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 12, 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

Mechanistic interpretability of Protein Language Models (PLMs) offers a significant pathway to understanding the biological rules learned by these models. This project focuses on the encoder of ProtT5 [2], a state-of-the-art PLM. ProtT5 takes in protein sequences as input and generates embeddings that contain information about the protein's structure and function. These embeddings are then used as input for downstream tasks. Our goal is to disentangle the internal representations across all layers of the ProtT5 encoder to understand how biological concepts are represented and evolved throughout the network. While PLMs effectively capture evolutionary patterns, standard analysis methods like linear probes often fail to distinguish true causal features from correlated confounds. Recent advances using Sparse Autoencoders (SAEs), such as InterPLM [5], InterProt [1], and SAEFold [4], have shown promise in disentangling the internal representations of PLMs into distinct human-interpretable features. However, training independent SAEs for every layer of a deep network is computationally inefficient and fails to explicitly model the shared nature of features across network depth. Standard SAEs struggle to resolve cross-layer superposition or track features that persist throughout the residual stream. To address this, we propose applying sparse crosscoders [3], a novel architecture proposed by a team at Anthropic that learns a single shared dictionary across the internal representations of all layers of a deep network. This approach enables the identification of cross-layer features and facilitates circuit simplification by removing duplicate features and allowing meaningful connections to "jump" across trivial identity transformations. This approach allows for more robust feature interpretation and opens the door to using these features to guide inputs, enabling interpretable steerability for biological design tasks like directed evolution.

2 Related Works and Background

2.1 Protein Language Models (PLMs)

Protein Language Models (PLMs) like ProtT5, which was trained using unsupervised pre-training on a huge corpus of protein sequences the BFD (Big Fantastic Database) and UniRef50, have established themselves as standard tools for per-residue prediction, feature extraction, and embedding generation [2].

2.2 SAEs in Computational Biology

In the realm of interpretability, SAEs in Computational Biology have seen recent application. Works such as *InterPLM* [5] and *InterProt* [1] have applied standard SAEs to models like ESM-2, demonstrating that latent directions in the model space can correspond to specific active sites, protein domains, and functional properties. Similarly, *SAEFold* [4] has extended this approach to structure prediction models.

2.3 Crosscoders (The Novelty)

The specific novelty of this proposal lies in the application of sparse crosscoders. Originally introduced by Anthropic for large language models, sparse crosscoders offer superior efficiency and interpretability for cross-layer analysis compared to independent SAEs [3]. To our knowledge, this architecture has not yet been applied to any PLM, making this project the first to transfer this methodology to the protein domain.

3 Methods

3.1 Model

We will analyze the **ProtT5-XL-U50** encoder, extracting internal representations from all layers to understand how information is transformed depth-wise.

3.2 Architecture: Sparse Crosscoder

The core of our approach is the **sparse crosscoder** architecture. The model will take the internal activations after each of the 24 layers, excluding the final layer, of the ProtT5 encoder as a unified input. We will employ the **L2-of-norms** loss function. Unlike the L1-of-norms penalty used in the original sparse crosscoder architecture [3], which weights the L1 regularization penalty by the L1 norm of the per-layer decoder weight norms to enable apples-to-apples loss comparisons with independent SAEs and to surface layer-specific features, we will use the L2-of-norms loss. The L2-of-norms loss treats the decoder weights as a single vector and weights by its L2 norm. While this sacrifices direct loss comparability with per-layer SAEs, it more efficiently optimizes the frontier between reconstruction error (MSE) and global sparsity across all layers. Since our primary goal is to discover shared, interpretable biological concepts rather than to perform model diffing or precise loss benchmarking against baselines, this efficiently incentivizes the model to discover shared features that persist across the network depth.

3.3 Dataset

For the **Dataset**, we aim to compile approximately 10 million protein sequences, primarily sourced from **UniRef50**, potentially augmented with data from **BFD** (Big Fantastic Database). This composition attempts to mimic the original training distribution of ProtT5, ensuring that the features discovered are "native" to the model's learned internal representation.

3.4 Analysis & Evaluation

Our **Analysis & Evaluation** strategy involves three main components. First, we will generate **automated interpretations** by using Large Language Models (LLMs) to annotate features based on the sequences that maximally activate them. Second, we will perform **biological validation** by cross-referencing these active features with UniProt annotations to ground them in established biological knowledge. Finally, we will explore **In-Silico Directed Evolution**. We aim to integrate the sparse crosscoder as a **surrogate model** within a ProteusAI-style evolutionary loop. Rather than predicting fitness via a traditional classifier, the sparse crosscoder will measure the activation of specific, desirable "feature circuits" (e.g., binding site integrity), directly guiding the selection of sequences during the evolutionary process.

4 Preliminary Work

4.1 Infrastructure

We have already established the necessary **infrastructure** for this project. This includes setting up the core `crosscoder` training codebase, adapted from Anthropic's open-source implementations, and configuring the ProtT5 inference pipelines required for large-scale extraction of hidden states.

5 Work Plan and Time Schedule

(3-Month Guided Research Project)

5.1 Month 1: Implementation & Data

The first month will focus on implementation and data preparation. This includes compiling the training

dataset (UniRef50 and possibly BFD), extracting hidden states from all layers of ProtT5, and implementing the sparse crosscoder architecture with the specified L2-of-norms loss.

5.2 Month 2: Training & Interpretation

Month 2 will be dedicated to training and interpretation. We will train the sparse crosscoder, carefully tuning hyperparameters such as the expansion factor and sparsity penalty. Concurrently, we will generate automated interpretations of the learned features and map them to biological databases.

5.3 Month 3: Application & Reporting (Optional Paths)

The final month will explore applications and reporting. We will investigate the utility of the learned features for directed evolution, either by manually guiding mutations or by implementing the Crosscoder as a surrogate model in ProteusAI for automated sequence optimization. The project will conclude with the finalization of the exposé and 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.