Chair of Bioinformatics
TUM School of Computation, Information and Technology
Technical University of Munich

Disentangling and Interpreting ProtT5 using Sparse Crosscoders

Guided Research Exposé

Sohrab Tawana

TUM School of Computation, Information and Technology, Technical University of Munich

sohrab.tawana@tum.de

December 11, 2025

Abstract — Protein Language Models (PLMs) like ProtT5 drive state-of-the-art performance in biology but remain opaque "black boxes". While their embeddings are useful, the rich biological knowledge they have likely acquired such as rules governing stability, binding, and function remains locked within their weights. This research aims to ask: Can Sparse Crosscoders effectively disentangle interpretable features from the internal representation of the ProtT5 encoder to provide mechanistic insight? We propose training a Sparse Crosscoder on the hidden states of all layers of the ProtT5 encoder. This architecture learns a shared dictionary of features, allowing us to interpret how biological concepts are represented and manipulated throughout the network. The expected result is a set of interpretable features that can be used to understand the model's internal logic, potentially shedding light into the features that the ProtT5 encoder extracts from the protein sequences it sees. As a use case, we will explore using the Crosscoder as a surrogate model within a ProteusAI-style evolutionary loop. Instead of a traditional surrogate classifier predicting fitness, our Crosscoder can directly measure the activation of desirable "feature circuits" (e.g., binding site integrity) to guide the selection of best-fitting protein sequences during evolution.

1 Introduction

Topic: Mechanistic interpretability of large Protein Language Models (PLMs).

Focus: The encoder of ProtT5 (fine-tuned for downstream tasks), specifically analyzing **all layers**.

Relevance/Problem: PLMs capture evolutionary patterns, but standard linear probes often fail to disentangle causal features from correlated confounds. Sparse Autoencoders (SAEs) have shown promise (InterPLM, SAEFold) in isolating distinct features.

Specific Gap: Training independent SAEs for every layer is inefficient and doesn't explicitly model the shared nature of features across depth.

Approach: We will apply **Sparse Crosscoders** [3]. By training a single shared dictionary across all layers, we can interpret features more robustly. We aim to use these features not just for observation, but to guide inputs—enabling interpretable steerability for biological design (directed evolution).

2 State of the Art of Research

2.1 Protein Language Models (PLMs)

ProtT5 (trained on BFD + UniRef50) is a standard for per-residue prediction and embedding generation [2].

2.2 SAEs in Biology

- *InterPLM* [5] & *InterProt* [1]: Applied standard SAEs to ESM-2, validating that latent directions correspond to active sites, domains, and functional properties.
- *SAEFold* [4]: Applied to structure prediction models.

2.3 Crosscoders (The Novelty)

Recent work by Anthropic ("Sparse Crosscoders") demonstrates superior efficiency and interpretability for cross-layer analysis compared to independent SAEs [3].

Novelty: To our knowledge, Sparse Crosscoders have **not yet been applied to any PLM**. This project would be the first to transfer this architecture to the protein domain.

3 Methods

3.1 Model

ProtT5-XL-U50 (Encoder only). We will extract and analyze hidden states from **all layers**.

3.2 Architecture: Sparse Crosscoder

- *Input:* Hidden states from all N layers defined as a unified input vector or batched appropriately.
- *Loss Function:* We will use the **L2-of-norms** loss. Unlike the L1 version used for direct baseline comparisons, utilizing L2 more efficiently optimizes the frontier of MSE and global sparsity, encouraging the model to find the most efficient shared features without explicit constraints to match per-layer SAE formulations.
- *State:* We will target the residual stream or full layer states, following the protocol of reference crosscoder literature.

3.3 Dataset

Target: ~10 million protein sequences primarily from **UniRef50**, potentially augmented with a significant portion of **BFD** (Big Fantastic Database).

Rationale: This mix aims to mimic the original pre-training data distribution of ProtT5 (which used BFD for pre-training and UniRef50 for fine-tuning), ensuring the features we discover are “native” to the model’s learned representation.

3.4 Analysis & Evaluation

- **Automated Interpretation:** Use LLMs to annotate features based on maximizing sequences.
- **Biological Validation:** Cross-reference active features with UniProt annotations.
- **In-Silico Directed Evolution (ProteusAI Integration):** We will explore using the Crosscoder as a **surrogate model** within a ProteusAI-style evolutionary loop. Instead of a traditional surrogate classifier predicting fitness, our Crosscoder can directly measure the activation of desirable "feature circuits" (e.g., binding site integrity) to guide the selection of best-fitting protein sequences during evolution.

4 Preliminary Work

4.1 Infrastructure

- Setup of the core `crosscoder` training codebase (based on Anthropic/open-source implementations).
- Configuration of ProtT5 inference pipelines for large-scale hidden state extraction.

5 Work Plan and Time Schedule

(3-Month Guided Research Project)

5.1 Month 1: Implementation & Data

- Prepare the training dataset (UniRef50 and possibly BFD).
- Extract hidden states from all layers of ProtT5.
- Implement the Sparse Crosscoder architecture with L2-of-norms loss.

5.2 Month 2: Training & Interpretation

- Train the Crosscoder (tuning expansion factor and sparsity penalty).
- Generate automated interpretations and map to biological databases.

5.3 Month 3: Application & Reporting (Optional Paths)

- *Option A (Directed Evolution):* Investigate using features to manually guide specific mutations (“restricted wet-lab in silico”).
- *Option B (ProteusAI):* Implement the Crosscoder as a surrogate model in ProteusAI to drive automated sequence optimization.
- Finalize the exposé/report.

References

[1] Etowah Adams, Liam Bai, Minji Lee, Yiyang Yu, and Mohammed AlQuraishi. From mechanistic interpretability to mechanistic biology: Training, evaluating, and interpreting sparse autoencoders on protein language models. *bioRxiv*, 2025.

[2] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, Oct 2022.

[3] Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse crosscoders for cross-layer features and model diffing.

[4] Nithin Parsan, David J. Yang, and John J. Yang. Towards interpretable protein structure prediction with sparse autoencoders, 2025.

[5] Elana Simon and James Zou. InterPLM: discovering interpretable features in protein language models via sparse autoencoders. *Nature Methods*, 22(10):2107–2117, 2025.