Decoding Loss-of-Function Variants with Sparse Concept Features of ESM-2

Sungnam Kim [1] **Doyeon Ha** [1]

Abstract

Protein language models (PLMs) excel at variant-effect prediction, yet their dense residue embeddings lack interpretable axes, making the basis for predictions unclear. To address this, we trained sparse autoencoders (SAEs) on wild-type embeddings and aligned the resulting features with functional residue annotations, yielding interpretable concept features—SAE dimensions aligned with biological annotations (e.g., kinase domain, ATP binding site). Without any deep mutational scanning (DMS) supervision, we found that the more residue positions where a missense variant silences a concept feature below its wild-type activation, the greater the loss of function (LOF) observed in DMS assays. Furthermore, across five kinase activity assays, suppression of these concept features localized to subdomain IX—a region whose disruption is known to substantially impair kinase activity. These findings show SAEs decompose dense language model representations into interpretable signals that align with known mechanisms underlying variant effects.

1. Introduction

Protein language models (PLMs) (Rives et al., 2021; Lin et al., 2023) have become the standard residue-level substrate for variant-effect prediction (Frazer et al., 2021; Cheng et al., 2023; Notin et al., 2023). However, the dense hidden vectors driving these predictions reside in hundreds to thousands of dimensions without semantic axes. This opacity limits clinical applicability, as clinicians and biologists require a mechanistic understanding to evaluate the reliability of variant effect predictions. To address this, we decode per-residue ESM-2 embeddings via a sparse autoencoder (SAE) and align the resulting features post-hoc with UniProtKB/SwissProt (UniProt Consortium, 2023) residue annotations, yielding interpretable concept features. We use ESM-2 8M as the main backbone for compute efficiency in the per-layer SAE training, and verify that the kinase-domain trends reproduce at ESM-2 650M. We then investigate whether these concept features, trained exclusively on wild-type sequences, can identify deep-mutational-scanning (DMS) loss-of-function (LOF) variants by counting the residue positions at which a missense substitution silences a concept feature below its wild-type activation.

SAEs applied to PLM activations have been utilized for biological feature discovery (Adams et al., 2025; Garcia & Ansuini, 2025; Gujral et al., 2025; Simon & Zou, 2025; Liu et al., 2026). Whereas InterPLM focuses on characterizing wild-type features, our work extends this by demonstrating their zero-shot utility in decoding missense variant effects. The closest precedent is the recent work by Corominas et al. (2025), who fine-tuned a sparse autoencoder on α-amylase DMS data and used the resulting features as control surfaces for sequence generation in a single case study. However, while DMS assays are resource- and time-intensive, wild-type reference sequences are already curated at scale across species; we therefore ask whether a sparse autoencoder trained solely on such sequences can decode missense variant effects across a broad range of unseen DMS assays.

The central contribution of this work is a zero-shot, supervision-free read-out of functional silencing derived entirely from wild-type-trained concept features. Across diverse protein–concept pairs, the number of residue positions silenced by a missense substitution tracks DMS-measured loss of function in 9 of 10 protein–concept pairs. Furthermore, in five kinase activity assays, the mutation sites of variants producing such widespread silencing localize to subdomain IX—a region whose disruption is independently known to substantially impair catalytic activity (Serizawa et al., 2016). Two complementary checks support the specificity of this read-out: the signal is not restricted to sequence conservation, and a random-forest regressor trained on the interpretable concept basis matches the dense ESM-2 backbone in supervised DMS prediction. We further report a kinase-restricted up-shift signal in the same read-out. Figure 1 summarizes the framework.

[1]Department of Artificial Intelligence, Research and Development Center, 3billion, Inc., Seoul, Republic of Korea. Correspondence to: Doyeon Ha <biologysaves@gmail.com>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026)

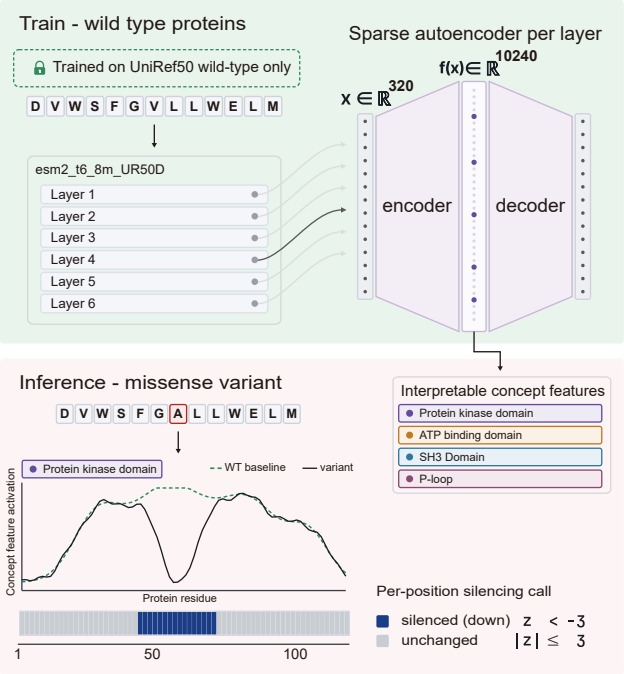

Figure 1. **End-to-end pipeline. Top (training, green):** Per-residue ESM-2 embeddings from each of six layers are passed through per-layer sparse autoencoders trained on wild-type sequences, and each retained feature is post-hoc aligned to a UniProtKB/SwissProt residue annotation, yielding interpretable concept features. **Bottom (inference, pink):** A missense variant is run through the same pipeline; for a chosen concept feature, each residue position is called *silenced* ($z < -3$), *up-shifted* ($z > +3$), or *unchanged* relative to the wild-type baseline.

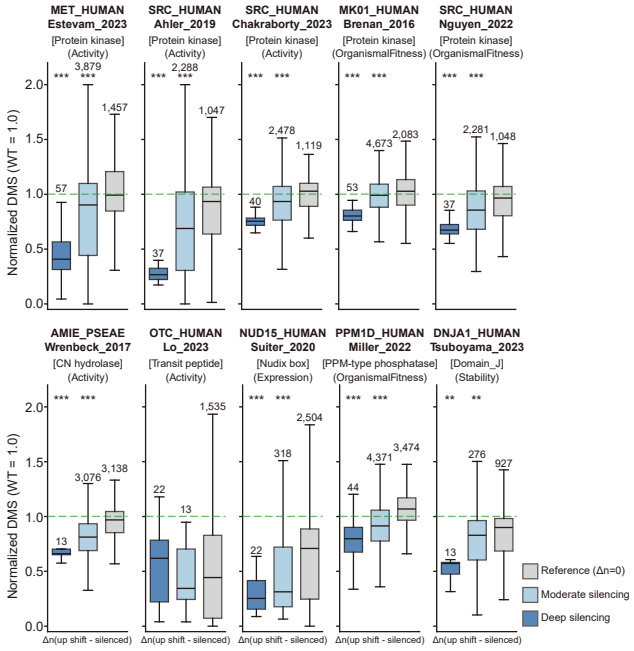

Figure 2. **Concept feature silencing tracks DMS loss of function across diverse proteins.** Normalized DMS scores (Appendix C) are plotted against the net concept shift $\Delta n = n^{\mathrm{up}} - n^{\mathrm{down}}$ for 10 protein-concept pairs: 5 kinase assays (top row) and 5 non-kinase assays (bottom row). Variants with net silencing ($\Delta n \leq 0$) are grouped into a zero-shift reference ($\Delta n = 0$) and two active silencing bins (moderate and deep; Section 3.1). Numbers above each box indicate per-bin variant counts. In 9 of 10 pairs, the silenced bins show a statistically significant decrease relative to the zero-shift reference (two-sided Mann–Whitney U test with Benjamini–Hochberg FDR correction applied within each panel; *, **, *** indicate $p < 0.01, 0.001, 10^{-5}$).

2. Methods

2.1. Sparse Autoencoder

Our SAE follows the design of InterPLM (Simon & Zou, 2025). We extract per-residue embeddings from every transformer layer of ESM-2 `esm2_t6_8M_UR50D` (Lin et al., 2023) and train a separate ReLU + ℓ_1 (Bricken et al., 2023) SAE for each layer on UniRef50 (Suzek et al., 2015) wild-type proteins. Each per-residue input $\mathbf{x} \in \mathbb{R}^{320}$ is mapped to a sparse code and a reconstruction with expansion factor $r = 32$ and dictionary size $D_h = 10{,}240$ per layer (full training protocol in Appendix A).

$$\begin{aligned} \mathbf{f}(\mathbf{x}) &= \mathrm{ReLU}\big(\mathbf{W}_e(\mathbf{x} - \mathbf{b}_d) + \mathbf{b}_e\big) \\ \hat{\mathbf{x}} &= \mathbf{W}_d \mathbf{f}(\mathbf{x}) + \mathbf{b}_d, \end{aligned} \tag{1}$$

2.2. Concept-Aligned Read-Out

To obtain interpretable features, we align SAE features post-hoc to UniProtKB (UniProt Consortium, 2023) residue annotations, retaining 168 features with $F_1 \geq 0.5$. For LOF analysis, each concept is represented by its F_1-best feature. For each feature j and position t, we quantify the impact of each variant i, regardless of its mutation site, by computing a z-score relative to the wild-type activation $A^{\mathrm{WT}}_{j,t}$:

$$z_{i,j,t} = \frac{A^{\mathrm{var}}_{i,j,t} - A^{\mathrm{WT}}_{j,t}}{\sigma_{j,t}}, \quad \sigma_{j,t} = \mathrm{SD}_i\big(A^{\mathrm{var}}_{i,j,t} - A^{\mathrm{WT}}_{j,t}\big) \tag{2}$$

where $\sigma_{j,t}$ is the standard deviation calculated across the entire variant population. To suppress noise, we define a feature as *silenced* if $z_{i,j,t} < -3$ and *up-shifted* if $z_{i,j,t} > +3$. These per-position events are then aggregated into variant-level counts, n_i^{down} and n_i^{up}, to facilitate downstream tasks.

2.3. Kinase Case Study

The kinase case study uses the F_1-best SAE feature aligned to the Protein kinase concept (layer-3, $F_1 = 0.834$). The analysis covers the five ProteinGym (Notin et al., 2023) activity-cohort assays—MET, MK01, and three SRC sets—that carry the UniProtKB/SwissProt "Protein kinase" annotation. We build an SRC-anchored MSA over 3,683 SwissProt kinase-domain entries; per-position conservation is computed as $1 - H(\mathrm{col})/\log_2 20$ with Henikoff position-based sequence weights (Henikoff & Henikoff, 1994) (full procedure in Appendix B).

Figure 3. **Deep silencing converges on kinase subdomain IX.** SRC-anchored MSA of the five kinase activity assays (MET, MK01, three SRC sets) with NTRK1 and FLT3 as literature-validated references. Dark-blue cells mark positions containing variants in the deep-silencing bin (Section 3.1); light-blue cells mark residues within subdomain IX of NTRK1 and FLT3 with substitutions shown to impair kinase activity (Serizawa et al., 2016).

3. Results and Discussion

3.1. Concept Feature Silencing Tracks Loss of Function

For each (protein, concept) pair, we define the net concept shift as $\Delta n = n^{\text{up}} - n^{\text{down}}$, where n^{up} and n^{down} denote the number of up-shifted and silenced residue positions, respectively (Section 2.2). Focusing on the silencing regime ($\Delta n \leq 0$), we plot the normalized DMS scores (Appendix C) against Δn and partition variants into three bins based on the maximum observed net-negative shift $N_{\max}$: a zero-shift reference ($\Delta n = 0$), a moderate-silencing bin (from -1 to $-N_{\max}/2$), and a deep-silencing bin (from $-(N_{\max}/2 + 1)$ to $-N_{\max}$).

Figure 2 presents results for ten protein-concept pairs that met the minimum threshold of 1,000 single-substitution variants in the DMS dataset (ENVZ_ECOLI_Ghose_2023 was excluded as no variants produced silencing). These comprise five kinase pairs (top row) and five non-kinase pairs spanning diverse selection types such as Expression and Stability (bottom row). In 9 of the 10 pairs, the conditional medians of the normalized DMS scores decrease monotonically as the silencing magnitude grows, and the deep-silencing bins of these 9 pairs show a statistically significant functional loss relative to the zero-shift reference (two-sided Mann–Whitney U test, Benjamini–Hochberg FDR correction (Benjamini & Hochberg, 1995)). The same trend holds in the five kinase pairs at an alternative model scale (Appendix D).

These results demonstrate that an interpretable concept feature, trained solely on wild-type sequences, can track DMS loss-of-function scores at inference time (Fowler & Fields, 2014) across diverse protein families and selection types: when a missense variant disrupts the wild-type sequence regularities captured by a concept feature at many residues, the variant tends to manifest as LOF in the corresponding DMS assay.

3.2. Silencing Signals Localize to Kinase Subdomain IX

Across the five kinase activity assays (MET, MK01, and three SRC sets; Section 2.3), we mapped variants from the deep-silencing bin (Section 3.1) onto a multiple sequence alignment (MSA) anchored on the SRC (Figure 3).

Across all five assays, the deep-silencing variants clustered in a single contiguous MSA region: subdomain IX of the kinase catalytic core (Hanks & Hunter, 1995; Taylor & Kornev, 2011). In a previous study, this subdomain was shown to stabilize the conformation of the catalytic loop, and specific substitutions within it—such as NTRK1 V710A and FLT3 K868N—were shown to substantially impair kinase activity (Serizawa et al., 2016). We visualized this signal by projecting the union of these residues onto the AlphaFold2-predicted SRC structure (Appendix E). The same signal on subdomain IX is reproduced at an alternative model scale (Appendix D).

The co-localization of the deep-silencing positions with this independently characterized functional anchor indicates that the kinase-domain concept feature aligns with a known mechanism of catalytic loss of function, despite the underlying SAE being trained without any DMS or functional supervision.

3.3. Concept-Feature Silencing Carries Information Beyond Sequence Conservation

A natural concern is whether the silencing signal recapitulates sequence conservation, since the underlying PLM implicitly captures conservation patterns from its pre-training corpus. To address this, we compared the deep-silencing positions identified in Section 3.2 against per-column sequence conservation (Capra & Singh, 2007). For each of the five kinase activity assays, we plot per-column Shannon conservation, computed over 3,683 SwissProt kinase-domain entries (Section 2.3), against the per-position median normalized DMS score (Figure 4). Positions containing deep-silencing-

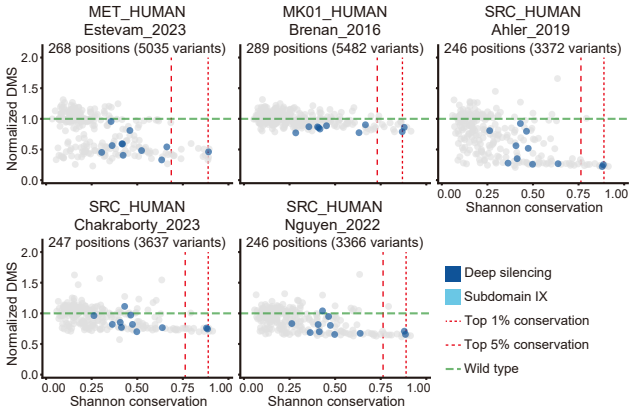

Figure 4. **Concept-feature silencing carries information beyond sequence conservation.** For each of the five kinase activity assays, per-position median normalized DMS scores (y-axis) are plotted against per-column Shannon conservation over 3,683 SwissProt protein kinase domain entries (x-axis; Section 2.3). Blue points mark positions containing deep-silencing-bin variants (criterion as in Figure 3); gray points mark all other positions. Vertical red dashed lines mark the top 5% and top 1% conservation percentile cutoffs commonly used in conservation-based functional residue prediction (Panchenko et al., 2004). Blue points span a wide range of conservation values, indicating that the concept feature flags positions that per-column conservation alone would not.

bin variants are colored blue; all other positions are gray. Vertical lines mark the top 5% and top 1% conservation percentile cutoffs commonly used in conservation-based functional residue prediction (Panchenko et al., 2004).

The deep-silenced (blue) positions span a broad range of conservation values, including many positions that fall outside the top-5% percentile cutoff. Applying a standard top-percentile cutoff for conservation-based functional residue prediction would therefore miss these positions, indicating that the kinase-domain concept feature captures functional constraints that conservation cutoffs alone do not identify. The two readouts thus provide complementary evidence for variant effects. In particular, the convergence on subdomain IX shown in Section 3.2 reflects critical functional residues that wild-type-trained concept features can flag but per-column conservation alone does not.

3.4. Concept Features Preserve Variant-Effect Prediction Performance

To evaluate whether these concept features can reliably power interpretable predictive models without sacrificing accuracy, we trained random-forest regressors across $N = 201$ ProteinGym substitution assays (Appendix F). Despite reducing the per-position feature pool by an order of magnitude (from $L \times d_{\text{model}} = 1{,}920$ to 168 features), the SAE representation achieved predictive performance (median Spearman $\rho = 0.719$, Pearson $r = 0.756$) matching the dense ESM-2 backbone ($\rho = 0.715$, $r = 0.753$; Appendix F). The paired difference $\Delta\rho = -0.003 \pm 0.026$ (mean $\pm$ SD over $N = 201$ assays) is statistically equivalent within an a priori margin of ± 0.01 (TOST paired-t, $p < 10^{-3}$). This demonstrates that concept features successfully capture the variant-effect signal, making them a reliable basis for building transparent predictive models.

3.5. Concept Feature Up-Shifts Track Functional Gain in Kinases

The preceding silencing analyses focused on $\Delta n \leq 0$. Because the read-out also records the symmetric up-side event ($z > +3$; Section 2.2), we extended the box-plot analysis of Figure 2 to the full range of $\Delta n = n^{\text{up}} - n^{\text{down}}$ for the five kinase activity assays (Section 2.3). Variants are partitioned into five symmetric bins—deep silencing, moderate silencing, zero-shift reference, moderate up-shift, and deep up-shift—using thresholds analogous to those defined in Section 3.1 (Appendix G).

In all five assays, the conditional medians of the normalized DMS scores increase monotonically as Δn becomes more positive, mirroring the decrease observed in the silencing regime. The deep up-shift bins show a statistically significant elevation in functional score relative to the zero-shift reference (two-sided Mann–Whitney U test, Benjamini–Hochberg FDR-corrected).

Driving a wild-type-aligned concept feature above its baseline activation ($z > +3$) tracks variants with above-wild-type DMS scores in the five kinase activity assays, consistent with gain-of-function substitutions.

4. Conclusion

We have shown that the dense per-residue embeddings of PLMs can be decomposed, via a sparse autoencoder trained solely on wild-type sequences, into an interpretable basis of concept features aligned with curated annotations. Without DMS supervision, the number of residue positions at which a missense substitution silences such a concept feature tracks DMS-measured loss of function in 9 of 10 protein–concept pairs. In a kinase case study, these silencing signals clustered in subdomain IX of the catalytic core, a region whose disruption is independently known to impair catalytic activity (Serizawa et al., 2016), and this signal is not recoverable from conservation alone. The interpretable concept basis matches the dense ESM-2 backbone in supervised DMS prediction, showing that interpretability need not cost predictive utility. Together with a kinase-restricted up-shift signal, these results support concept features as a mechanistically transparent lens on missense variant effects. Despite remaining limitations in model scale, annotation coverage, and SAE design (Appendix H), this work offers a concrete starting point for interpretable variant-effect tools.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Adams, E., Bai, L., Lee, M., Yu, Y., and AlQuraishi, M. From mechanistic interpretability to mechanistic biology: Training, evaluating, and interpreting sparse autoencoders on protein language models. *bioRxiv*, 2025.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Breiman, L. Random forests. *Machine learning*, 45(1): 5–32, 2001.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Capra, J. A. and Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23 (15):1875–1882, 2007.

Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L. H., Zielinski, M., Sargeant, T., et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381(6664):eadg7492, 2023.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422, 2009.

Corominas, G. B., Stocco, F., and Ferruz, N. Sparse autoencoders in protein engineering campaigns: Steering and model diffing. In *ICML 2025 Generative AI and Biology (GenBio) Workshop*, 2025.

Fowler, D. M. and Fields, S. Deep mutational scanning: a new style of protein science. *Nature methods*, 11(8): 801–807, 2014.

Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J. K., Brock, K., Gal, Y., and Marks, D. S. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021.

Garcia, E. N. V. and Ansuini, A. Interpreting and steering protein language models through sparse autoencoders. *arXiv preprint arXiv:2502.09135*, 2025.

Gujral, O., Bafna, M., Alm, E., and Berger, B. Sparse autoencoders uncover biologically interpretable features in protein language model representations. *Proceedings of the National Academy of Sciences*, 122(34):e2506316122, 2025.

Hanks, S. K. and Hunter, T. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification 1. *The FASEB journal*, 9(8):576–596, 1995.

Henikoff, S. and Henikoff, J. G. Position-based sequence weights. *Journal of molecular biology*, 243(4):574–578, 1994.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

Liu, X., Lei, H., Liu, Y., Liu, Y., and Hu, W. ProtSAE: Disentangling and interpreting protein language models via semantically-guided sparse autoencoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 773–781, 2026.

Notin, P., Kollasch, A., Ritter, D., Van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Orenbuch, R., Weitzman, R., et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in neural information processing systems*, 36:64331–64379, 2023.

Panchenko, A. R., Kondrashov, F., and Bryant, S. Prediction of functional sites by analysis of sequence and structure conservation. *Protein science*, 13(4):884–892, 2004.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the national academy of sciences*, 118(15):e2016239118, 2021.

Serizawa, M., Kusuhara, M., Ohnami, S., Nagashima, T., Shimoda, Y., Ohshima, K., Mochizuki, T., Urakami, K., and Yamaguchi, K. Novel tumor-specific mutations in receptor tyrosine kinase subdomain ix significantly reduce extracellular signal-regulated kinase activity. *Anticancer Research*, 36(6):2733–2744, 2016.

Simon, E. and Zou, J. InterPLM: discovering interpretable features in protein language models via sparse autoencoders. *Nature methods*, 22(10):2107–2117, 2025.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.

Taylor, S. S. and Kornev, A. P. Protein kinases: evolution of dynamic regulatory proteins. *Trends in biochemical sciences*, 36(2):65–77, 2011.

UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic acids research*, 51(D1):D523–D531, 2023.

A. Sparse Autoencoder Training and Pipeline Detail

Training hyperparameters. Table 1 lists the per-layer sparse-autoencoder training configuration; the same configuration applies to every layer $\ell \in \{1, \ldots, 6\}$. A linear warm-up on both the learning rate and the sparsity coefficient prevents early dead-feature collapse without dead-neuron resampling (Bricken et al., 2023). The non-default scikit-learn settings used by the per-fold variant-effect-prediction pipeline of Appendix F ($n_{\mathrm{NN}} = 5$ for the mutual-information estimator, $n_{\mathrm{est}} = 200$ trees with `min_samples_leaf` $= 2$ for the random-forest regressor) are retained from preliminary tuning on a held-out protein subset; all other parameters use scikit-learn defaults.

Table 1. Sparse autoencoder training hyperparameters (per ESM-2 layer).

Quantity	Symbol	Value
Input dimension	D	320
Expansion factor	r	32
Dictionary size	D_h	10,240
Batch size	B	2,048
Total training steps	T	5×10^5
Base learning rate	η_0	10^{-6}
Sparsity coefficient	λ_0	0.08
Warm-up fraction	ρ_w	0.05 ($T_w = 25{,}000$)
Gradient ℓ_2-clip	—	1.0
Random seed	—	42
Optimiser	—	ConstrainedAdam (Algorithm 1)

ConstrainedAdam. The decoder unit-norm constraint of Section 2.1 is maintained by a first-order retraction onto the product of unit spheres. Algorithm 1 composes (i) a tangent-space projection of the decoder-column gradient that drops its radial component, (ii) a standard Adam update on all parameters, and (iii) a re-projection of decoder columns to unit norm.

Algorithm 1 ConstrainedAdam (one optimiser step on the SAE).

Input: parameters $(\mathbf{W}_e, \mathbf{b}_e, \mathbf{W}_d, \mathbf{b}_d)$, gradients $(\mathbf{g}_{W_e}, \mathbf{g}_{b_e}, \mathbf{g}_{W_d}, \mathbf{g}_{b_d})$
for $j = 1$ **to** D_h **do**
  $\hat{\mathbf{w}}_j \leftarrow \mathbf{W}_d[:, j] / (\|\mathbf{W}_d[:, j]\|_2 + \varepsilon)$
  $\mathbf{g}_{W_d}[:, j] \leftarrow \mathbf{g}_{W_d}[:, j] - (\mathbf{g}_{W_d}[:, j] \cdot \hat{\mathbf{w}}_j)\, \hat{\mathbf{w}}_j$
end for
$(\mathbf{W}_e, \mathbf{b}_e, \mathbf{W}_d, \mathbf{b}_d) \leftarrow \mathrm{Adam}(\mathbf{W}_e, \mathbf{b}_e, \mathbf{W}_d, \mathbf{b}_d;\ \mathbf{g}_{W_e}, \mathbf{g}_{b_e}, \mathbf{g}_{W_d}, \mathbf{g}_{b_d})$
for $j = 1$ **to** D_h **do**
  $\mathbf{W}_d[:, j] \leftarrow \mathbf{W}_d[:, j] / (\|\mathbf{W}_d[:, j]\|_2 + \varepsilon)$
end for

B. ProteinGym Subset and Exclusion List

ProteinGym subset. We cap inputs at $L_{\max} = 1{,}024$ residues (the ESM-2 positional-embedding limit); sixteen ProteinGym substitution assays exceed this cap and are excluded a priori. Table 2 lists each excluded assay with its sequence length and ProteinGym `coarse_selection_type`. The cohort distribution (8 OrganismalFitness, 5 Activity, 2 Expression, 1 Binding) is consistent with the over-representation of large multi-domain proteins among long-sequence assays (BRCA1/2, the SARS-CoV-2 spike, Cas9 from Streptococcus pyogenes, the Zika virus envelope, picornavirus and HCV polymerases, NPC1).

Database releases. UniRef50 (Suzek et al., 2015), downloaded 2026_02; UniProtKB/SwissProt (UniProt Consortium, 2023), downloaded 2026_02; ESM-2 model checkpoint `facebook/esm2_t6_8M_UR50D` (Lin et al., 2023); ProteinGym substitution release v1.1, downloaded 2026_02 (Notin et al., 2023). Specific release tags are pinned in the project repository configuration and will be released with the camera-ready code package.

SwissProt kinase-domain MSA. The 3,683-record SwissProt kinase universe used in Section 2.3 and Figure 4 is the set of UniProtKB/SwissProt entries (downloaded 2026_02; UniProt query `reviewed:true`) carrying `FT DOMAIN /note="Protein kinase"` as the kinase-domain definition. The SRC reference was chosen so every aligned column

Table 2. Sixteen ProteinGym substitution assays excluded a priori under the ESM-2 t6 context cap ($L > 1{,}024$). Sorted by sequence length.

Assay (DMS_id)	L	Cohort
CAR11_HUMAN_Meitlis_2020_gof	1,154	OrgFitness
CAR11_HUMAN_Meitlis_2020_lof	1,154	OrgFitness
KCNH2_HUMAN_Kozek_2020	1,159	Activity
UBE4B_MOUSE_Starita_2013	1,173	Activity
ERBB2_HUMAN_Elazar_2016	1,255	Expression
SPIKE_SARS2_Starr_2020_binding	1,273	Binding
SPIKE_SARS2_Starr_2020_expression	1,273	Expression
NPC1_HUMAN_Erwood_2022_HEK293T	1,278	Activity
NPC1_HUMAN_Erwood_2022_RPE1	1,278	Activity
CAS9_STRP1_Spencer_2017_positive	1,368	Activity
BRCA1_HUMAN_Findlay_2018	1,863	OrgFitness
SCN5A_HUMAN_Glazer_2019	2,016	OrgFitness
POLG_CXB3N_Mattenberger_2021	2,185	OrgFitness
POLG_HCVJF_Qi_2014	3,033	OrgFitness
BRCA2_HUMAN_Erwood_2022_HEK293T	3,418	OrgFitness
A0A140D2T1_ZIKV_Sourisseau_2019	3,423	OrgFitness

maps to a SRC residue index for variant lookup. The SRC-anchored alignment was constructed with Biopython `1.87` (Cock et al., 2009) via the `PairwiseAligner` in global mode, using the BLOSUM62 substitution matrix with NCBI-default affine gap penalties (gap open $= -10$, gap extend $= -1$); each of the 3,683 kinase entries was globally aligned to the SRC reference (UniProt P12931), yielding an SRC-anchored row matrix from which per-position effective coverage n_{eff} and Henikoff position-based sequence weights are computed by the same script. Each ProteinGym kinase assay is matched to its row by exact ungapped-sequence equality.

C. DMS Score Rescaling for Visualization

DMS assays differ in selection type, dynamic range, and absolute scale, so raw fitness scores are not directly comparable across assays. To enable cross-assay visual comparison in our figures, we apply a per-assay piecewise-linear rescaling that anchors three reference points: the per-assay minimum to 0, the wild-type score to 1, and the per-assay maximum to 2. For a variant i in a given assay with raw DMS score s_i, wild-type score s_{WT}, per-assay minimum $s_{\min}$, and per-assay maximum $s_{\max}$, the normalized DMS score is defined as

$$\tilde{s}_i = \begin{cases} \dfrac{s_i - s_{\min}}{s_{\text{WT}} - s_{\min}} & \text{if } s_i < s_{\text{WT}}, \\ 1 + \dfrac{s_i - s_{\text{WT}}}{s_{\max} - s_{\text{WT}}} & \text{if } s_i \geq s_{\text{WT}}, \end{cases} \tag{3}$$

yielding values in $[0, 2]$ with the wild-type anchored at 1. This rescaling is applied only for visualization; the rank-based statistical tests reported throughout the paper are invariant to it and therefore yield identical results on raw DMS scores.

D. Robustness to Model Scale

To assess the robustness of the wild-type-aligned concept silencing read-out across model scales, the analyses of Figures 2 and 3 are repeated under an alternative model checkpoint. Specifically, the SAE pipeline is re-trained on `esm2_t33_650M_UR50D` (Lin et al., 2023) per-residue embeddings using the same training configuration as Table 1, with silencing threshold $|z| > 2$. For this 650M setting, the SAE is trained at layer 33 only (rather than across all transformer layers as in the 8M analysis), and the protein kinase–aligned feature at that layer attains $F_1 = 0.654$ (vs. $F_1 = 0.834$ at 8M layer 3). The bin definitions of Section 3.1 are recomputed per assay, with $N_{\max}$ determined separately for each assay. All other procedures—DMS-score rescaling (Appendix C), MSA construction (Section 2.3), and statistical testing (Mann–Whitney U with Benjamini–Hochberg FDR correction)—are held identical to the main analyses. These two figures are reported as sensitivity analyses; the BH correction is applied within each panel and is not pooled with the main-figure family.

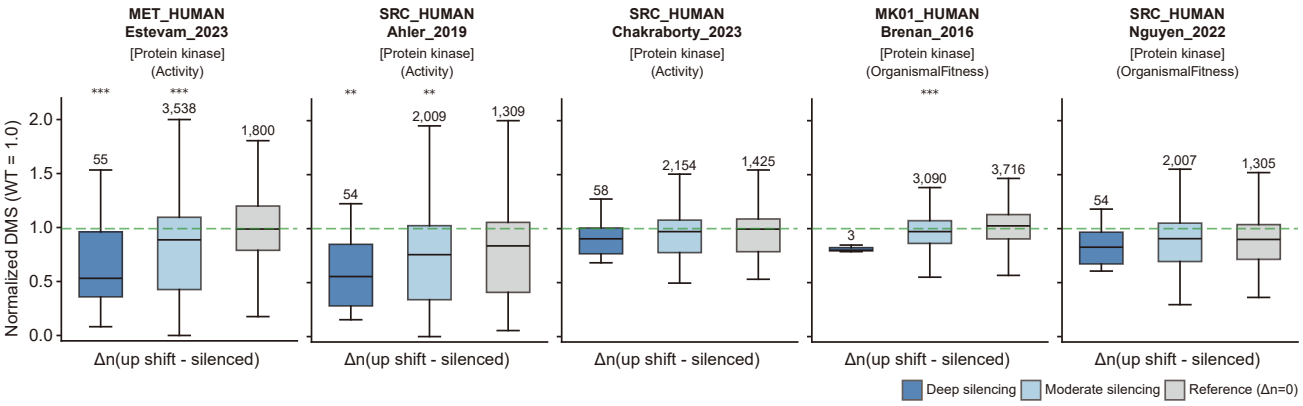

Figure 5. **Concept feature silencing tracks DMS loss of function in kinase pairs at ESM-2 650M.** The box-plot construction of Figure 2 is re-run on the five kinase pairs (the top-row assays of Figure 2) using `esm2_t33_650M_UR50D` layer 33 ($F_1 = 0.654$ against the SwissProt Protein kinase annotation) with silencing threshold $|z| > 2$. The moderate and deep-silencing bins are recomputed with $N_{\max}$ determined separately per assay; the zero-shift reference and statistical procedure (two-sided Mann–Whitney U test with Benjamini–Hochberg FDR correction applied within each panel; *, **, *** indicate $p < 0.01, 0.001, 10^{-5}$) are otherwise identical to Figure 2. Numbers above each box indicate per-bin variant counts. The non-kinase pairs of Figure 2 are not included in this robustness check.

Figure 6. **Deep silencing converges on kinase subdomain IX at ESM-2 650M.** The SRC-anchored MSA of Figure 3 (MET, MK01, and three SRC sets, with NTRK1 and FLT3 reference rows) is re-generated using deep-silencing positions identified from variants in the 650M setting (`esm2_t33_650M_UR50D` layer 33; $|z| > 2$; bins recomputed per assay) of Figure 5; the MSA construction (Section 2.3) and position-marking criterion are otherwise identical to Figure 3. Dark-blue cells mark positions containing deep-silencing-bin variants; light-blue cells mark NTRK1 and FLT3 subdomain IX residues whose substitutions are reported to impair kinase activity (Serizawa et al., 2016).

E. Structural Projection of Deep-Silencing Positions

In Figure 7, the deep-silencing positions from Section 3.2 cluster within a contiguous region of the SRC kinase fold, consistent with the subdomain IX localization shown in Figure 3.

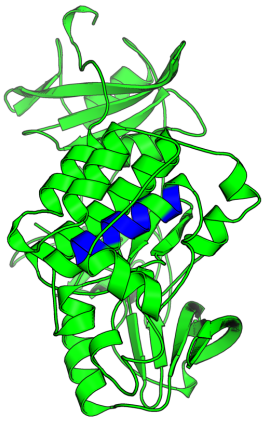

Figure 7. **Structural projection of deep-silencing positions onto the SRC kinase fold.** The deep-silencing positions identified in Figure 3 are mapped onto the AlphaFold2-predicted SRC structure and shown as dark-blue residues, while all remaining positions are colored green for contrast. The N-terminal segment (residues 1–70), which is predicted to be unstructured and lies outside the kinase domain, is hidden.

F. Variant-Effect Prediction Protocol and Parity

For each protein and each single-substitution variant i, we form a feature vector spanning the entire residue support of the protein and all six trained transformer layers. For the SAE representation, we restrict the features to the $|\mathcal{S}| = 168$ concept-feature subset; flattening the per-residue activations over (position, SAE-feature) yields a vector of length $N \times 168$, where $N \leq 1{,}024$ is the per-protein sequence length. The matched raw-ESM-2 baseline applies flattening over (layer, position, hidden-dim), yielding a vector of length $L \times N \times d_{\text{model}}$, where $L = 6$ and $d_{\text{model}} = 320$.

For each protein, the variant set is partitioned into five disjoint train/test splits with `KFold(n_splits=5, shuffle=True, random_state=42)`, and the same fold seed is used for both representations. On each training fold, columns are scored by mutual information (Pedregosa et al., 2011, `mutual_info_regression`, $n_{\text{NN}} = 5$, training half only), and the top $n_{\text{select}} = 1{,}000$ columns are retained. A random-forest regressor (Breiman, 2001, `n_estimators` $= 200$, `min_samples_leaf` $= 2$, `max_depth=None`, otherwise `scikit-learn` defaults) is then fit on the retained features and evaluated on the held-out fold, yielding Spearman ρ and Pearson r per fold. Both approaches share the fold splits, the MI estimator and its parameters, and the random-forest hyperparameters; the only asymmetry at the input is the candidate pool before MI selection (168 features per position for the SAE representation versus $L \times d_{\text{model}} = 1{,}920$ for the ESM-2 representation).

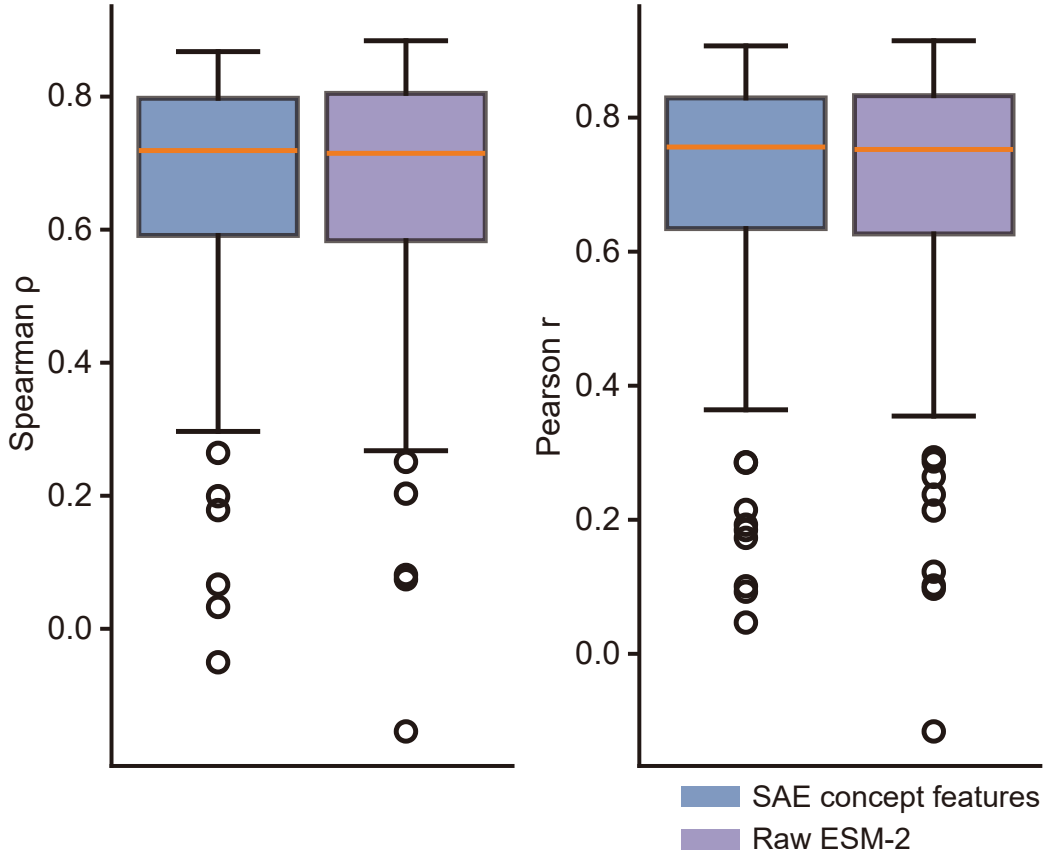

Figure 8. **The interpretable concept basis preserves variant-effect signal.** Distributions of Spearman ρ (left) and Pearson r (right) of random-forest predictions of normalized DMS scores across $N = 201$ ProteinGym substitution assays under matched protocols. Each box summarizes per-assay performance; orange line, median; box, interquartile range; whiskers, $1.5\times$IQR; circles, outliers. The two representations show closely matched medians (SAE concept features: Spearman $\rho = 0.719$, Pearson $r = 0.756$; raw ESM-2: Spearman $\rho = 0.715$, Pearson $r = 0.753$).

G. Bidirectional Concept Shifts in Kinase Activity Assays

In Figure 9, normalized DMS scores in the five kinase activity assays increase monotonically as the net concept shift Δn becomes more positive, mirroring the decrease seen in the silencing regime.

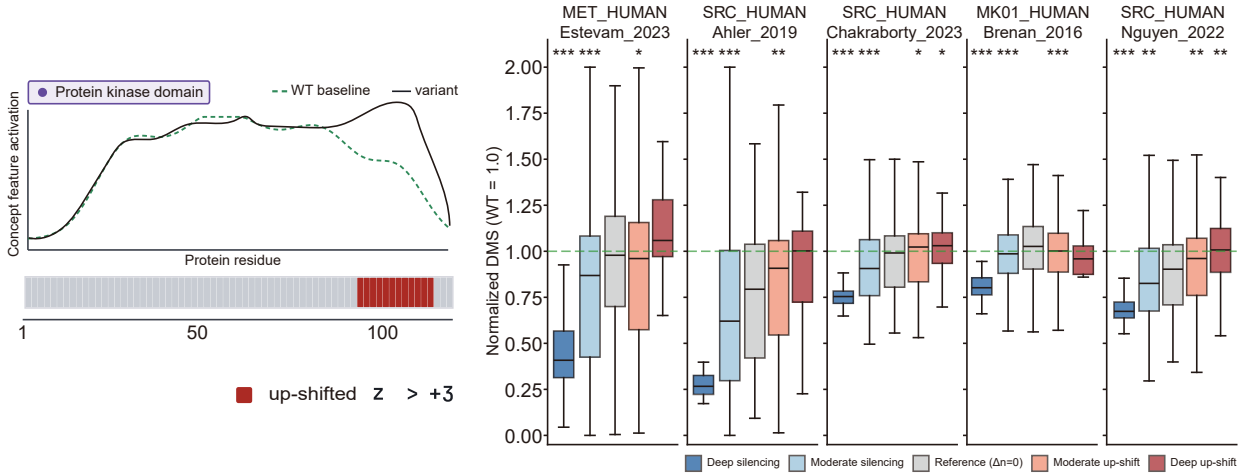

Figure 9. **Bidirectional concept shifts track DMS scores in kinase activity assays.** The box-plot construction from Figure 2 is extended to the full range of the net concept shift $\Delta n = n^{\mathrm{up}} - n^{\mathrm{down}}$ for the five kinase activity assays (single row). Variants are partitioned into five symmetric bins: deep silencing (dark blue), moderate silencing (light blue), zero-shift reference (gray), moderate up-shift (light red), and deep up-shift (dark red). The dashed green line indicates the wild-type baseline (Normalized DMS $= 1.0$; Appendix C). In all five assays, normalized DMS scores increase monotonically as Δn becomes more positive, mirroring the decrease observed in the silencing regime (Figure 2). Statistical significance relative to the zero-shift reference was assessed by a two-sided Mann–Whitney U test with Benjamini–Hochberg FDR correction applied within each panel (* $p < 0.01$, ** $p < 0.001$, *** $p < 0.00001$).

H. Limitations and Future Work

Model scale and SAE design choices. This study trains the SAE primarily on the 8M-parameter variant of ESM-2 (320-dimensional embeddings); larger PLMs may capture additional biological concepts under SAE analysis (Simon & Zou, 2025). A robustness check at ESM-2 650M reproduces the kinase-domain trends of Figures 2 and 3 (Appendix D), but a full assessment across the ESM-2 scale ladder remains future work. We also adopt a single SAE design (ReLU + ℓ_1, expansion factor $r = 32$); alternative architectures such as TopK and JumpReLU SAEs may yield concept feature sets with different coverage and specificity. Even with this design, sparse autoencoders do not yet achieve a one-to-one mapping between dictionary atoms and biological concepts—feature splitting and residual polysemanticity remain unresolved (Bricken et al., 2023), and a feature with $F_1 \geq 0.5$ may still partially absorb signal from neighboring or related concepts. In addition, our concept features are bounded by the SwissProt annotation vocabulary and a one-feature-per-concept aggregation, leaving mechanisms outside this vocabulary—and regions lacking curated annotations such as intrinsically disordered regions—invisible to the read-out.

From a zero-shot read-out to interpretable variant-effect models. The present read-out is intentionally minimal: it counts residue positions at which a concept feature falls below a wild-type baseline. This simplicity is by design—it keeps the read-out fully zero-shot and supervision-free—but it also leaves substantial headroom on the modeling side. A natural next step is to build supervised machine-learning or deep-learning variant-effect models that take concept-feature activations (rather than raw ESM-2 embeddings) as input, so that each prediction comes with per-residue and per-concept attribution drawn from a biologically grounded vocabulary. Such concept-feature-based models would (i) enable head-to-head comparison against specialized variant predictors—for example, AlphaMissense (Cheng et al., 2023) or EVE (Frazer et al., 2021)—on shared benchmarks, and (ii) operationalize the interpretability advantage in a way the present diagnostic read-out only foreshadows. We view this transition from a single zero-shot signal to a family of concept-feature-based interpretable models as the most direct path from our findings to clinically usable variant-effect tools.

I. Code and Data Availability

All code and the trained sparse autoencoders are publicly available at `https://github.com/Doyeon-Ha/lof_sae` under the MIT License. Evaluation pipelines—the wild-type-trained concept-feature read-out on ProteinGym substitution

assays and the SwissProt kinase-domain MSA analysis—are reproducible from public resources per Appendices A and B: ProteinGym (Notin et al., 2023) substitution data and UniProtKB/SwissProt (UniProt Consortium, 2023) annotations are publicly available, and specific release versions together with the resulting set of $N = 201$ retained assays (after the $L_{\max} = 1{,}024$ filter; see Table 2) are recorded in Appendix B. The 168-feature concept-alignment table and the kinase MSA construction script will be released with the same repository.