A166 · Sparse autoencoders recover molecular mechanisms of disease in protein language models

Presenter: Karna Mendonca — Northeastern University

Authors: Karna Mendonca, Maria Clara de Paolis Kazula, Ross Stewart, Jannik Brinkmann, Predrag Radivojac

Session: Day 1 · Thu Oct 1, 2026 · 2:15–4:15 PM

Protein language models (pLMs) have become an integral part of the variant effect prediction toolkit, enabling state-of-the-art performance in predicting the effects of missense variants. However, most computational tools merely predict pathogenicity, without providing insight into the molecular mechanism of disease in pathogenic variants. To this end, we train sparse autoencoders (SAEs) on residue embeddings from ProtTransT5 and ESM2, identifying latent activations that are associated with specific types of functional residues such as phosphorylation and metal-binding sites, with significant enrichment. We also introduce variant-SAE, which is trained to reconstruct the differences between wildtype and mutant residue embeddings for missense population variants from gnomAD. Through linear probing on the difference of embeddings, we first demonstrate that residue-level embeddings from pLMs do capture loss of stability, achieving 0.931 AUROC when distinguishing highly destabilizing from neutral variants with experimentally-profiled effects. We also discover a small set of variant-SAE latent activations are significantly associated with destabilizing effects. To further validate the causal effects of latent activations on mechanisms of disease, we performed activation patching on variant-SAE latents by patching individual latent activations for neutral variants with the mean activation of all destabilizing variants. This resulted in the performance of the binary classifier to substantially drop by up to 20 percentage points. These results show that mechanistic effects of missense variants are sufficiently captured in pLM representations, and can be decomposed into interpretable features through unsupervised representation learning.

Sparse autoencoders recover molecular mechanisms of disease in protein language models

Karna Mendonca[1], Maria Clara de Paolis Kazula[1], Ross Stewart[1], Jannik Brinkmann[2], Predrag Radivojac[1]

Protein language models (pLMs) have become an integral part of the variant effect prediction toolkit, enabling state-of-the-art performance in predicting the effects of missense variants. However, most computational tools merely predict pathogenicity, without providing insight into the molecular mechanism of disease in pathogenic variants. To this end, we train sparse autoencoders (SAEs) on residue embeddings from ProtTransT5 and ESM2, identifying latent activations that are associated with specific types of functional residues such as phosphorylation and metal-binding sites, with significant enrichment. We also introduce variant-SAE, which is trained to reconstruct the differences between wildtype and mutant residue embeddings for missense population variants from gnomAD. Through linear probing on the difference of embeddings, we first demonstrate that residue-level embeddings from pLMs do capture loss of stability, achieving 0.931 AUROC when distinguishing highly destabilizing from neutral variants with experimentally-profiled effects. We also discover a small set of variant-SAE latent activations are significantly associated with destabilizing effects. To further validate the causal effects of latent activations on mechanisms of disease, we performed activation patching on variant-SAE latents by patching individual latent activations for neutral variants with the mean activation of all destabilizing variants. This resulted in the performance of the binary classifier to substantially drop by up to 20 percentage points. These results show that mechanistic effects of missense variants are sufficiently captured in pLM representations, and can be decomposed into interpretable features through unsupervised representation learning.

[1] Northeastern University
[2] University of Mannheim