RESIDUE-LEVEL TEXT CONDITIONING FOR PROTEIN LANGUAGE MODEL MUTATION EFFECT PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

To augment protein sequence models with language, we introduce Conditioning on Residue-level Annotations from TExt (CRATE), a fine-tuning method that fuses two models using feature-wise linear modulation. We fine-tune protein language models at a large scale, first constructing a dataset (CRATE-train) joining annotations from InterPro and UniProtKB with sequences from UniRef90 resulting in approximately 105 million sequences each with at least three annotations and nearly 100% sequence coverage on average. Applying CRATE to mutation effect prediction improves performance on the ProteinGym over prior benchmarks. Leveraging these improvements, we show CRATE can be used to select annotations with the largest positive impact on mutation effect prediction and estimate the deep mutational scan (DMS) scores tested over multiple different assay selection types.

1 INTRODUCTION

Recent studies (Zheng et al., 2023; Lin et al., 2022; Meier et al., 2021; Madani et al., 2023) of protein language models (PLMs) have been shown to exhibit complex behavior that demostrates performative modeling of protein sequence, structure, function, and evolution (Rives et al., 2021; Zhang et al., 2024; Chowdhury et al., 2022; Huo et al., 2024). These findings elicit interest in augmenting and controlling protein language model behavior using additional contextual information, such as functional labels, text, or structure (Su et al., 2024; Hayes et al., 2024; Liu et al., 2025; Dai et al., 2024; Zhou et al., 2024; Duan et al., 2024). The increasing interest in multimodal representations of proteins increase the importance of investigating the modalities that, in addition to sequence, may provide contextual information that enriches the foundation model representation and result in performance improvements on downstream tasks.

In this work, we propose a framework for Conditioning on Residual-level Annotations from Text (CRATE) and evaluate the influence on model performance of residue-level text annotations on the ProteinGym benchmark. In addition to increased performance, CRATE-trained models benefit from being able to automatically identify the most influential annotations to provide as context on a per-task basis. CRATE fuses a protein foundation network with a text conditioning network using feature-wise linear modulation (FiLM) (Perez et al., 2017). Given a sequence and a set of residue-level annotations, the text conditioning network processes the set into annotation tracks and produces a contextual representation that is used for the FiLM calculations at each layer. We fine-tune CRATE models on CRATE-train, an inner join of the InterPro annotation database (Blum et al., 2024) and UniRef90 (Suzek et al., 2014) sequence cluster.

We train a CRATE model and evaluate its performance under different conditions on the ProteinGym benchmark (Notin et al., 2023). To enable text conditioning during evaluation, we propagate wildtype InterPro scan annotations onto the mutants. We find that in all but one assay category, we are able to improve performance by introducing some subset of the available annotations. We also demonstrate the capability of CRATE to accommodate different tasks by showcasing the model's performance on distinct selection assays of mutants with a common genotype.

2 METHODS

We hypothesize that incorporating text information from residue-level annotations from site-specific protein family models, such as HMM signatures for homologous proteins in InterPro (Blum et al., 2024), will sharpen the per-residue logit distributions along annotated intervals. Furthermore, we believe that after pre-training, the sharpened logit distributions are useful for downstream applications such as mutation effect prediction, given that the signal provided by the annotation may contribute to a more nuanced likelihood estimate.

2.1 MODEL

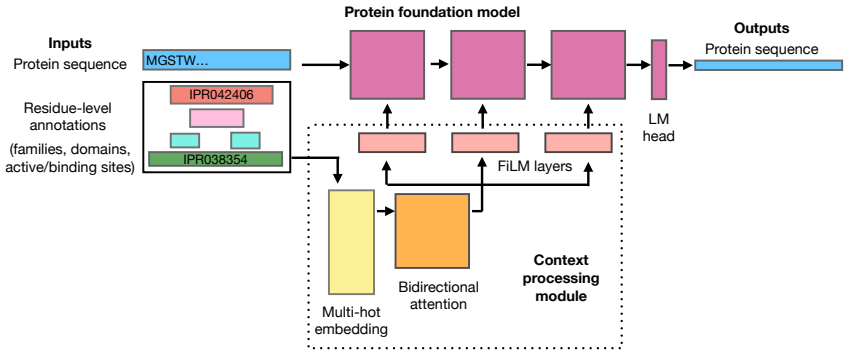

Figure 1: CRATE Model schematic. For a given sequence with residue-level annotations, we process the annotations by constructing a multi-hot embedding matrix and applying a single bidirectional Llama (Touvron et al., 2023) block trained from scratch. At each layer of the protein foundation model, we use FiLM (Perez et al., 2017) to inject the contextual information.

As seen in Fig. 1, CRATE fuses a text conditioning network (the context processing module) and a protein foundation model and is fine-tuned on a text-augmented dataset (see Section 2.2) using the foundation model's pre-training objective.

2.1.1 TEXT CONDITIONING NETWORK

Given a set of labels $\{y_0, \ldots, y_{N-1}\}$, we first construct a description embedding matrix $\boldsymbol{E} \in \mathbb{R}^{N \times m}$ so that the k-th row $\boldsymbol{E}_{k,:} \in \mathbb{R}^m$ corresponds to the description embedding for y_k. Let S be a protein sequence of length ℓ and $C_S = \{(i, j, k) \mid 0 \leq i \leq j < \ell,\ 0 \leq k < N\}$ be the set of residue level annotations of S, where some $(i, j, k) \in C_S$ is meant to represent that the closed index interval $[i, j]$ is annotated with label y_k. The text conditioning network consumes the annotation set C_S and constructs a coarse conditioning matrix $\tilde{\boldsymbol{C}}^S \in \mathbb{R}^{\ell \times m}$ by assigning each residue position i to the sum of the description embeddings that annotate it.

$$\tilde{\boldsymbol{C}}^S_{r,:} := \sum_{\{(i,j,k) \in C_S \mid\ i \leq r \leq j\}} \boldsymbol{E}_{k,:} \tag{1}$$

The coarse conditioning matrix then passes through a bidirectional attention module implemented as a single Llama (Touvron et al., 2023) block so that the final representation captures the mutual information between position-specific annotations. The final representation is a matrix $C \in \mathbb{R}^{\ell \times d}$, where d is the hidden dimension of the foundation model.

2.1.2 PROTEIN FOUNDATION MODEL

At each hidden layer of the foundation model, we insert a FiLM (Perez et al., 2017) module, which shifts and scales the input hidden representation by the conditioning input C (see Section 2.1.1). Importantly, the FiLM parameters are initialized to zero in order to avoid a high deviation from the input at the beginning of training.

2.1.3 GREEDY ANNOTATION SUBSET SELECTION

There is no guarantee that every label positively influences the correlation between model likelihood and DMS score. For example, the presence of an annotation may overly collapse the logit distribution in a specific position and therefore assign a mutant residue an artifically low log-probability. In order to find a subset of label types that positively impact the Spearman correlation, we employ a greedy algorithm for optimal subset selection. Starting from an empty set of labels, the greedy algorithm populates a "greedy subset" under the invariant that the label in question together with the running greedy subset improves the running best average Spearman correlation. This construction allows the user to focus the algorithm on a specific subset of assays, for example the group of assays with the same selection and mutation type, allowing for adaptation of CRATE to select different contextual information depending on the selection assay or mutation type.

2.2 DATA

We construct two datasets for pre-training (`CRATE-train`) and evaluation (`CRATE-vet`) by augmenting existing sequence databases with residue-level text annotations. For pre-training, we inner joined UniRef90 (Suzek et al., 2014) sequences with InterPro (Blum et al., 2024) annotations and UniProtKB (Bateman et al., 2024) binding/active site annotations, resulting in 43729 different InterPro label types, 552 different site label types, and 105,347,199 sequences with a significant amount of annotation coverage per sequence (Appendix A). For evaluation, we first obtain InterProScan annotations for each wildtype sequence tested in the ProteinGym (Notin et al., 2023) deep mutational scanning (DMS) assays. We then propagated the residue-level wildtype annotations onto their respective mutants by either direct transfer (in the case of substitutions) or by adjusting the annotation intervals appropriately by aligning the wildtype and mutant and inferring the sites of indels.

2.2.1 ANNOTATION TEXT EMBEDDINGS

We hypothesize that information described in natural-language text contextualized along the sequence is useful to sharpen the logit distribution along an annotated residue interval. In order to condition on text information at the residue-level, we embed the long-form descriptions of each InterPro/site label type present in `CRATE-train` using the state-of-the-art embedding model on the Massive Text Embedding Benchmark (MTEB) leaderboard (Muennighoff et al., 2023), NV-Embed-v2 (Lee et al., 2025). In order to save memory, we apply dimensionality reduction using principal component analysis (PCA) to reduce the original NV-Embed-v2 hidden dimension from 4096 to 1024. For the UniProt sites that did not already have long-form abstracts, we generated them by prompting Llama-3-1B (MetaAI Llama Team, 2024) to elaborate on the short-form name of the term (see Appendix D).

3 RESULTS

3.1 DEEP MUTATIONAL SCANNING BENCHMARK

Table 1: Performance comparison for different CRATE settings and the baseline averaged per (selection type, mutation type) pair.

	Expression		**Organismal Fitness**		**Stability**		**Activity**		**Binding**
Model	Indels	Subs.	Indels	Subs.	Indels	Subs.	Indels	Subs.	Subs.
Baseline	0.3594	0.3923	0.4212	0.3449	0.4885	0.3655	0.5109	0.3619	**0.2822**
$\text{CRATE}_{\emptyset}$	**0.3645**	0.3937	0.4244	0.3573	0.4936	0.3646	0.5245	0.3664	0.2550
$\text{CRATE}_{\text{all}}$	0.3571	**0.4062**	0.4476	**0.3724**	0.5072	0.3910	0.5158	**0.4062**	0.2680
$\text{CRATE}_{\text{greedy}}$	**0.3645**	0.3995	**0.4708**	0.3628	**0.5666**	**0.4264**	**0.5304**	0.3702	0.2770

We CRATE-trained ProGen2-small by fine-tuning the pre-trained weights using the procedure described in Section 2.1 and evaluated its performance on `CRATE-vet`. For fairness in evaluation, we also fine-tune a baseline ProGen2-small architecture with no modifications on the `CRATE-train` sequences. We compare three different inference modes of CRATE to the baseline: $\mathrm{CRATE}_{\emptyset}$ (no annotations are provided), $\mathrm{CRATE}_{\mathrm{all}}$, (all annotations are provided), and $\mathrm{CRATE}_{\mathrm{greedy}}$, (only the greedy optimal subset of annotations is provided). Table 1 shows that CRATE either outperforms or ties the baseline model on all but one assay/mutation type category. We also note that in certain cases, the greedy subset of annotations outperforms the all-in setting, $\mathrm{CRATE}_{\mathrm{all}}$, and sometimes vice versa. The cases where the greedy subset outperforms the all-in settings indicate that indeed a certain tuned subset of labels are perhaps independently correlated with the output signal. The latter cases, where the all-in setting exceeds the greedy one, may be explained by the existence of dependencies between the labels that were not tested during the course of the greedy algorithm.

Table 2: Performance comparison for different CRATE settings and the baseline averaged per group of (mutation-type, taxon division) pairs.

Taxon	Mutation type	$\mathrm{CRATE}_{\mathrm{all}}$	$\mathrm{CRATE}_{\emptyset}$	Baseline
Eukaryote	Indels	0.5057	0.5299	**0.5372**
	Subs.	**0.3981**	0.3758	0.3831
Human	Indels	**0.5438**	0.5015	0.4921
	Subs.	**0.4091**	0.3846	0.3791
Prokaryote	Indels	**0.4148**	0.3802	0.3647
	Subs.	**0.3507**	0.3375	0.3210
Virus	Indels	0.4698	**0.5068**	0.5007
	Subs.	**0.2956**	0.2808	0.2721

Table 2 similarly shows favorable performance of CRATE models to the baseline in all but one setting. These ablations provide evidence that conditioning on residue-level information improves downstream mutation effect prediction performance.

3.2 TASK-SPECIFIC CONDITIONING

VKOR1 is a transmembrane protein that drives the vitamin K cycle playing a role in blood clotting. VKOR1 contains 3-4 transmembrane domains and 4 conserved functional cysteine residues (Chiasson et al., 2020), depicted in Fig. 2. CRATE-trained models can be adapted at inference time to task-specific conditions; here, we compared the performance of two different CRATE variants to the baseline model on two different assay selection types, abundance and activity, for the same VKOR1 phenotype. Not only do the CRATE variants outperform the baseline on each assay, they also correctly exhibit different behaviors from distinct annotation inputs.

In order to demonstrate the differences in CRATE performance for the separate tasks, we illustrate the logit distributions in two different ways. For the abundance assay, we visualize the logit standard deviations calculated per-position, per-token across each mutant to focus on the variance across the sequence (see Fig. 2B and Fig. 2C). The greedy annotation set and propagated VKOR1 labels have one label in common "Vitamin K epoxide reductase complex subunit I" (IPR042406). As shown in Fig. 2, the annotation collapses token distributions in and surrounding the transmembrane domains, contributing to a slightly more robust mutation effect prediction. By contrast, the baseline model has much higher variance per token which may contribute to a more erroneous prediction. For the activity assay, we visualize the softmax probabilities per-token computed across each mutant (see Fig. 2E and Fig. 2F). Notably, CRATE correctly places most probability mass on 3/4 of the conserved cysteine residues (as opposed to the baseline), which echoes the findings in Chiasson et al. (2020) that only 3/4 of the conserved cysteines may be relevant to retain activity, supporting the notion that there is relevant information in the annotations that further collapses the softmax probabilities, resulting in more effective mutation effective prediction.

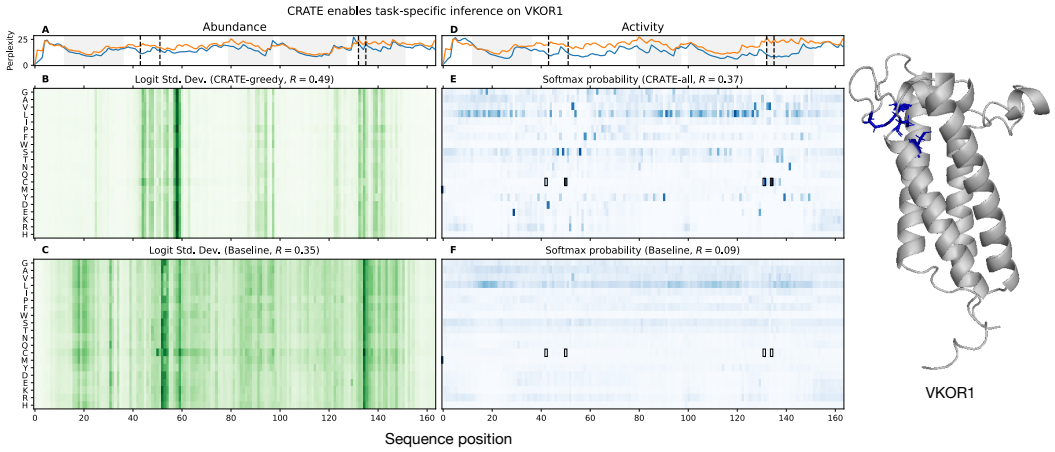

Figure 2: *Left*. VKOR1 case study on task specific inference. *Right*. AlphaFold 2 (Jumper et al., 2021) predicted structure of VKOR1. Highly conserved functional cysteine residues are shaded in blue. **(A, D)** Position specific perplexities for CRATE (blue) and the baseline (orange) averaged over abundance **(A)** and activity **(D)** phenotypes. Shaded regions represent trans-membrane domains and dashed lines represent positions of conserved, functional cysteine residues. **(B, C)** Logit standard deviations across abundance phenotypes for CRATE-greedy **(B)** and the baseline **(C)**. Positions with higher variance are shaded with more intensity. **(E, F)** Softmax probabilities calculated from activity phenotypes for CRATE-all **(E)** and the baseline **(F)**. Positions with higher probability mass are shaded with more intensity. Black boxes represent the softmax probability of conserved cysteines.

4 Conclusions & Future work

We proposed conditioning on residue-level annotations from text (CRATE), that fuses two networks to incorporate positional annotation information during training. We create two new annotation augmented datasets for training and inference, `CRATE-train` and `CRATE-vet`. We introduce the CRATE approach and demonstrate its improvement over the baseline as both a mutation effect predictor and task-specific method that can be modified to leverage different contextual information at test time. In the latter case, we greedily optimized label sets on a per-mutation-type, per-selection-type basis and showed that whereas all available annotations seem to contribute to improved predictions of VKOR1 activity phenotypes, the greedy annotation set improves abundance prediction for the same phenotype. We hypothesize that the annotations contribute to a more focused concentration of probability mass on a smaller subset of tokens over the baseline. Furthermore, in cases where $\mathrm{CRATE}_{\mathrm{greedy}}$ performs better, each label may independently contribute to the concentration of mass. By the same intuitions, higher-order dependencies between annotations may better explain instances where $\mathrm{CRATE}_{\mathrm{all}}$ performs better.

4.1 Future work

A promising next step would be investigating not only residue-level and sequence-level annotation conditioning in the context of protein engineering, but also studying applications of CRATE to other foundation models such as masked language models. In particular, we are interested in incorporating more annotations sources than InterPro (e.g., taxonomic and structure derived) as well as developing more principled frameworks for testing the effect of different annotation encodings. Finally, we are interested in testing CRATE models on other downstream tasks, such as protein function prediction and design.

References

Alex Bateman, Maria-Jesus Martin, Sandra Orchard, Michele Magrane, Aduragbemi Adesina, Shadab Ahmad, Emily H Bowler-Barnett, Hema Bye-A-Jee, David Carpentier, Paul Denny, Jun Fan, Penelope Garmiri, Leonardo Jose da Costa Gonzales, Abdulrahman Hussein, Alexandr

Ignatchenko, Giuseppe Insana, Rizwan Ishtiaq, Vishal Joshi, Dushyanth Jyothi, Swaathi Kandasaamy, Antonia Lock, Aurelien Luciani, Jie Luo, Yvonne Lussi, Juan Sebastian Martinez Marin, Pedro Raposo, Daniel L Rice, Rafael Santos, Elena Speretta, James Stephenson, Prabhat Totoo, Nidhi Tyagi, Nadya Urakova, Preethi Vasudev, Kate Warner, Supun Wijerathne, Conny Wing-Heng Yu, Rossana Zaru, Alan J Bridge, Lucila Aimo, Ghislaine Argoud-Puy, Andrea H Auchincloss, Kristian B Axelsen, Parit Bansal, Delphine Baratin, Teresa M Batista Neto, Marie-Claude Blatter, Jerven T Bolleman, Emmanuel Boutet, Lionel Breuza, Blanca Cabrera Gil, Cristina Casals-Casas, Kamal Chikh Echioukh, Elisabeth Coudert, Beatrice Cuche, Edouard de Castro, Anne Estreicher, Maria L Famiglietti, Marc Feuermann, Elisabeth Gasteiger, Pascale Gaudet, Sebastien Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Arnaud Kerhornou, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Lucille Pourcel, Sylvain Poux, Monica Pozzato, Manuela Pruess, Nicole Redaschi, Catherine Rivoire, Christian J A Sigrist, Karin Sonesson, Shyamala Sundaram, Anastasia Sveshnikova, Cathy H Wu, Cecilia N Arighi, Chuming Chen, Yongxing Chen, Hongzhan Huang, Kati Laiho, Minna Lehvaslaiho, Peter McGarvey, Darren A Natale, Karen Ross, C R Vinayaka, Yuqi Wang, and Jian Zhang. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, November 2024. ISSN 1362-4962. doi: 10.1093/nar/gkae1010. URL http://dx.doi.org/10.1093/nar/gkae1010.

Matthias Blum, Antonina Andreeva, Laise Cavalcanti Florentino, Sara Rocio Chuguransky, Tiago Grego, Emma Hobbs, Beatriz Lazaro Pinto, Ailsa Orr, Typhaine Paysan-Lafosse, Irina Ponamareva, Gustavo A Salazar, Nicola Bordin, Peer Bork, Alan Bridge, Lucy Colwell, Julian Gough, Daniel H Haft, Ivica Letunic, Felipe Llinares-López, Aron Marchler-Bauer, Laetitia Meng-Papaxanthos, Huaiyu Mi, Darren A Natale, Christine A Orengo, Arun P Pandurangan, Damiano Piovesan, Catherine Rivoire, Christian J A Sigrist, Narmada Thanki, Françoise Thibaud-Nissen, Paul D Thomas, Silvio C E Tosatto, Cathy H Wu, and Alex Bateman. InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Research*, 53(D1): D444–D456, November 2024. ISSN 1362-4962. doi: 10.1093/nar/gkae1082. URL http://dx.doi.org/10.1093/nar/gkae1082.

Melissa A Chiasson, Nathan J Rollins, Jason J Stephany, Katherine A Sitko, Kenneth A Matreyek, Marta Verby, Song Sun, Frederick P Roth, Daniel DeSloover, Debora S Marks, Allan E Rettie, and Douglas M Fowler. Multiplexed measurement of variant abundance and activity reveals vkor topology, active site and human variant impact. *eLife*, 9, September 2020. ISSN 2050-084X. doi: 10.7554/elife.58026. URL http://dx.doi.org/10.7554/eLife.58026.

Ratul Chowdhury, Nazim Bouatta, Surojit Biswas, Christina Floristean, Anant Kharkar, Koushik Roy, Charlotte Rochereau, Gustaf Ahdritz, Joanna Zhang, George M. Church, Peter K. Sorger, and Mohammed AlQuraishi. Single-sequence protein structure prediction using a language model and deep learning. *Nature Biotechnology*, 40(11):1617–1623, October 2022. ISSN 1546-1696. doi: 10.1038/s41587-022-01432-w. URL http://dx.doi.org/10.1038/s41587-022-01432-w.

Fengyuan Dai, Yuliang Fan, Jin Su, Chentong Wang, Chenchen Han, Xibin Zhou, Jianming Liu, Hui Qian, Shunzhi Wang, Anping Zeng, Yajie Wang, and Fajie Yuan. Toward De Novo Protein Design from Natural Language. August 2024. doi: 10.1101/2024.08.01.606258. URL http://dx.doi.org/10.1101/2024.08.01.606258.

Haonan Duan, Marta Skreta, Leonardo Cotta, Ella Miray Rajaonson, Nikita Dhawan, Alán Aspuru-Guzik, and Chris J. Maddison. Boosting the Predictive Power of Protein Representations with a Corpus of Text Annotations. July 2024. doi: 10.1101/2024.07.22.604688. URL http://dx.doi.org/10.1101/2024.07.22.604688.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. July 2024. doi: 10.1101/2024.07.01.600583. URL http://dx.doi.org/10.1101/2024.07.01.600583.

Mingjia Huo, Han Guo, Xingyi Cheng, Digvijay Singh, Hamidreza Rahmani, Shen Li, Philipp Gerlof, Trey Ideker, Danielle A. Grotjahn, Elizabeth Villa, Le Song, and Pengtao Xie. Multi-Modal Large Language Model Enables Protein Function Prediction. August 2024. doi: 10.1101/2024.08.19.608729. URL http://dx.doi.org/10.1101/2024.08.19.608729.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, July 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL http://dx.doi.org/10.1038/s41586-021-03819-2.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models, 2025. URL https://arxiv.org/abs/2405.17428.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic level protein structure with a language model. July 2022. doi: 10.1101/2022.07.20.500902. URL http://dx.doi.org/10.1101/2022.07.20.500902.

Shengchao Liu, Yanjing Li, Zhuoxinran Li, Anthony Gitter, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Arvind Ramanathan, Chaowei Xiao, Jian Tang, Hongyu Guo, and Anima Anandkumar. A Text-guided Protein Design Framework, 2025. URL https://arxiv.org/abs/2302.04611.

Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, January 2023. ISSN 1546-1696. doi: 10.1038/s41587-022-01618-2. URL http://dx.doi.org/10.1038/s41587-022-01618-2.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 29287–29303. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f51338d736f95dd42427296047067694-Paper.pdf.

MetaAI Llama Team. The Llama 3 Herd of Models, 2024. URL https://arxiv.org/abs/2407.21783.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive Text Embedding Benchmark, 2023. URL https://arxiv.org/abs/2210.07316.

Erik Nijkamp, Jeffrey Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. ProGen2: Exploring the Boundaries of Protein Language Models, 2022. URL https://arxiv.org/abs/2206.13517.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora Marks. ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 64331–64379. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/cac723e5ff29f65e3fcbb0739ae91bee-Paper-Datasets_and_Benchmarks.pdf.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer, 2017. URL https://arxiv.org/abs/1709.07871.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15), April 2021. ISSN 1091-6490. doi: 10.1073/pnas.2016239118. URL http://dx.doi.org/10.1073/pnas.2016239118.

Jin Su, Xibin Zhou, Xuting Zhang, and Fajie Yuan. ProTrek: Navigating the Protein Universe through Tri-Modal Contrastive Learning. June 2024. doi: 10.1101/2024.05.30.596740. URL http://dx.doi.org/10.1101/2024.05.30.596740.

Baris E. Suzek, Yuqi Wang, Hongzhan Huang, Peter B. McGarvey, and Cathy H. Wu. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, November 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu739. URL http://dx.doi.org/10.1093/bioinformatics/btu739.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, 2023. URL https://arxiv.org/abs/2302.13971.

Zhidian Zhang, Hannah K. Wayment-Steele, Garyk Brixi, Haobo Wang, Dorothee Kern, and Sergey Ovchinnikov. Protein language models learn evolutionary statistics of interacting sequence motifs. *Proceedings of the National Academy of Sciences*, 121(45), October 2024. ISSN 1091-6490. doi: 10.1073/pnas.2406285121. URL http://dx.doi.org/10.1073/pnas.2406285121.

Zaixiang Zheng, Yifan Deng, Dongyu Xue, Yi Zhou, Fei YE, and Quanquan Gu. Structure-informed Language Models Are Protein Designers, 2023. URL https://arxiv.org/abs/2302.01649.

Hanjing Zhou, Mingze Yin, Wei Wu, Mingyang Li, Kun Fu, Jintai Chen, Jian Wu, and Zheng Wang. ProtCLIP: Function-Informed Protein Multi-Modal Learning, 2024. URL https://arxiv.org/abs/2412.20014.

A Annotation coverage

We investigated the coverage, or the distribution of percentages of positions annotated by at least one residue-level annotation, in `CRATE-train`. As shown in Fig. 3, on average the majority of residues in the bulk of sequences have at least one annotation. Importantly, the overlapping annotations may contain redundant information, in which case we expect the model to preference certain labels over others.

B More information on CRATE fine-tuning

We fine-tune our ProGen2-small variants according to the recommendations in Nijkamp et al. (2022). In particular, we reduce the learning rate to 4×10^{-5}, implement a linear warmup schedule for 3000 steps followed by cosine annealing, and use the same optimizer weight decay and momentum parameters.

During training, we regularize the model by dropping out all annotations of a given sample with probability $p_{\text{drop}} = 0.5$. For next token prediction models such as ProGen2-small, we also reverse the sequence (and associated annotations) with probability $p_{\text{rev}} = 0.5$.

C ProteinGym evaluation

The ProteinGym benchmark (Notin et al., 2023) uses the assay-level Spearman rank correlations between the model estimate of the mutant log-likelihood and its experimental DMS score. The final

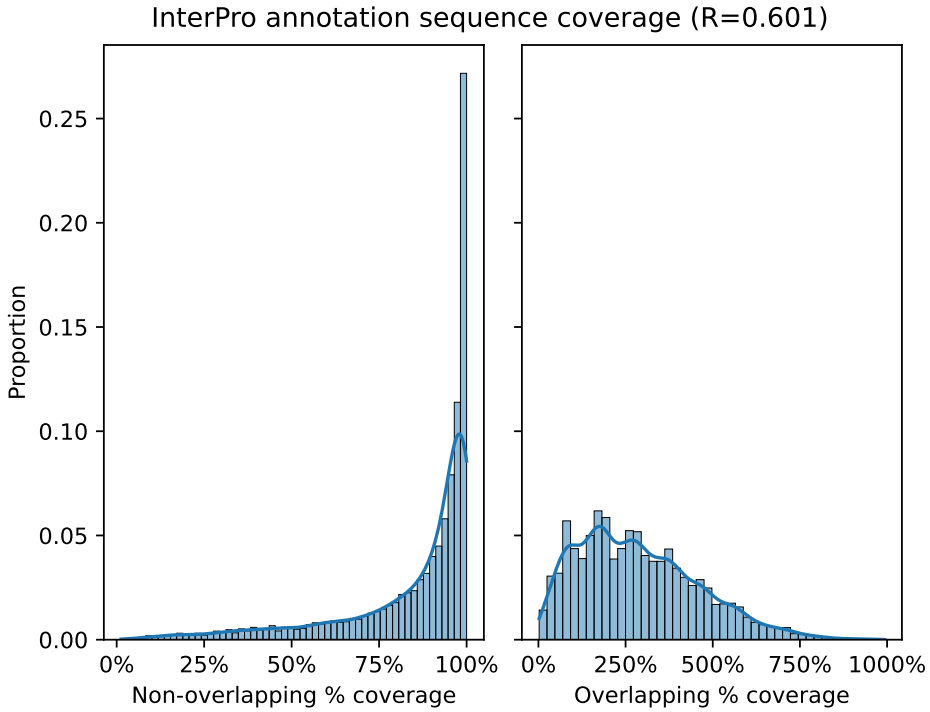

Figure 3: Annotation coverage distributions across a random subsample (10%) of `CRATE-train`. *Left*. The distribution of percent coverages per sequence (covered are counted once per sequence). *Right*. The distribution of percent coverage when we repeat the count of residues participating in multiple annotations.

evaluation scores are averages grouped over specific metadata properties of the assays, such as the mutation type (substitutions or indels), the DMS selection signal, or the taxon.

In all of our experiments using ProGen2-small, we follow Nijkamp et al. (2022) and compute the bidirectional sequence likelihood score by evaluating the model/annotations twice in forward and reverse configuration.

D MORE INFORMATION ON LONG FORM DESCRIPTION GENERATION

We condition CRATE model on natural-language embeddings of annotation descriptions. In order to generate description embeddings for the 552 different UniProt sites without natural language abstracts we first constructed for each site type a short form phrase of the form `site_type: site_name` if a site name was provided or otherwise just `site_type`. Using the short form phrase, we prompted Llama-3-1B to generate abstracts based on the following prompt.

> You are an expert molecular biologist with over 30 years of experience and an extensive knowledge of protein structure and function. You will be provided with a technical, short-form phrase. Generate a long-form description (no more than 500 words) that expands on the subject. Provide only the long-form description, without any explanation or comment otherwise.

We expand the short-form phrase into a long-form description with the intuition that the longer context creates a more informative description embedding.