A039 · Natural compensatory variation reveals how protein language models represent pathogenic epistasis and their ability to generate druggable targets via compensation

Presenter: Shivam Gandhi — Sunyaev Lab, Harvard DBMI

Authors: Shivam Gandhi, Carles Boix, Shamil Sunyaev

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

Protein language models (pLMs) achieve strong variant-effect prediction, but whether they learn context-dependent epistasis remains unclear. We introduce compensated pathogenic deviations (CPDs), human pathogenic variants that occur as wild-type alleles in orthologous proteins, as natural counterfactuals for testing whether pLM predictions change appropriately across sequence backgrounds. We construct a catalog of 425 high-confidence CPDs across more than 800 placental mammals.

Across models, contextual rescue depends strongly on information supplied at inference time. ESM2 and MSA-Pairformer predict rescue for 27.3% and 45.6% of CPDs, respectively, while adding structure to ESM3 significantly increases rescue (paired Wilcoxon $p=3.6\times10-7$). In MSA-Pairformer, removing CPD-carrying sequences from the input alignment leaves only 40.6% of full-MSA rescue magnitude, and matched random sequence removal yields significantly greater rescue than CPD-specific removal ($p=2.08\times10-15$), demonstrating contributions from both pretrained representations and inference-time evolutionary context.

Mechanistically, ESM2 preferentially attends from pathogenic sites to structurally contacting compensatory residues, with matched enrichment increasing to 10 percentile points in the final layer; 94.2% of variants show a positive effect. Yet context sensitivity is not equivalent to learned epistasis: full-background ESM rescue is largely explained by additive marginal effects of individual substitutions (Pearson $r=0.887$), and zero-shot ESM3 epistasis is essentially uncorrelated with experimental GB1 double-mutant epistasis (Spearman $\rho=0.005$).

These results establish CPDs as a natural benchmark for contextual protein prediction and reveal both where pLMs capture compensatory interactions and where current models fall short of genuine epistatic reasoning.

Natural compensatory variation reveals how protein language models represent pathogenic epistasis and their ability to generate druggable targets via compensation

Protein language models (pLMs) achieve strong variant-effect prediction, but how they represent the epistatic sequence context that determines variant effects remains unclear. We introduce compensated pathogenic deviations (CPDs), human pathogenic variants that occur as wild-type alleles in orthologous proteins, as natural counterfactuals for studying context-dependent prediction and compensatory epistasis. We construct the largest catalog to date, comprising 425 high-confidence ClinVar pathogenic variants observed across more than 800 placental mammals.

We compare zero-shot pathogenicity of the same amino-acid state in human and naturally compensated orthologous contexts across single-sequence, structure-conditioned, and MSA-aware pLMs. ESM2 and MSA-Pairformer predict rescue for 27.3% and 45.6% of CPDs, respectively. Providing structure to ESM3 significantly increases contextual rescue, with 238 of 391 variants improving (paired Wilcoxon $p=3.6\times10^{-7}$). In MSA-Pairformer, removing CPD-carrying sequences from the input alignment reduces rescue to 40.6% of its full-MSA magnitude, while removing an equal number of random sequences preserves significantly more rescue ($p=2.08\times10^{-15}$). Nevertheless, 83.8% of variants retain positive rescue following CPD-sequence removal, demonstrating contributions from both pretrained representations and evolutionary context supplied at inference time.

We next ask what forms of compensatory interactions pLMs preferentially capture. ESM2 rescue is strongly biased toward local structural interactions. After matching for sequence distance and amino-acid identity, attention from pathogenic residues to direct-contact compensatory sites is enriched by 10 percentile points in the final layer, with 94.2% of variants showing a positive effect, whereas distant non-contact compensations show no robust enrichment. Charge rebalancing emerges as a particularly interpretable mode of compensation. ESM-derived counterfactual representations perform best on variants whose compensators oppose the pathogenic charge change, comprising 45.4% of strong rescues among high-performing variants versus 27.7% among low-performing variants. More broadly, charge-rebalancing substitutions produce stronger predicted compensation than steric rebalancing. A classifier using only 30 of 1,280 counterfactual residual features recovers ESM2 compensation rankings with mean per-variant AUROC 0.864, suggesting that compensation-relevant information is highly compressible.

These successes coexist with important limitations. Full-background ESM rescue is largely predictable from additive marginal effects of individual substitutions (Pearson $r=0.887$), while zero-shot ESM3 epistasis is essentially uncorrelated with experimentally measured GB1 double-mutant epistasis (Spearman $\rho=0.005$). Thus, context sensitivity does not necessarily imply generalizable prediction of higher-order epistasis. Together, our results establish CPDs as a natural benchmark for contextual protein prediction, reveal structural and biochemical interactions represented by pLMs, and identify limitations of current zero-shot epistasis prediction. Beyond variant interpretation, learning these compensatory relationships provides a route toward prioritizing suppressor mutations and molecular interactions that could inform therapeutic strategies for rescuing pathogenic protein states.