee, S., Walker, B. D., Chakraborty, A., and Ndung'u, T. The fitness landscape of HIV-1 Gag: advanced modeling approaches and validation of model predictions by in vitro testing. *PLoS Computational Biology*, 10(8):e1003776, 2014. ISSN 1553-7358.

McCoy, L. E., Rutten, L., Frampton, D., Anderson, I., Granger, L., Bashford-Rogers, R., Dekkers, G., Strokappe, N. M., Seaman, M. S., and Koh, W. Molecular evolution of broadly neutralizing llama antibodies to the CD4-binding site of HIV-1. *PLoS Pathogens*, 10(12): e1004552, 2014. ISSN 1553-7374.

McMahon, C., Baier, A. S., Pascolutti, R., Wegrecki, M., Zheng, S., Ong, J. X., Erlandson, S. C., Hilger, D., Rasmussen, S. G., Ring, A. M., Manglik, A., and Kruse, A. C. Yeast surface display platform for rapid discovery of conformationally selective nanobodies. *Nature Structural & Molecular Biology*, 25(1545-9985):289–296, 2018.

Mighell, T. L., Evans-Dutson, S., and O'Roak, B. J. A saturation mutagenesis approach to understanding PTEN lipid phosphatase activity and genotype-phenotype relationships. *The American Journal of Human Genetics*, 102(5):943–955, 2018.

Muyldermans, S. Nanobodies: natural single-domain antibodies. *Annual Review of Biochemistry*, 82:775–797, 2013. ISSN 0066-4154.

van den Oord, A. and Vinyals, O. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pp. 6306–6315, 2017.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Pascanu, R., Mikolov, T., and Bengio, Y. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pp. 1310–1318, 2013.

Pokusaeva, V., Usmanova, D., Putintseva, E., Espinar, L., Sarkisyan, K., Mishin, A., Bogatyreva, N., Ivankov, D., Filion, G., Carey, L., and Kondrashov, F. Experimental assay of a fitness landscape on a macroevolutionary scale. *bioRxiv*, pp. 222778, 2018.

Puchta, O., Cseke, B., Czaja, H., Tollervey, D., Sanguinetti, G., and Kudla, G. Network of epistatic interactions within a yeast snoRNA. *Science*, 352(6287):840–844, 2016.

Riesselman, A. J., Ingraham, J. B., and Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nature Methods*, 15(10):816–822, 2018. ISSN 1548-7091.

Rives, A., Goyal, S., Meier, 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. *bioRxiv*, 2019. doi: 10.1101/622803.

Salimans, T. and Kingma, D. P. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 901–909, 2016.

Säll, A., Walle, M., Wingren, C., Müller, S., Nyman, T., Vala, A., Ohlin, M., Borrebaeck, C. A. K., and Persson, H. Generation and analyses of human synthetic antibody libraries and their application for protein microarrays. *Protein Engineering Design and Selection*, 29(10): 427437, 2016. doi: 10.1093/protein/gzw042.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. ISSN 1532-4435.

Steinegger, M. and Söding, J. Clustering huge protein sequence sets in linear time. *Nature Communications*, 9 (1):2542, 2018. doi: 10.1038/s41467-018-04964-5.

Sutskever, I., Martens, J., and Hinton, G. E. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 1017–1024, 2011.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pp. 3104–3112, 2014.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D., and Batra, D. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., and Bengio, S. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

Yang, Z., Hu, Z., Salakhutdinov, R., and Berg-Kirkpatrick, T. Improved variational autoencoders for text modeling using dilated convolutions. *arXiv preprint arXiv:1702.08139*, 2017.

Zhang, T., Ramakrishnan, R., and Livny, M. BIRCH: an efficient data clustering method for very large databases. In *ACM SIGMOD Record*, volume 25, pp. 103–114. ACM, 1996. ISBN 0897917944.