SPLIT, NOT SPARSE: FEATURE RECOVERY ACROSS SCIENTIFIC DOMAINS AND MODEL SCALE

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecular and protein transformers predict properties such as hERG potassium-channel blockade and DNA binding from chemical structures and amino-acid sequences. Explaining these predictions requires identifying which scientific properties the models represent and how the associated directions affect their outputs. Sparse autoencoders (SAEs) express internal activations through learned features, with only a small subset active per input. Evaluating these features requires measuring both their coverage of a property across examples and prediction sensitivity to removing their associated directions. We introduce a *paired recovery-and-removal protocol* that ranks features against an external annotation, decodes it from feature groups, and removes the corresponding decoder directions with a random-removal control. We report three findings. **(1) Group Recovery:** groups improve property decoding in a ChemBERTa hERG predictor and three ESM-2 DNA-binding predictors. For the molecular conjunction of a basic nitrogen and an aromatic system, top-64 features reach 0.842 AUROC, versus 0.740 for one feature and 0.853 for the full layer. This recovery curve supports feature-budget selection. **(2) Recovery–Intervention Dissociation:** ESM-2 top-64 decoding stays within 0.026 AUROC of the full-layer reference, while targeted-minus-random removal effects range from 0.041 at 35M to 0.002 at 650M under the evaluated budgets. The paired measurements distinguish property recovery from removal sensitivity. **(3) Target-Dependent Coverage:** an aromatic-plus-basic protein target leaves reference gaps up to 0.116 AUROC at top-64, identifying settings where the selected group provides incomplete coverage. Together, the protocol and findings support choosing feature groups for a specified scientific property and evaluating their associated prediction effects.

1 INTRODUCTION

Predicting hERG potassium-channel blockade from a molecule or DNA binding from a protein sequence leaves an interpretability question: which properties of the input does the model represent, and how do the associated representations influence its predictions? Molecular-transformer analysis and hERG attribution address this question through chemical components and learned computations (Varadi et al., 2025; Lee & Yoo, 2025).

Sparse autoencoders (SAEs) approximate a model's activation vector using learned dictionary features, with only a small subset active for each input (Bricken et al., 2023; Huben et al., 2024; Templeton et al., 2024). A feature can then be examined for alignment with a recognizable property. Two questions remain: how much of that property is recovered across examples, and how sensitive are predictions to the associated directions? Feature splitting and absorption motivate looking beyond one dictionary entry (Chanin et al., 2025; Bussmann et al., 2025); downstream probing studies motivate comparison against strong reference representations (Kantamneni et al., 2025).

We study properties defined independently of the SAE. In a ChemBERTa hERG predictor, an RDKit indicator labels molecules containing both a basic nitrogen and an aromatic system. In three ESM-2 DNA-binding predictors, the target is basic-residue composition. These annotations fix what information to recover. Removing associated decoder directions then measures prediction sensitivity. A separate aromatic-plus-basic protein target tests recovery across targets and representation settings.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We pair linear probes on selected SAE features with removal of their decoder directions, comparing each readout with its own controls. We report three findings:

(1) Group Recovery. Combining aligned features recovers more target information than the highest-alignment single feature. The hERG top-64 probe improves AUROC from 0.740 to 0.842; the three protein task models gain 0.151–0.236. A nine-configuration chemistry sweep tests SAE family, capacity, and sparsity. These results motivate evaluating target coverage across a feature budget.

(2) Recovery–Intervention Dissociation. Similar group decoding accompanies different prediction sensitivity to decoder-direction removal. The protein intervention margin, defined as targeted drop minus random-control drop, declines from 0.041 at 35M to 0.002 at 650M. The paired readouts distinguish the information accessible in a feature group from prediction sensitivity to its associated decoder span.

(3) Target-Dependent Coverage. The remaining gap to a full-layer probe depends on the annotation and representation setting. Top-64 gaps are at most 0.026 for the task-conditioned targets but reach 0.116 for the separate protein target. This finding makes target choice part of dictionary evaluation: group improvement and near-reference recovery are separate outcomes.

The paired protocol supports feature-budget selection, identifies incomplete target coverage, and directs follow-up interventions. Chemistry supplies the detailed case study; proteins extend the analysis across family and size. AMES provides an auxiliary attention-head localization check.

2 Method: Paired Recovery-and-Removal Analysis

Our method evaluates an SAE explanation through a pair of measurements: recovery of a specified property and prediction sensitivity to removal of associated decoder directions. The input is a predictor, a layer representation, an SAE dictionary, and an external property annotation. Its output is a recovery profile across feature budgets, paired with a targeted-versus-random intervention contrast. Figure 1 summarizes the shared ranking and the two measurement paths.

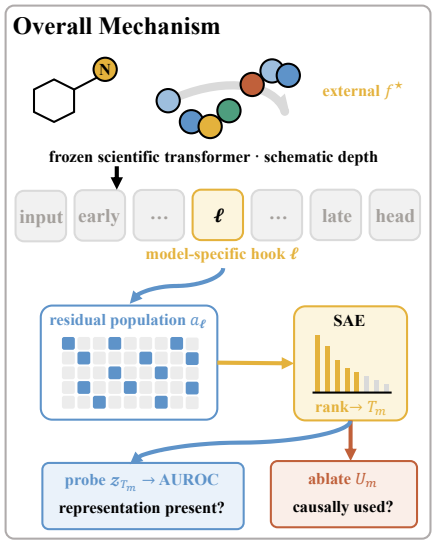

Figure 1: Study overview. External annotations define a molecular or sequence property. An SAE encodes pooled activations; target alignment ranks its features. The probe decodes the property from selected activations, while intervention removes associated decoder directions and reevaluates predictions. Hooks are model-specific.

2.1 Why These Models and Targets?

Chemistry uses a ChemBERTa-based hERG predictor (Chithrananda et al., 2020; Singh et al., 2026) and an RDKit-defined indicator for the joint presence of a basic nitrogen and an aromatic system (RDKit, n.d.). This conjunction provides a concrete test of compositional coverage: a representation of either component captures only part of the target. The predictor uses ChemBERTa-100M-MLM with 12 layers and residual width $d = 768$ (DeepChem, 2025). Task adaptation uses LoRA on query and value projections, with rank 8 and scaling parameter $\alpha = 16$ (Hu et al., 2022).

The chemistry analysis uses TDC hERG and AMES (Huang et al., 2021). hERG is the detailed case study: it supports the SAE sweep, the feature-budget curve, and decoder-span interventions. AMES supplies a second endpoint for the attention-head localization analysis. This design separates the detailed evaluation of a molecular property from the cross-endpoint comparison of component sensitivity.

Biology uses ESM-2 at 35M, 150M, and 650M parameters (Lin et al., 2023), with residual widths 480, 640, and 1280. The DNA-binding task uses UniProt keyword KW-0238 to define positives (The UniProt Consortium, 2023). The property decoded in these predictors is basic-residue composition (K+R), binarized at the median. A separate pretrained-model analysis evaluates aromatic-plus-basic

composition, including human and *E. coli* sequences. The first analysis pairs recovery with task interventions; the second examines coverage across target and representation settings.

Task prediction and property decoding use different labels. The hERG predictor estimates blockade, whereas its probe reads the molecular indicator; the DNA-binding predictors classify sequences, whereas their probes read residue composition. Keeping these labels separate makes a high probe score interpretable as property recovery and leaves the relationship to task behavior for the intervention analysis.

The molecular and protein models provide two scientific input types for testing the same analysis procedure. Within ESM-2, the size series compares recovery and intervention profiles across increasingly wide representations. The additional protein target examines how coverage changes with the annotation, organism, and layer. Finally, a synthetic experiment varies concept density and spread to examine the geometry of single-feature recovery. Together, these studies connect a detailed molecular example, cross-family recovery evidence, a protein scale comparison, and a controlled geometric analysis.

2.2 Feature selection and decoding

Let $h_\ell(x,t) \in \mathbb{R}^d$ denote a token-level residual activation and $a_\ell(x)$ its mean-pooled representation. An externally specified binary annotation is $f^\star(x)$. We fit an SAE with latent activations $z_j(a)$ and decoder columns v_j, so that $a \approx b + \sum_j z_j(a) v_j$. The default configuration is TopK with an $8\times$ dictionary and $k = 32$ active latents; the chemistry sweep also includes ReLU and JumpReLU (Gao et al., 2025; Rajamanoharan et al., 2024).

The alignment score and selected indices are

$$s_j = |\operatorname{corr}(z_j(a_\ell(x)), f^\star(x))|, \qquad T_m = \operatorname{top}_m\{s_j\}. \tag{1}$$

Write $D(T_m)$ for the AUROC of a linear probe on the selected *latent activations* $z_{T_m}(a)$, and $D(\text{full})$ for a linear probe on the pooled layer representation a_ℓ. The highest-correlation latent T_1 provides the single-feature comparator. These definitions fix three useful reference points: the strongest target-aligned dictionary entry, a selected feature group, and the full-layer representation. Feature ranking, probe fitting, and held-out scoring are separate stages of the evaluation protocol.

We quantify group recovery using its gain and remaining gap,

$$G_m = D(T_m) - D(T_1), \qquad H_m = D(\text{full}) - D(T_m). \tag{2}$$

We call G_m the *group-recovery gain* and H_m the *reference gap*. The gain measures the benefit of combining features; the gap measures the remaining difference from the full-layer probe. Together, the gain and gap distinguish improvement over one latent from decoding performance close to the full-layer reference. Random groups of the same size provide a feature-selection control. We use AUROC for property decoding, absolute correlation for latent alignment, and reconstruction R^2 for SAE fidelity.

2.3 Interventions and controls

For the selected decoder matrix $V_m = [v_j]_{j \in T_m}$, let U_m contain an orthonormal basis for its column span, with rank $r_m \le m$. The projection intervention is

$$h' = h - U_m U_m^\top h, \qquad \kappa(T_m) = Q(M) - Q(M \mid \operatorname{remove}(U_m)). \tag{3}$$

Here Q denotes the prediction or task score evaluated before and after intervention, with positive κ indicating a decrease. The intervention removes the selected span at a residual-stream hook and resumes the downstream computation. The right panel of Figure 2 illustrates this operation. Each intervention margin compares targeted and random removal using the same prediction score and evaluation setting. The reported effects retain that score's units.

The two measurements use complementary parts of the SAE. Its encoder supplies the latent activations for property decoding; its decoder supplies the directions for removal from the model representation. Shared feature indices connect the measurements to the same target-based ranking. The probe therefore evaluates the information accessible from the selected features, while the intervention evaluates prediction sensitivity to the associated directions. Random removal measures the background sensitivity to an edit of comparable size.

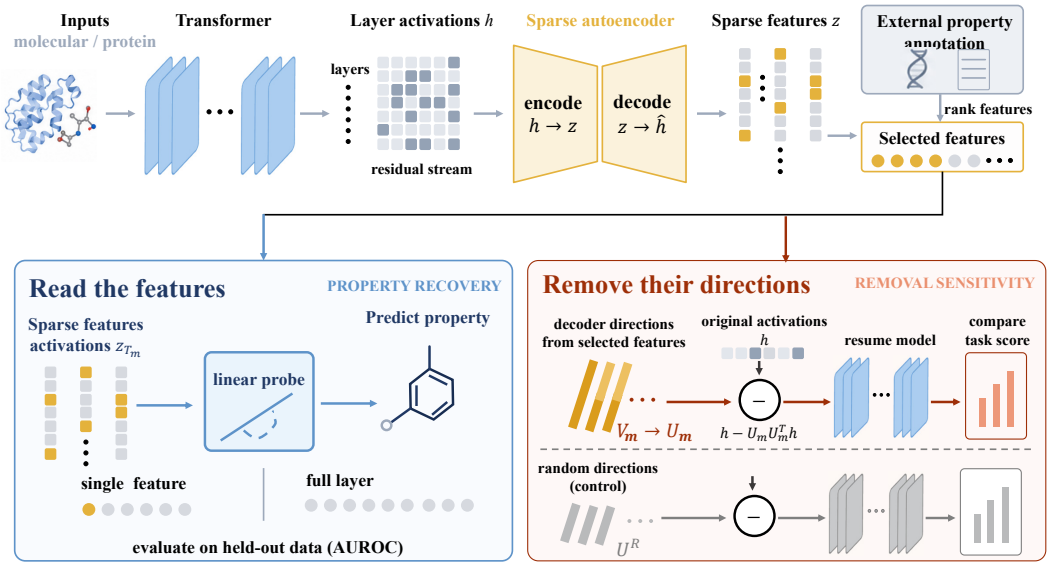

Figure 2: Paired recovery-and-removal analysis. External property annotations rank SAE features extracted from layer activations. Left: a linear probe decodes the property from selected feature activations, with single-feature and full-layer comparators evaluated on held-out data. Right: selected decoder columns define a span, whose orthonormal basis U_m is used to remove its projection from the original activations before prediction resumes. Random-direction removal provides a control. For hERG, single-feature, top-64, random-64, and full-layer decoding reach 0.740, 0.842, 0.577, and 0.853 AUROC; targeted and random removal effects are 0.067 and 0.008.

For a control span R of the same rank within a model, define

$$\mu_R = \mathbb{E}_R[\kappa(R)], \qquad \rho_m = \kappa(T_m) - \mu_R. \tag{4}$$

We call ρ_m the *intervention margin*. It measures the additional drop associated with targeted removal relative to the random control. The targeted/random *effect ratio* gives the corresponding relative effect. We report the original drops alongside these summaries so that a large ratio can be distinguished from a large absolute change. The control specification includes the span-sampling rule and numerical rank; these determine which background perturbations the comparison represents.

Interventions select $m = 64, 32, 32, 64$ columns for chemistry, 35M, 150M, and 650M, respectively. Relative to residual width, these budgets are $8.3\%, 6.7\%, 5.0\%, 5.0\%$. The selected-column budget m and numerical span rank r_m describe the intervention's size in different ways. Property decoding uses 64 features for the cross-model comparison; removal uses the model-specific budgets above. Reporting both budgets preserves the link between the recovery profile and the edit actually evaluated.

Intervention implementation. SAEs are fitted to mean-pooled activations. Decoder-span removal is applied at a residual-stream hook, after which the forward computation resumes. For the component analysis, an attention head is ablated by zeroing its context slice through a forward pre-hook on the attention output projection. The selected head's contribution is therefore removed before the output projection combines head outputs. These operations provide the representation-level and component-level interventions evaluated in the chemistry analysis. Head-level effects are evaluated on both hERG and AMES to compare localization across the two endpoints.

2.4 Reading the Paired Profile

The procedure consists of four stages. First, choose a scientific annotation and rank SAE features by their absolute correlation with it. Second, fit linear probes at several feature budgets and compare their scores with a single-feature probe, a full-layer probe, and available random-group controls. Third, project out the selected decoder span and evaluate the predictor against a random-removal control. Fourth, summarize the result through group gain, reference gap, and intervention margin. This produces a target-specific evaluation of a dictionary and its associated directions.

Each output supports a concrete analysis choice. The recovery curve indicates how many selected features are useful for decoding the target. The reference gap measures the difference in target-

decoding AUROC between the full-layer probe and the selected-group probe. The intervention margin measures how strongly the selected directions affect the evaluated prediction pathway relative to the control. For example, a large group gain with a substantial reference gap identifies a useful but incomplete feature group; near-reference recovery with a small intervention margin identifies a group that decodes the annotation well while its decoder-span removal has a small additional effect over random removal.

For hERG, the top-64 gain, reference gap, and intervention margin are 0.102, 0.011, and 0.059, respectively. The protein analyses use the same definitions. Appendix A gives the model configuration.

A feature-budget criterion for the recovery path in Figure 2 is $D(T_m) \geq (1-\varepsilon)D(\text{full})$, where ε is the tolerated relative shortfall in decoding AUROC. Equivalently, the reference gap must satisfy $H_m \leq \varepsilon D(\text{full})$. The reference and tolerance therefore determine the required decoding score before comparing feature budgets. The $|r| = 0.50$ alignment threshold concerns alignment of an individual SAE feature with the annotation. Using the corresponding metric for each measurement separates single-feature alignment, group decoding, and prediction sensitivity. The intervention panel then compares the effect of removing the selected decoder span against random removal, using the same prediction score within that comparison.

3 Group Recovery: combining features improves target decoding

Group Recovery denotes the observed gain from combining SAE features. The hERG case supplies the detailed feature-budget curve; the DNA-binding models repeat the gain at three sizes. The separate protein-composition analysis shows where the remaining gap becomes substantial.

3.1 Chemistry: recovery improves across a feature group

On hERG, decoding improves smoothly as more aligned features enter the probe: 0.740, 0.771, 0.801, 0.819, and 0.842 AUROC for $m = 1, 8, 16, 32, 64$. The top-64 gain is 0.102, and its gap to the full-layer reference is 0.011. Random-64 decoding is 0.577, so the aligned group exceeds that control by 0.265. The left path in Figure 2 illustrates the decoding procedure used for these comparisons; random groups additionally control for feature selection.

Table 1: Complete chemistry SAE sweep. Dictionary size is relative to residual width. Compare reconstruction R^2 with best single-feature alignment $|r|$; all nine values of $|r|$ are below 0.50.

SAE	Dict.	Setting	R^2	$\lvert r\rvert$
ReLU	$4\times$	$\lambda_{L1}{=}4\times10^{-3}$	.93	.28
TopK	$8\times$	$k{=}16$	.97	.27
TopK	$8\times$	$k{=}32$	.98	.39
TopK	$8\times$	$k{=}64$	.98	.41
TopK	$16\times$	$k{=}32$	.98	.30
TopK ($L2$)	$8\times$	$k{=}32$	.98	.33
JumpReLU	$8\times$	$c{=}2\times10^{-4}$	.97	.39
JumpReLU	$8\times$	$c{=}5\times10^{-4}$	.97	.33
JumpReLU	$8\times$	$c{=}1\times10^{-3}$	.96	.38

The chemistry sweep varies SAE family, dictionary capacity, and activation sparsity. Across its nine configurations, the best single-feature alignment ranges from 0.27 to 0.41, while reconstruction R^2 ranges from 0.93 to 0.98 (Table 1). The dictionaries thus preserve most activation variance while giving different degrees of alignment with the molecular target. This comparison evaluates whether a configuration that reconstructs well also provides a useful individual feature for the annotation.

The sweep separates capacity from sparsity. At $8\times$ dictionary size, increasing TopK's active-latent budget from 16 to 32 to 64 raises alignment from 0.27 to 0.39 to 0.41. Doubling dictionary size at $k = 32$ instead lowers alignment from 0.39 to 0.30, while reconstruction remains 0.98. The larger dictionary preserves reconstruction fidelity while its best individual feature has lower target alignment. JumpReLU alignment is also nonmonotone in c. Target alignment distinguishes these dictionaries even when their reconstruction scores coincide.

The feature-budget curve reveals how this target is recovered across dictionary entries. Adding seven features raises AUROC by 0.031; extending the group from eight to sixteen adds 0.030. The next two expansions add 0.018 and 0.023. Recovery therefore continues to improve throughout the measured

range. The top-64 score of 0.842 approaches the full-layer value of 0.853, making the selected group a compact representation of this particular molecular indicator.

The budget curve also supports an explicit coverage decision. A budget of 16 features yields a reference gap of 0.052; increasing it to 32 reduces the gap to 0.034, and increasing it to 64 reduces it to 0.011. For this target, a reader seeking a gap below 0.02 AUROC would select 64 among the evaluated budgets. A budget of 32 offers a smaller group with a larger remaining gap.

The chemistry intervention links this recovery result to prediction sensitivity. Removing the top-64 decoder span produces an effect of 0.067, compared with 0.008 for random removal. The intervention margin is therefore 0.059, and the effect ratio is approximately 8.4. Read together with the probe curve, this shows that the selected group both decodes the molecular conjunction and identifies directions whose removal affects the hERG predictor more than the random control.

Component localization. The component analysis provides context for the group-level result. Early MLP components dominate the hERG localization, whereas the largest attention-head effect is 0.009. The corresponding maximum on AMES is 0.005. These endpoint analyses provide a component-level view of the predictors, complementing the target-based SAE analysis. The auxiliary circuit summary retains 1024 of 6144 neurons at 98% faithfulness, with attribution approximately $104\times$ the random-k baseline (Appendix G). The dominant MLP contributions occur at layers 1–2. The accompanying neuron-level readout is $R^2 = 0.047$; this recovery measure complements the circuit summary by describing the target information captured at the neuron level (Appendix G).

The single-feature operating-point summary provides another view of coverage: precision is approximately 0.87 and recall approximately 0.60 for the molecular conjunction (Appendix C). Thus, annotated examples are concentrated among those selected by the feature, while a substantial portion of the annotated set lies outside that selection. The group-probe curve evaluates how target prediction improves when several aligned features are combined, complementing this individual feature's precision and recall. Precision measures the proportion of feature-positive examples carrying the annotation; recall measures the proportion of annotated examples selected by the feature. At the reported operating point, the feature covers about three fifths of the annotated set. These operating-point measures complement AUROC, which evaluates discrimination across score thresholds.

3.2 Target-Dependent Coverage in protein representations

For the basic-residue target in the DNA-binding models, top-64 decoding exceeds the single-feature comparator by 0.236, 0.151, and 0.196 AUROC at 35M, 150M, and 650M. The remaining gaps are 0.011, 0.026, and 0.026 (Table 2). The three sizes therefore show a consistent group benefit, with top-64 scores close to their respective full-layer references.

Table 2: Task-conditioned target recovery. Task AUROC evaluates hERG or DNA binding; the next three columns evaluate the external target. All scores are AUROC. Gains and reference gaps are given in the text.

Model	Task	Single	Top-64	Full
ChemBERTa	0.940	0.740	0.842	0.853
35M	0.962	0.606	0.842	0.853
150M	0.975	0.666	0.817	0.843
650M	0.983	0.627	0.823	0.849

Target-Dependent Coverage is evident in the separate aromatic-plus-basic analysis. In 35M ESM-2, the top-64 scores are 0.815 and 0.811 at layer 6 for *E. coli* and human proteins, and 0.818 at layer 9 for *E. coli*; their reference gaps are 0.038, 0.040, and 0.035. At 150M layer 15, recovery improves from 0.598 to 0.733 but remains 0.116 below the reference. At 650M layer 16, it improves from 0.643 to 0.780 with a 0.076 gap. Table 3 compares the single-feature, group, and full-layer scores for each setting.

The two analyses separate group benefit from near-reference recovery. Aromatic-plus-basic gains are 0.122, 0.123, 0.143, 0.135, and 0.137 AUROC in table order. At 150M, the group gains 0.135 but leaves a 0.116 gap; the task-conditioned basic-residue analysis gains 0.151 and leaves only 0.026. Similar group gains therefore coexist with different gaps to the full-layer decoding reference. These profiles describe the combined target and representation settings, including the selected layer and task adaptation.

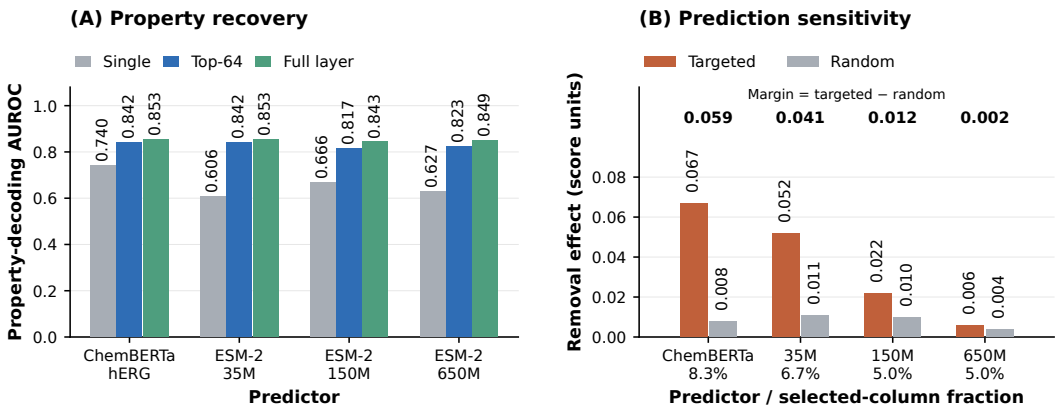

Figure 3: Recovery and intervention readouts across models. Panel (A) compares single-feature, top-64, and full-layer property-decoding AUROC from Table 2. Panel (B) plots targeted and random removal effects from Table 4; bold values give their differences. The selected-column fractions are $8.3\%, 6.7\%, 5.0\%, 5.0\%$, respectively. Protein recovery remains near the reference as intervention margins decrease across model sizes. Effects use the prediction-score units of each evaluation.

Across both protein analyses, combining features improves decoding. The task-conditioned basic-residue target has reference gaps of 0.011–0.026, compared with 0.035–0.116 for aromatic-plus-basic composition in pretrained models. Target coverage is therefore a property of the target–representation pair. This distinction matters when choosing a dictionary for scientific analysis: a feature budget that works well for one annotation can leave a substantial recovery gap for another.

Table 3: Boundary: aromatic-plus-basic composition. These pretrained-model probes use a different target from Table 2. Gap is full minus top-64. A dash denotes unspecified organism coverage.

Model	Layer / data	Single	Top-64	Full	Gap
35M	L6 / *E. coli*	.693	.815	.853	.038
35M	L6 / human	.688	.811	.851	.040
35M	L9 / *E. coli*	.675	.818	.853	.035
150M	L15 / —	.598	.733	.849	.116
650M	L16 / —	.643	.780	.856	.076

4 Recovery–Intervention Dissociation across Model Sizes

Across the three protein models, top-64 decoding stays close to the full-layer reference while the intervention margin decreases from 0.041 to 0.012 to 0.002. We call this pattern *recovery–intervention dissociation*. Figure 3 places the recovery scores and removal effects side by side; Table 4 records the budgets and original effects.

4.1 Removal Sensitivity across Protein Model Sizes

The targeted effect decreases from 0.052 at 35M to 0.006 at 650M, while the random effect decreases from 0.011 to 0.004. Subtracting the controls gives margins of 0.041 and 0.002. The corresponding effect ratios are 4.73 and 1.50. Thus, the largest model has the smallest additional sensitivity to targeted removal under the evaluated intervention. The targeted and random drops show the absolute prediction changes underlying each margin and effect ratio.

The 150M and 650M models provide a comparison at the same nominal column fraction: each intervention selects 5% of residual width. Their top-64 AUROCs are 0.817 and 0.823, with the same reference gap of 0.026. Their intervention margins are 0.012 and 0.002. At this shared column fraction, similar recovery profiles accompany different targeted-versus-random removal effects. The 35M intervention selects 6.7% of width and yields the largest protein margin.

The size series is an observational comparison across checkpoints and selected layers, with model-specific intervention budgets. Uncertainty estimates and matched-fraction sweeps are needed to estimate the robustness of the small 650M contrast and isolate a scale effect.

The prediction task and decoded property also change differently across size. DNA-binding task AUROC rises from 0.962 to 0.975 to 0.983, while basic-residue top-64 decoding is 0.842, 0.817, and 0.823. Together, these scores distinguish task performance, property recovery, and removal sensitivity. In this series, task performance increases, property recovery stays near its layer-specific reference, and the targeted removal contrast becomes smaller.

Table 4: Intervention scope and reported effects. Selected-column budgets are $64/768$, $32/480$, $32/640$, and $64/1280$ (m/d), respectively. Margin is targeted minus random; ratio is targeted divided by random.

Model	Target	Random	Margin	Ratio
ChemBERTa	.067	.008	.059	8.38
35M	.052	.011	.041	4.73
150M	.022	.010	.012	2.20
650M	.006	.004	.002	1.50

4.2 Information remaining after removal

A useful next step is to probe the target after removing the selected span, then evaluate a second controlled removal of directions predictive of the remaining target information. The probe trajectory would measure how much property information survives successive edits, while the prediction trajectory would measure sensitivity to those edits. This test would assess whether information outside the first selected span contributes to the smaller intervention margins in the larger protein models.

5 Single-feature recovery decreases with concept density

In the toy grid, single-feature recovery falls as density increases. At density 0.25, scores are 0.82–0.84 across concept spreads; at density 8, they are 0.52–0.53. Figure 4 and Appendix F show the same grid. The appropriate comparison follows a column while changing density, then compares columns at fixed density. The contrast across density is larger than the contrast across spread.

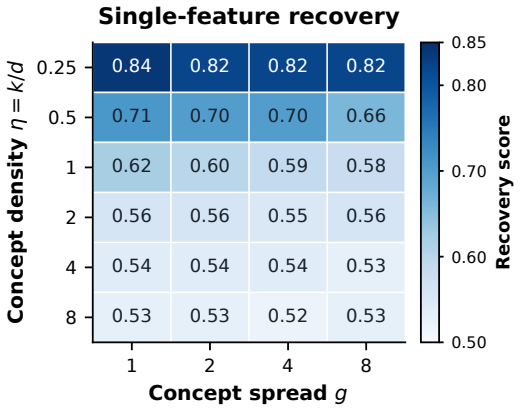

Figure 4: Toy single-feature recovery. Rows vary concept density and columns vary the number of sub-directions per concept. Read down a column to compare density at fixed spread: recovery decreases in all four columns. Density is $\eta = k/d$, and spread is g. Appendix F lists the cell values.

Density is the number of active concepts per representation dimension; spread is the number of sub-directions assigned to each concept. These variables connect the analysis to toy studies of superposition (Elhage et al., 2022). At spread $g = 1$, the recovery score falls from 0.84 to 0.71, 0.62, 0.56, 0.54, and 0.53 as density increases. At $g = 8$, the corresponding values are 0.82, 0.66, 0.58, 0.56, 0.53, and 0.53. The density trend appears across all four spread settings.

At fixed density, changing spread produces smaller differences. At $\eta = 0.25$, scores span 0.82–0.84; at $\eta = 8$, they span 0.52–0.53. The intermediate $\eta = 2$ row spans 0.55–0.56. These comparisons identify concept density as the stronger axis of variation in the reported grid. The intermediate rows show how the spread contrast changes with density: at $\eta = 0.5$, recovery is 0.71 for $g = 1$ and 0.66 for $g = 8$; at $\eta = 2$, both settings score 0.56 (Appendix F).

Group recovery in the toy. The auxiliary top-16 analysis reports a peak gain of approximately 0.27 over single-feature recovery at moderate density ($\eta = 0.5$–1) and spread $g \geq 2$. This group-minus-single contrast complements the single-feature grid by measuring the benefit of combining several selected features in that part of the toy parameter space.

Per-input sparsity and target coverage describe different aspects of a representation. Sparse activation describes how many SAE features respond to an individual input. Target recovery describes how

effectively one feature or a group predicts an annotation across the dataset. In the scientific cases, a group combines information from several entries; in the toy, increasing concept density reduces the score obtained by an individual feature. These observations motivate examining feature groups when a target is represented through several directions.

6 Related Work

Sparse autoencoders and feature coverage. SAEs provide learned dictionaries for decomposing neural activations into sparsely active features (Bricken et al., 2023; Huben et al., 2024; Templeton et al., 2024). TopK and JumpReLU architectures offer different ways to control activation sparsity and reconstruction (Gao et al., 2025; Rajamanoharan et al., 2024). Our chemistry sweep evaluates these choices against a fixed molecular annotation, making target alignment directly comparable across configurations. The group probes then assess how much of that annotation is accessible across several dictionary entries.

Feature splitting, absorption, and dictionary stability. Feature splitting describes finer decompositions across dictionary configurations; absorption concerns broad features failing to activate on examples assigned to more specific entries (Chanin et al., 2025). Matryoshka SAEs learn nested dictionaries to address hierarchical structure and evaluate their features through probing and concept erasure (Bussmann et al., 2025). Studies of canonicity and training variation examine dictionary completeness, atomicity, and stability across seeds (Leask et al., 2025; Paulo & Belrose, 2026). We use *group recovery* for the measured gain within a fitted dictionary, keeping this observable distinct from the mechanisms of splitting and absorption. External molecular and sequence annotations give the comparison a fixed scientific target.

Probing and causal interventions. Linear probes measure information accessible in learned representations (Alain & Bengio, 2016; Belinkov, 2022; Gurnee et al., 2023). Downstream SAE evaluations compare feature usefulness against alternative representations and baselines (Kantamneni et al., 2025). Circuit frameworks describe computations through connected features and attention pathways (Olah et al., 2020; Elhage et al., 2021). Intervention methods test behavioral contributions of components, edges, and representations (Wang et al., 2023; Conmy et al., 2023; Bhaskar et al., 2024; Meng et al., 2022). Our protocol joins property decoding and decoder-span removal through a common feature ranking, and reports a recovery gap alongside the intervention contrast. This pairing supports comparisons between accessible property information and prediction sensitivity within the same scientific application.

Interpretability in scientific models. Molecular-transformer analysis studies chemical features and computations (Varadi et al., 2025), while hERG attribution relates toxicity predictions to molecular components (Lee & Yoo, 2025). Binding-prediction research also emphasizes the effects of dataset structure on generalization (Chatterjee et al., 2023). We study a concrete molecular conjunction in an hERG predictor and extend the recovery analysis to residue-composition properties in ESM-2. The chemistry case supplies the detailed dictionary and intervention comparisons; the protein analyses examine feature coverage across sizes and target settings. Appendix B summarizes these connections.

7 Conclusion

We introduce a paired recovery-and-removal protocol for evaluating SAE explanations against scientific annotations. Feature groups improve target decoding across the molecular and protein cases. In the protein size series, near-reference decoding accompanies decreasing intervention margins, while a separate protein target leaves larger reference gaps. These findings connect feature-group recovery, target-dependent coverage, and prediction sensitivity to the associated decoder span.

The recovery curve supports feature-budget selection, the full-layer comparison measures the remaining decoding gap, and controlled removal evaluates prediction sensitivity. Together, these measurements help assess a feature group as an explanation of a specified scientific property. Post-removal probes and matched-fraction interventions would extend the analysis to information surviving an edit and its contribution to prediction.

Reproducibility Statement

The ranking, probing, and projection operations are defined in Section 2; model configurations and numerical summaries are provided in the tables and appendices. The accompanying materials contain manuscript source and figure assets.

Ethics Statement

The study analyzes molecular and protein predictors using public TDC and UniProt resources. Molecular and sequence descriptors serve as operational annotations for interpretability experiments. Toxicity assessment and biological applications require task-specific domain validation; the present evaluation concerns model representations and prediction sensitivity.

AI Use Statement

AI tools assisted with manuscript drafting, editing and translation, literature searches, methodological feedback, result interpretation, and code writing. The authors are responsible for the final manuscript and code.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. URL https://arxiv.org/abs/1610.01644.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Adithya Bhaskar, Alexander Wettig, Dan Friedman, and Danqi Chen. Finding transformer circuits with edge pruning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 18506–18534, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/20fdaf67581e6d7157376d1ed584040a-Abstract-Conference.html.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with Matryoshka sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6077–6101. PMLR, 2025. URL https://proceedings.mlr.press/v267/bussmann25a.html.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *Advances in Neural Information Processing Systems*, volume 38, pp. 82318–82355, 2025. doi: 10.52202/085713-2755. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/764ff7477b8e24dbe01531f6791e8bdf-Abstract-Conference.html.

Ayan Chatterjee, Robin Walters, Zohair Shafi, Omair Shafi Ahmed, Michael Sebek, Deisy Gysi, Rose Yu, Tina Eliassi-Rad, Albert-László Barabási, and Giulia Menichetti. Improving the generalizability of protein-ligand binding predictions with AI-Bind. *Nature Communications*, 14 (1):1989, 2023. doi: 10.1038/s41467-023-37572-z. URL https://doi.org/10.1038/s41467-023-37572-z.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020. URL https://arxiv.org/abs/2010.09885.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 16318–16352, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

DeepChem. ChemBERTa-100M-MLM: Model configuration, 2025. URL https://huggingface.co/DeepChem/ChemBERTa-100M-MLM/blob/20370e671b72b25bfd0ffdbac855b52472349e21/config.json. Configuration published August 18, 2025; accessed September 22, 2026.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/42ef3308c230942d223c411adf182c88-Abstract-Conference.html.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=JYs1R9IMJr.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/4c56ff4ce4aaf9573aa5dff913df997a-Abstract-round1.html.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1fa1ab11f4bd5f94b2ec20e794dbfa3b-Abstract-Conference.html.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? A case study in sparse probing. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29018–29049. PMLR, 2025. URL https://proceedings.mlr.press/v267/kantamneni25a.html.

Abdul Karim, Matthew Lee, Thomas Balle, and Abdul Sattar. CardioTox net: a robust predictor for hERG channel blockade based on deep learning meta-feature ensembles. *Journal of Cheminformatics*, 13(1):60, 2021. doi: 10.1186/s13321-021-00541-z. URL https://doi.org/10.1186/s13321-021-00541-z.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/84ca3f2d9d9bfca13f69b48ea63eb4a5-Abstract-Conference.html.

Dohyeon Lee and Sunyong Yoo. hERGAT: predicting hERG blockers using graph attention mechanism through atom- and molecule-level interaction analyses. *Journal of Cheminformatics*, 17(1):11, 2025. doi: 10.1186/s13321-025-00957-x. URL https://doi.org/10.1186/s13321-025-00957-x.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://doi.org/10.1126/science.ade2574.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in/.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3c1fe56b043848b211030c202764c6a7-Abstract-Conference.html.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024. URL https://arxiv.org/abs/2407.14435.

RDKit. RDKit: Open-source cheminformatics, n.d. URL https://www.rdkit.org. Accessed September 22, 2026.

Riya Singh, Aryan Amit Barsainyan, Rida Irfan, Connor Joseph Amorin, Stewart He, Tony Davis, Arun Thiagarajan, Shiva Sankaran, Seyone Chithrananda, Walid Ahmad, Derek Jones, Kevin McLoughlin, Hyojin Kim, Anoushka Bhutani, Shreyas Vinaya Sathyanarayana, Venkat Viswanathan, Jonathan E. Allen, and Bharath Ramsundar. ChemBERTa-3: An open source training framework for chemical foundation models. *Digital Discovery*, 5:662–685, 2026. doi: 10.1039/D5DD00348B. URL https://doi.org/10.1039/D5DD00348B.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan.

Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2023. doi: 10.1093/nar/gkac1052. URL https://doi.org/10.1093/nar/gkac1052.

Therapeutics Data Commons. Toxicity: Dataset documentation, n.d. URL https://tdcommons.ai/single_pred_tasks/tox/. Accessed September 22, 2026.

Kristof Varadi, Mark Marosi, and Péter Antal. Circuits, features, and heuristics in molecular transformers. *arXiv preprint arXiv:2512.09757*, 2025. URL https://arxiv.org/abs/2512.09757.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

A Model and intervention configuration

The molecular predictor uses ChemBERTa-100M-MLM with 12 layers and residual width 768, adapted with LoRA on query and value projections ($r = 8$, $\alpha = 16$). The protein predictors use the ESM-2 t12, t30, and t33 variants, with 35M, 150M, and 650M parameters and residual widths 480, 640, and 1280. The reported hardware is one RTX 4090, with adaptation batch size 8 for the 650M model.

SAE training uses mean-pooled representations. The default TopK dictionary has eight times the residual width and 32 active latents per input. Table 1 lists the chemistry alternatives. Decoder-span removal is applied through a residual-stream hook; individual attention heads are ablated by zeroing their context slice through a pre-hook on the attention output projection. These operations evaluate representation-level and component-level sensitivity, respectively.

The datasets are TDC hERG and AMES for molecules (Therapeutics Data Commons, n.d.) and a UniProt DNA-binding task for proteins. KW-0238 defines DNA-binding positives. The molecular annotation is the conjunction of a basic nitrogen and an aromatic system; the task-conditioned protein annotation is median-thresholded basic-residue composition. The additional pretrained-model analysis uses aromatic-plus-basic composition.

The chemistry analysis uses a "Karim split" (Karim et al., 2021), and the protein task uses a cached UniProt dataset with KW-0238 positives. The experimental summaries report reconstruction R^2 of 0.90–0.98 overall; the nine chemistry configurations in Table 1 span 0.93–0.98.

B CONNECTIONS TO RELATED WORK

Table 5: Related questions and their connection to this study.

Question	Connection
Splitting and absorption	Hierarchical concepts can span or disappear among dictionary entries (Chanin et al., 2025; Bussmann et al., 2025); here annotations are fixed by molecular or sequence rules.
Dictionary stability	Canonicity and seed variation limit a unique-feature interpretation (Leask et al., 2025; Paulo & Belrose, 2026); group recovery is an empirical property of the fitted SAE.
Probe usefulness	Strong alternative baselines matter (Kantamneni et al., 2025); we separate full-layer recovery from decoder-removal sensitivity.
Scientific model analysis	Molecular-transformer analysis (Varadi et al., 2025) and hERG attribution (Lee & Yoo, 2025) motivate task-specific tests; our protein series adds a descriptive scale comparison.

Table 5 connects the protocol to four areas of interpretability research. The common annotation links the recovery and removal measurements, while the scientific case studies provide concrete targets for evaluating coverage.

C CHEMISTRY DICTIONARY COMPARISONS

The nine configurations in Table 1 vary architecture, dictionary size, and activation sparsity. Reconstruction R^2 ranges from 0.93 to 0.98, and best-feature alignment ranges from 0.27 to 0.41. Within the $8\times$ TopK dictionary, increasing k from 16 to 64 improves alignment by 0.14. At $k = 32$, expanding the dictionary from $8\times$ to $16\times$ changes alignment from 0.39 to 0.30 with reconstruction fixed at 0.98. The target-based comparison therefore distinguishes configurations with similar reconstruction fidelity.

Single-feature precision and recall. The best-aligned chemistry feature has reported precision of approximately 0.87 and recall of approximately 0.60 for the basic-nitrogen-plus-aromatic indicator. Precision describes the proportion of feature-positive examples carrying the annotation; recall describes the proportion of annotated examples captured by that feature. At the reported operating point, the feature selects a comparatively concentrated subset of positive examples while covering about three fifths of the annotated set.

Additional hERG probe assessment. An initial hERG probe assessment reports AUROC 0.82. The full-layer reference used in the main feature-budget comparison is 0.853, alongside 0.740 for the single-feature probe and 0.842 for top-64.

D PROTEIN TARGET COMPARISONS

Table 2 evaluates basic-residue composition in the task-conditioned DNA-binding predictors. Table 3 evaluates aromatic-plus-basic composition in pretrained ESM-2 representations. Its 35M rows compare two organisms at layer 6 and two layers for *E. coli*. Their top-64 AUROCs span 0.811–0.818, and their reference gaps span 0.035–0.040. The larger-model settings retain positive group gains with wider reference gaps. These results show how the reference gap varies across the evaluated protein targets and representations.

E INTERVENTION BUDGETS AND EFFECTS

The chemistry intervention selects 64 columns at width 768 (8.3%). The protein interventions select $32/480$ (6.7%), $32/640$ (5.0%), and $64/1280$ (5.0%). Table 4 gives the targeted and random effects. Their differences are 0.059, 0.041, 0.012, and 0.002; their ratios are 8.38, 4.73, 2.20, and 1.50. The main scale comparison uses the protein series. These fractions describe selected-column budgets; numerical span rank is the dimension of the projection in Eq. 3.

Population-level intervention. A complementary population-level analysis reports an effect of 0.069 against an on-manifold null of 0.008. For decoder-span removal, the targeted and random effects are 0.067 and 0.008, respectively (Table 4).

F TOY RECOVERY GRID

Table 6 gives the single-feature recovery grid plotted in Figure 4. Density $\eta = k/d$ measures active concepts per dimension; spread g counts sub-directions per concept. Each column holds spread fixed while increasing density. The score decreases down every column, with the largest losses occurring between the first few density settings.

The fixed-density comparisons in Section 5 summarize variation across columns of this grid.

Table 6: Single-feature recovery over density η (rows) and spread g (columns) in the toy experiment. Larger values indicate higher recovery on the reported score.

$\eta \backslash g$	1	2	4	8
0.25	0.84	0.82	0.82	0.82
0.5	0.71	0.70	0.70	0.66
1.0	0.62	0.60	0.59	0.58
2.0	0.56	0.56	0.55	0.56
4.0	0.54	0.54	0.54	0.53
8.0	0.53	0.53	0.52	0.53

G AUXILIARY COMPONENT AND STABILITY ANALYSES

The attention-head analysis finds maximum effects of 0.009 for hERG and 0.005 for AMES, with dominant hERG MLP contributions at layers 1–2. The accompanying circuit summary reports 1024 of 6144 neurons at 98% faithfulness and attribution approximately $104\times$ a random-k baseline. The accompanying neuron-level readout is $R^2 = 0.047$. These component and circuit summaries describe behavioral localization and neuron-level recovery alongside the SAE property-decoding analysis.

An auxiliary subset analysis reports approximately 12% of calls in a low-deployment subset, with bootstrap Jaccard overlap 0.15. Jaccard overlap measures intersection size divided by union size, so the reported value describes limited overlap between resampled memberships. The low overlap indicates that subset membership varies across bootstrap samples. Shortcut learning concerns decision rules that work on standard benchmarks but fail under changed testing conditions (Geirhos et al., 2020). Here, the bootstrap result motivates checking membership stability before interpreting subgroup differences.