Automated Neuron Labelling Enables Generative Steering and Interpretability in Protein Language Models

Arjun Banerjee[1], David Martinez[1], Camille Dang[1], Ethan Tam[1]
[1]Department of EECS, University of California, Berkeley

Abstract

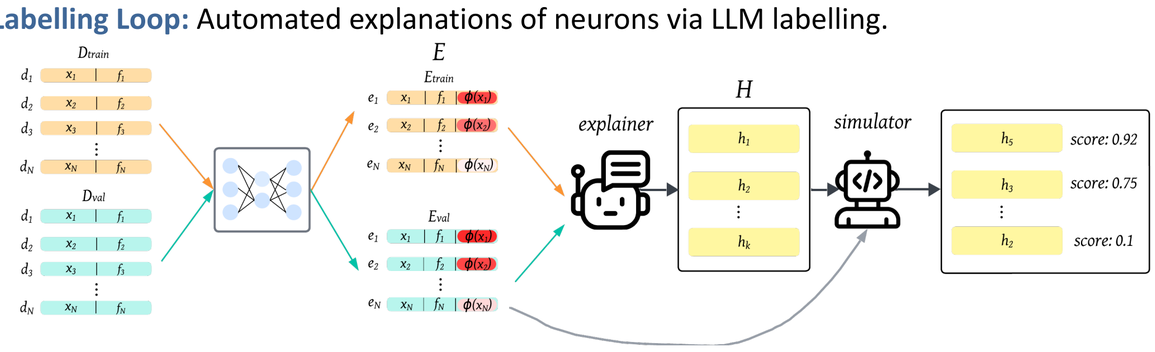
Protein language models (PLMs) encode rich biological information, yet their internal neuron representations are poorly understood. We introduce the first automated framework for labeling every neuron in a PLM with biologically grounded natural language descriptions. Unlike prior approaches relying on sparse autoencoders or manual annotation, our method scales to hundreds of thousands of neurons, revealing individual neurons are selectively sensitive to diverse biochemical and structural properties. We then develop a novel neuron activation-guided steering method to generate proteins with desired traits, enabling convergence to target biochemical properties like molecular weight and instability index as well as secondary and tertiary structural motifs, including alpha helices and canonical Zinc Fingers. We finally show that analysis of labeled neurons in different model sizes reveals PLM scaling laws and a structured neuron space distribution.

Introduction

- Existing interpretability efforts on PLMs use Sparse Autoencoders (SAEs) to identify features (binding sites, structural motifs, etc.) and steer models[1, 2,3]
- However, SAEs introduce optimization instability, architecture specific biases, sensitivity to initialization, and human-interpretation which all hinder effectiveness[4]
- Neuron-level labeling has emerged as a promising approach for interpreting model internals. While per-neuron approaches offer higher granularity, manually labeling individual neurons is prohibitively labor-intensive and does not scale to large models.
- Yet, recent works suggests that LLMs can explain neurons in language models[5,6] . This enables fine-grained feature identification and facilitates controlled steering
- Successes of neuron labelling in NLP settings raises key questions: Is neuron-level labelling for PLMs possible? Can it enable the same steering & interpretability success as in LLMs?

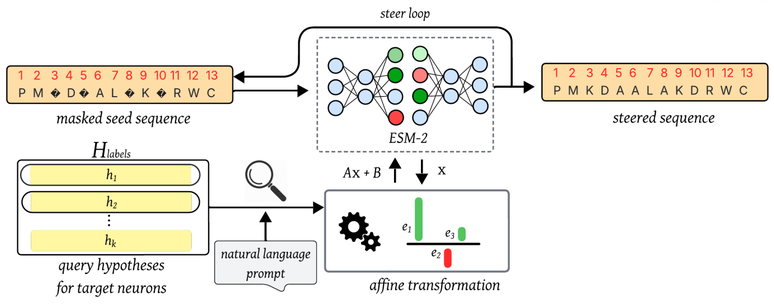
Methods: Labelling and Generation

Labelling Loop: Automated explanations of neurons via LLM labelling.



1. Constructed a dataset of 500,000 protein sequences[7] (x) annotated with features (f) and computed neuron activations (ϕ) through the forward pass of each ESM-2 model
2. Collect the m-top activating proteins and prompt and explainer LLM (E) to generate k possible hypothesis (h) explanations for each neuron
3. Score the hypotheses via a fine-tuned simulator LLM (S) that predicts an activation based on the features, sequence, and hypothesis. We take the hypothesis that maximizes the Pearson Correlation with the true activation

Generation Loop: Neuron-level intervention to steer towards natural language prompting



1. Given a natural language input, use an LLM to identify relevant neurons by their labels
2. Randomly mask a subset of amino acids in the (randomly initialized) input sequence
3. Pass the sequence through the ESM-2 model, replacing the activations X the with affine transformation $Ax + B$ for each relevant neuron. Sample new residues from the output logits using the model's softmax distribution, yielding a refined sequence
4. Repeat steps 2 and 3 for a user specified number of inputs

Learn More!

Paper: **Github:** **Labels:**

Steering Results

Characteristic Steering: can we target specific biochemical qualities (ex: "weight")?

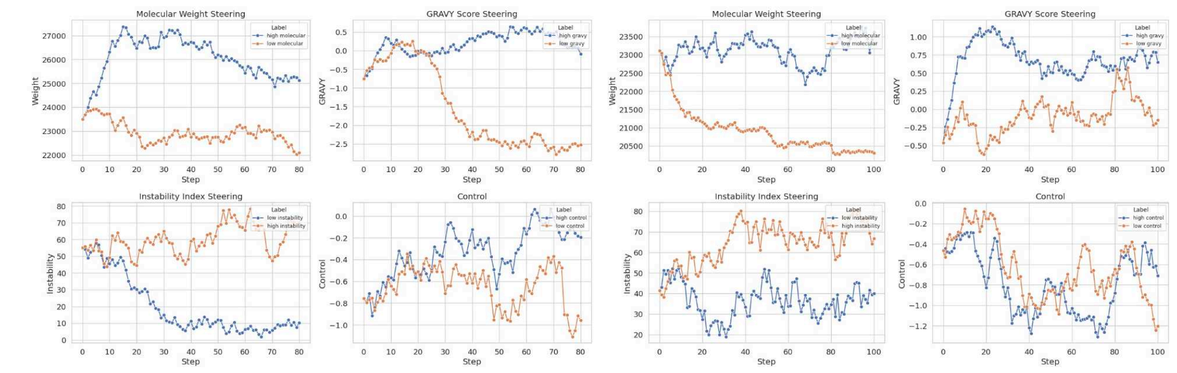


Chart 3. Successful steering to "high" and "low" values of various characteristics for ESM2-35M (Left) & ESM2-8M (Right)

Secondary Structure Steering: can we steer towards secondary structure (ex: "β-sheets")?

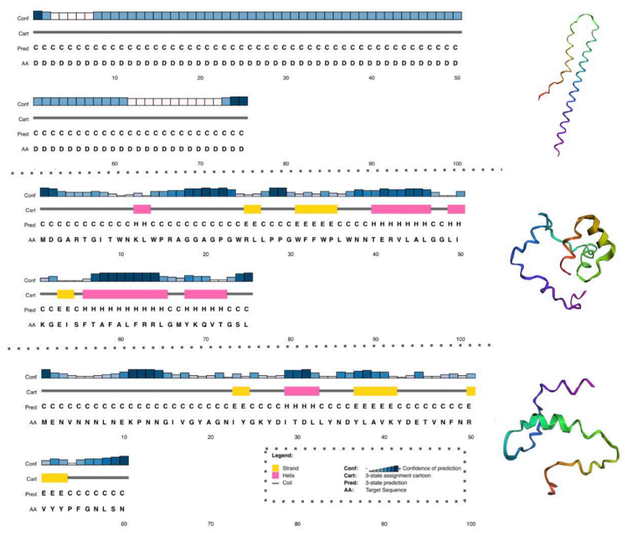


Chart 4. PsiPred structures for a base sequence (top), a sequence steered to α-helices (middle), and a sequence steered to β-sheets. Structures contain small amounts of the other characteristic as some neurons contain both features.

Tertiary Structure Steering: can we steer towards tertiary domains (ex: "Zinc Fingers")?

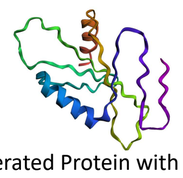


Chart 5. Generated Protein with a Zinc Finger — Cys-X(2–4)-Cys-X(12)-His-X(3–5)-His motif

Negative Steering: are linguistic opposites also neuron space opposites?

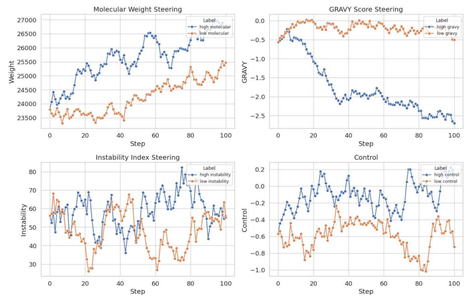


Chart 6. Negating the generation loop (ex: a ↦ -a) sometimes leads flipping characteristic effects

Interpretability Results

Generated Labels: what are the types of descriptions generated? Does a neuron correspond to one or many features?

Neuron	Description
(0, 160)	Strongly activates for secreted proteins with low to negative GRAVY scores.
(1, 323)	Strongly activates for flagellin proteins involved in bacterial flagellum structure.
(4, 204)	Strongly activates for proteins with high charge at pH 7 and a significant fraction of beta-sheet structure.
(7, 467)	Strongly activates for proteins with tryptophan synthase activity and negative gravy scores.
(9, 437)	Strongly activates for proteins with a specific role in DNA replication initiation and regulation.
(11, 473)	Strongly activates for chloroplastic proteins involved in RNA binding and processing.

Chart 7. Labels (layer, neuron number) are one sentence descriptors of structures, characteristics, and function. Neurons often contain multiple features.

Description Locations: what "types" of descriptions lie where? Does this change as model size grows?

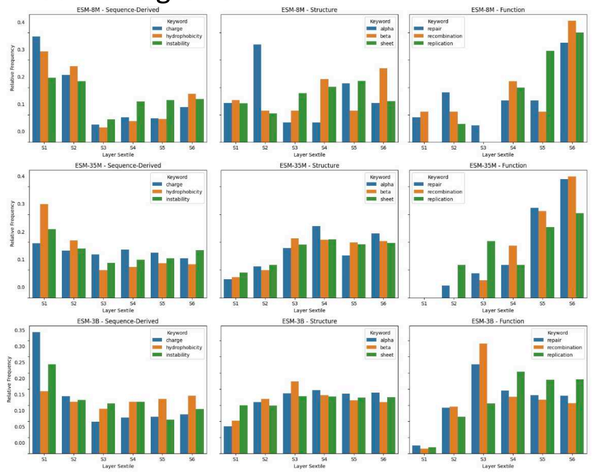


Chart 8. Distribution of neurons associated with 3 classes of descriptors across 3 different sizes of ESM-2. There exist general layer locations for each class of descriptor.

Scaling Laws: Does model size change label quality? Are more niche features present?

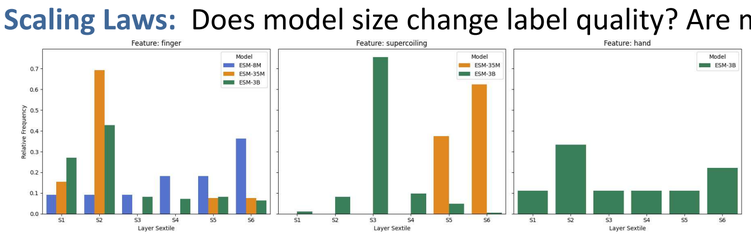


Chart 9. Neurons associated with niche structural motifs are only present in larger models, suggesting that larger models are able to better encode biophysical parameters.

Future Works

- Incorporate Structural Viability into the generation loop, potentially via an RL pipeline
- Explore the effects of LLM Oversimplification and mitigate potential effects of hallucination
- Label other PLMs and expand labelling to GLMs, exploring different model encodings
- Explore model pruning — try removing neurons that encode redundant information

We are all undergraduates applying to masters & PhD positions this Fall! Feel free to reach out!

Contacts

Arjun Banerjee: abaner@berkeley.edu
David Martinez: martinezdavid@berkeley.edu
Camille Dang: camillexdang@berkeley.edu
Ethan Tam: ethantam@berkeley.edu

References

[1] E. Adams, L. Bai, M. Lee, Y. Yu, M. AlQuraishi. "From Mechanistic Interpretability to Mechanistic Biology: Training, Evaluating, and Interpreting Sparse Autoencoders on Protein Language Models." bioRxiv, 2025. doi:10.1101/2025.02.06.636901
[2] E. N. V. Garcia, A. Ansuini. "Interpreting and Steering Protein Language Models Through Sparse Autoencoders." arXiv, 2025. https://arxiv.org/abs/2502.09135
[3] N. Parsan, D. J. Yang, and J. J. Yang. "Towards interpretable protein structure prediction with sparse autoencoders" arXiv preprint arXiv:2503.08764, 2025. URL https://arxiv.org/abs/2503.08764
[4] S. Kantamneni, J. Engels,, S. Rajamanoharan, M. Tegmark, N. Nanda. "Are sparse autoencoders useful?" A case study in sparse probing, 2025. https://arxiv.org/abs/2502.16681
[5] S. Bills, N. Cammarata, D. Mossing, H. Tillman, L. Gao, G. Goh, I. Sutskever, J. Leike, J. Wu, W. Saunders. "Language Models Can Explain Neurons in Language Models." OpenAI, 2023. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html
[6] D. Choi, Vi. Huang, K. Meng, D. D. Johnson, J. Steinhardt, S. Schwettman. "Scaling Automatic Neuron Description" 2024, https://transluce.org/neuron-descriptions
[7] The UniProt Consortium. "UniProt: The Universal Protein Knowledgebase in 2023." Nucleic Acids Res, 51(D1):D523–D531, 2023. doi:10.1093/nar/gkac1052