WHICH FEATURES MATTER? SPARSE LATENT INTERVENTION FOR PREDICTION-PRESERVING EXPLANATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks achieve strong predictive performance across a broad range of domains, yet their internal representations remain largely opaque: for a given prediction, it is generally unclear which internal features the model relies on, and whether they correspond to meaningful structure in the underlying domain. Sparse autoencoders (SAEs) decompose hidden representations into interpretable features, but reveal only what a layer represents, not which features a prediction actually depends on, while intervention-based methods typically assess features one at a time and overlook their interactions. We present Sparse Latent Intervention (SLI), a mechanistic interpretability framework that identifies, for a frozen model and a single input, a sparse set of SAE features that closely reproduces the model's prediction. SLI casts this inverse problem as a sparse selection problem: stochastic gates over the entire feature dictionary are jointly optimised under an expected-ℓ_0 penalty to preserve the prediction, while the counterfactual representation induced by the selected features is obtained in closed form through the frozen decoder, making the procedure end-to-end differentiable. We demonstrate SLI on a Transformer-based transcriptomic foundation model for two clinical prediction tasks, treatment response in inflammatory bowel disease and synovial endotype in rheumatoid arthritis, where curated domain knowledge enables independent validation, and compare it against established attribution methods. On both tasks, SLI recovers compact feature sets that closely reproduce the model's predictions, recur across patients, and align with domain-relevant structure.

1 INTRODUCTION

Deep neural networks now achieve strong predictive performance across a broad range of domains. Their internal representations, however, remain largely opaque: for a given prediction, it is generally unclear which of the many directions encoded in a hidden layer the model actually relies on, and whether they correspond to structure that is meaningful to a domain expert. This opacity limits the trust that can be placed in such models in high-stakes settings (Doshi-Velez & Kim, 2017; Rudin, 2019), and prevents their use as instruments for scientific discovery, where the goal is not only to predict but also to understand what the model has learned.

Two broad families of methods address this problem, with complementary shortcomings. *Post-hoc feature attribution* methods, such as saliency maps, Integrated Gradients, DeepLIFT and SHAP (Simonyan et al., 2014; Sundararajan et al., 2017; Shrikumar et al., 2017; Lundberg & Lee, 2017), assign an importance score to each *input* feature. They produce graded scores and require little access to the model, but they explain predictions in terms of raw input units rather than the abstractions the model forms internally, and their faithfulness has repeatedly been called into question (Hooker et al., 2019; Adebayo et al., 2018). *Mechanistic interpretability*, in contrast, opens the model and seeks human-interpretable units of computation within its hidden layers (Sharkey et al., 2025). Because networks represent more features than they have dimensions (Olah et al., 2020; Elhage et al., 2022), individual neurons are typically polysemantic; Sparse Autoencoders (SAEs) resolve this by decomposing activations into an overcomplete dictionary of sparsely active and largely monosemantic features (Bricken et al., 2023; Huben et al., 2024; Templeton et al., 2024; Gao et al., 2025). An SAE, however, only describes *what* a layer represents: that a feature is active on an input does not

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.

imply that the prediction depends on it. Establishing causal relevance requires intervening on the representation, and existing intervention-based approaches either ablate or patch components individually (Vig et al., 2020; Meng et al., 2022; Heimersheim & Nanda, 2024) or linearly approximate per-feature effects and then threshold them (Marks et al., 2025). Such one-at-a-time estimates overlook redundancy and interactions among features, and do not directly answer the question of *which set* of features, taken together, yields a given prediction.

We address this question by introducing Sparse Latent Intervention (SLI), a mechanistic interpretability framework that, for an arbitrary frozen model and a query input, identifies a sparse set of SAE features at a chosen layer that closely reproduces the model's prediction. SLI casts this inverse problem as a sparse selection problem: stochastic gates over the SAE dictionary are optimised under a prediction-fidelity objective and an expected-ℓ_0 sparsity penalty, while the counterfactual hidden representation induced by the selected features is obtained in closed form through the frozen SAE decoder. Because all features are selected jointly, redundant features tend to be pruned and complementary ones retained, and the resulting explanation is expressed in the model's own learned representation rather than in input space. SLI requires only the token-level embeddings of one layer and the differentiable forward map of the subsequent layers. It makes no assumption on the architecture, input modality or output type, and therefore applies to any model on which an SAE can be trained.

Assessing whether an interpretability method recovers meaningful features requires settings in which independent domain knowledge is available to validate them. We therefore demonstrate SLI on a scientific foundation model whose application domain provides such curated references: EVA-RNA (Bandasack et al., 2026), a transcriptomic foundation model applied to predicting treatment response in inflammatory bowel disease and synovial endotype in rheumatoid arthritis.

Our contributions are as follows:

- We formulate the identification of a sparse set of SAE features that closely reproduces a frozen model's prediction at a given layer as an ℓ_0-penalised selection problem over stochastic gates, with a counterfactual embedding computed in closed form that makes it end-to-end differentiable.
- We apply SLI to a frozen Transformer-based transcriptomic foundation model on two clinical prediction tasks, and show that it recovers compact feature sets that closely reproduce the model's predictions and overlap with domain-relevant structure established in the literature.
- We benchmark SLI against established attribution methods, namely SHAP (Lundberg & Lee, 2017) and Integrated Gradients (Sundararajan et al., 2017), and against feature sets of the same size chosen by activation strength or at random.

2 Related Work

Feature attribution and masking. Post-hoc attribution methods explain a prediction by scoring the features of the input. Gradient-based methods compute saliency from input gradients (Simonyan et al., 2014), integrate gradients along a path from a baseline (Sundararajan et al., 2017), or propagate activation differences with respect to a reference (Shrikumar et al., 2017); perturbation-based methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) fit local surrogate models or estimate Shapley values from model evaluations on perturbed inputs. Closest in spirit to SLI are mask-based explanations, which optimise a continuous mask over the input under a sparsity or area constraint, so that the preserved region suffices to maintain the prediction or its removal suffices to destroy it (Fong & Vedaldi, 2017; Dabkowski & Gal, 2017; Fong et al., 2019), and instance-wise feature selection, which learns to select a small informative subset of input features (Chen et al., 2018). De Cao et al. (2020) extended differentiable masking to hidden states, identifying which token representations at each layer can be masked out without changing the output. All of these methods explain predictions in terms of input units or token positions. Concept-based methods such as TCAV (Kim et al., 2018) operate at a higher level of abstraction, but require the concepts to be specified in advance. SLI retains the mask-optimisation principle but applies it to an unsupervised SAE dictionary rather than to the input: explanations are expressed in concepts learned from the model's own representations, the counterfactual is decoded from the retained features instead of

filled in with a blur or a baseline, and whether the retained set reproduces the prediction is verified directly by the model's forward pass.

Sparse autoencoders and mechanistic interpretability. Building on sparse coding (Olshausen & Field, 1996) and k-sparse autoencoders (Makhzani & Frey, 2014), SAEs learn overcomplete dictionaries of sparsely active features that are substantially more interpretable than individual neurons (Bricken et al., 2023; Huben et al., 2024). They have since been scaled to large language models (Templeton et al., 2024; Lieberum et al., 2024), with architectural variants such as TopK (Gao et al., 2025) and JumpReLU (Rajamanoharan et al., 2025) SAEs improving the trade-off between sparsity and reconstruction. SAEs have also been applied beyond language, notably to protein language models, where individual features align with known structural and functional annotations (Simon & Zou, 2025; Adams et al., 2025). A complementary line of work localises the internal components that causally mediate a behaviour by intervening on activations, through causal mediation analysis (Vig et al., 2020), activation patching (Meng et al., 2022; Heimersheim & Nanda, 2024) and circuit discovery (Wang et al., 2023; Conmy et al., 2023). Marks et al. (2025) bring the two lines together by discovering circuits over SAE features, estimating the indirect effect of each feature through linear approximations and retaining those above a threshold. SLI instead optimises jointly over the entire dictionary, and verifies that the retained set closely reproduces the prediction through the model's own forward pass rather than a linear approximation.

Sparse selection with stochastic gates. Selecting a subset of features that optimises an objective is a combinatorial problem. Louizos et al. (2018) made ℓ_0-penalised selection amenable to gradient-based optimisation with hard-concrete gates, a stretched and clipped version of the binary concrete relaxation (Maddison et al., 2017), originally to sparsify the weights of a network; the same gates have been used to learn masks over hidden states that preserve a model's output (De Cao et al., 2020). SLI applies them to the features of an SAE dictionary: the gates decide which interpretable features are retained in the counterfactual representation, and the expected-ℓ_0 penalty keeps the retained set small.

Scientific foundation models as testbeds. Transformer-based foundation models have been developed for a growing number of biological modalities, including transcriptomics (Theodoris et al., 2023; Cui et al., 2024; Hao et al., 2024; Bandasack et al., 2026) and epigenomics (Ying et al., 2024; de Lima Camillo et al., 2026). Their interpretation has so far relied mostly on attention analysis and input attribution (de Lima Camillo et al., 2022; Ying et al., 2024; Novakovsky et al., 2023). These domains are particularly well suited to evaluating interpretability methods because they provide external, curated biological references against which recovered features can be compared (Claye et al., 2025; Kendiukhov, 2026).

3 Background

Notation. We consider a frozen model $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$ composed of L layers, $\mathcal{F} = \mathcal{F}^{(L)} \circ \cdots \circ \mathcal{F}^{(1)}$, whose input is a sequence of N tokens, possibly extended with special tokens (e.g. a classification token) to $T \geq N$ tokens. At a layer ℓ, we write $\mathbf{H} = \mathbf{H}^{(\ell)}(\mathbf{x}) \in \mathbb{R}^{T \times d_\ell}$ for the *embedding* of an input $\mathbf{x}$, whose i-th row $\mathbf{h}_i$ is a *token-level embedding*, and $\mathcal{F}_{>\ell} = \mathcal{F}^{(L)} \circ \cdots \circ \mathcal{F}^{(\ell+1)}$ for the *tail model*, so that $\mathcal{F}(\mathbf{x}) = \mathcal{F}_{>\ell}(\mathbf{H})$.

Sparse autoencoders. An SAE maps a token-level embedding $\mathbf{h}$ to a sparse code $\mathbf{a} = \varphi(\mathbf{h}) \in \mathbb{R}^K$ over an overcomplete dictionary ($K \gg d_\ell$) and reconstructs it as $\psi(\mathbf{a}) = \mathbf{W}_\psi\, \mathbf{a} + \mathbf{b}_{\mathrm{pre}} \approx \mathbf{h}$; the k-th column $\mathbf{w}_k$ of $\mathbf{W}_\psi$ is the direction of feature k. We use the TopK SAE (Gao et al., 2025), whose encoder $\varphi(\mathbf{h}) = \mathrm{TopK}_k\big(\mathrm{ReLU}(\mathbf{W}_\varphi(\mathbf{h} - \mathbf{b}_{\mathrm{pre}}) + \mathbf{b}_\varphi)\big)$, which keeps the k largest pre-activations of each token, so sparsity is imposed by construction. The SAE is trained once per model and layer by minimising the mean squared reconstruction error on token-level embeddings from a corpus of inputs, and then frozen.

Hard-concrete gates. A binary mask $\mathbf{z} \in \{0,1\}^K$ penalised by $\|\mathbf{z}\|_0$ admits no useful gradient. Louizos et al. (2018) replace each entry by a *hard-concrete gate*, a stretched and clipped binary concrete variable (Maddison et al., 2017): with a learnable logit $\log \alpha_k$, a temperature $\beta > 0$ and

$\gamma < 0 < 1 < \zeta$,

$$u \sim \mathcal{U}(0,1), \quad s_k = \sigma\Big(\tfrac{\log u - \log(1-u) + \log \alpha_k}{\beta}\Big), \quad z_k = \mathrm{clip}\big(s_k(\zeta - \gamma) + \gamma,\, 0,\, 1\big). \tag{1}$$

Because of the stretch, z_k equals exactly 0 or 1 with non-zero probability, and the expected number of non-zero gates,

$$\mathbb{E}\,\|\mathbf{z}\|_0 = \textstyle\sum_{k=1}^{K} \sigma(\log \alpha_k - \beta \log(-\gamma/\zeta))\,, \tag{2}$$

is a differentiable surrogate for the ℓ_0 count.

4 Method

4.1 Overview

Given a frozen model $\mathcal{F}$, an input $\mathbf{x}$ and a layer ℓ, SLI seeks a sparse set of SAE features at layer ℓ that closely reproduces the prediction $\mathcal{F}(\mathbf{x})$. Each feature is scaled by a hard-concrete gate; the gated features are decoded into a counterfactual embedding (Section 4.2), and the gate logits are optimised so that the tail model reproduces its prediction on the input with as few gates open as possible (Section 4.3; Fig. 1).

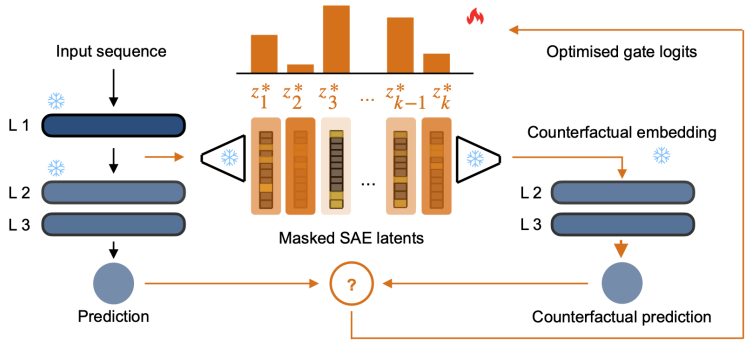

Figure 1: **Sparse Latent Intervention (SLI).** Each SAE latent of the layer-ℓ embedding is scaled by a hard-concrete gate z_k and decoded into a counterfactual embedding. Snowflakes denote frozen components; only the gate logits are optimised.

4.2 Intervention in Feature Space

The SAE code of the input, $\mathbf{A} = \varphi(\mathbf{H}) \in \mathbb{R}^{T \times K}$, is computed once and held fixed; its k-th column holds the activations of feature k across tokens. A gate vector $\mathbf{z} \in [0,1]^K$, shared across token positions, masks it as $\mathbf{A}_\mathbf{z} = \mathbf{A}\,\mathrm{diag}(\mathbf{z})$, so each feature is retained (z_k=1) or removed (z_k=0) for the input as a whole. The counterfactual embedding is the decoded masked code, $\tilde{\mathbf{H}}_\mathbf{z} = \psi(\mathbf{A}_\mathbf{z})$, an affine and hence differentiable function of $\mathbf{z}$. At $\mathbf{z} = \mathbf{1}$ it is the SAE reconstruction, from which the optimisation starts; its prediction error is our criterion for selecting the SAE hyper-parameters (Appendix A.2). At $\mathbf{z} = \mathbf{0}$ every intervened token collapses to $\mathbf{b}_{\mathrm{pre}}$. In between, the tail model's output is the prediction that the retained features induce on their own.

4.3 Sparse Feature Selection

SLI optimises the gate logits,

$$\log \boldsymbol{\alpha}^* = \arg\min_{\log \boldsymbol{\alpha}} \; \mathbb{E}_{\mathbf{z} \sim q(\cdot|\boldsymbol{\alpha})} \Big[\underbrace{d\big(\mathcal{F}_{>\ell}(\tilde{\mathbf{H}}_\mathbf{z}),\, \mathcal{F}_{>\ell}(\mathbf{H})\big)}_{\text{prediction fidelity}} \Big] + \lambda_z \underbrace{\mathbb{E}\,\|\mathbf{z}\|_0}_{\text{feature sparsity}}\,, \tag{3}$$

where q is the hard-concrete law equation 1, d a discrepancy on $\mathcal{Y}$ (e.g. the squared error) and $\lambda_z > 0$ sets the trade-off between fidelity and sparsity. We minimise equation 3 with Adam, starting with every gate open and drawing one gate sample per step; since each gate is a differentiable function of

its logit and of independent noise, back-propagating through the frozen tail model and SAE decoder gives an unbiased gradient estimate. After convergence, the gates are read out deterministically as $z_k^* = \mathrm{clip}(\sigma(\log \alpha_k)(\zeta - \gamma) + \gamma, 0, 1)$ and $\mathcal{K}^* = \{k : z_k^* > 0\}$. All reported fidelities are *binary*: they set z_k=1 on $\mathcal{K}^*$ and 0 elsewhere, so the retained features keep the activations they take in $\mathbf{x}$. As the tail model is non-linear, equation 3 is non-convex, and $\mathcal{K}^*$ is not guaranteed to be the smallest set that reproduces the prediction; its size, and the discrepancy between the prediction it induces and the original one, are governed by λ_z.

5 Experiments

5.1 SLI Explains Clinical Predictions of a Transcriptomic Foundation Model

EVA-RNA (Bandasack et al., 2026) is a Transformer encoder over gene tokens: a bulk RNA-seq profile is given to the model as a sequence of one token per gene, each combining the gene's identity with its expression level, and the sample is summarised by the 256-dimensional output of a `<cls>` token prepended to that sequence. Because its tokens are genes, every SAE feature can be named by the genes whose tokens activate it most strongly, so the selected set of features can be directly verified against the immunology of a cohort. We apply SLI to two tasks of the PRIMO benchmark (Kanbi et al., 2026): `t001`, endoscopic remission under adalimumab in inflammatory bowel disease (63 patients, binary outcome), and `t014`, rheumatoid arthritis synovial endotype (251 patients, three classes—fibroid, lymphoid and myeloid).

The predictor $\mathcal{F}$ that SLI explains is EVA-RNA's encoder followed by PRIMO's own probe, an ℓ_2-regularised logistic regression on the standardised `<cls>` token embedding. The probe is cross-fit over five stratified folds of each cohort, and each query is explained through the probe of the fold from which it was held out (Appendix A.1). On the public PRIMO leaderboard, EVA-RNA reaches an AUROC of 0.744 ± 0.031 on `t001` and 0.844 ± 0.011 on `t014`.

We intervene at layer ℓ=15 of the 24-layer encoder (d_ℓ=256) on the SAE features of the gene embeddings only. SLI would allow intervention on the `<cls>` embedding as well, but we keep it at its original value: the SAE is not trained on it, as its statistics here differ from those of the gene embeddings, and since the probe reads it at the last layer, reconstructing it poorly would shift the prediction by that error alone. The `<cls>` token attends to the intervened gene embeddings in the subsequent layers, so, given the information that `<cls>` already carries at layer ℓ, every change in the prediction comes from the features retained or removed from the gene embeddings.

Each cohort is split once, stratified, into a fitting pool (70%) on whose tokens the SAE is trained and an explanation pool (30%) from which queries are drawn; EVA-RNA is self-supervised and has seen neither set of labels. The SAE is a TopK autoencoder with K=2048 and k=64, selected on both tasks by cross-validation (Appendix A.2). We solve equation 3 with 2000 Adam steps and set λ_z=0.1 on `t001` and λ_z=0.03 on `t014`, chosen per task from the fidelity–sparsity trade-off (Appendix A.3). Every patient of the two explanation pools is explained, 19 and 76 patients respectively.

Small sets reproduce the prediction. On `t001`, 12.5 latents on average (0.6% of the dictionary) reproduce the probe's rescaled logit to a mean absolute error $\overline{|\Delta|} = 0.092$ and a worst case $|\Delta| = 0.196$ (Fig. 2a). The selected set gives the same predicted class as the full model on all 19 patients. On `t014`, 6.4 latents on average (0.3% of the dictionary) reproduce the three-way logit vector to a mean absolute error $\overline{|\Delta|} = 0.134$ per component (Fig. 2b), with the predicted class preserved on 72 of 76 patients. Keeping all 2048 latents ($\mathbf{z} = \mathbf{1}$) already shifts the prediction by $|\Delta| = 0.160 \pm 0.017$ on `t001` and $|\Delta| = 0.025 \pm 0.001$ on `t014`, the error of the SAE reconstruction itself (Appendix A.2). On `t001`, the dozen selected latents therefore reproduce the prediction more closely than the full reconstruction; on `t014`, their error is about five times this floor. The adopted λ_z values do not sit at abrupt points of the trade-off: across the sweep, the error decreases steadily as the support grows (Fig. 2c), so neighbouring values of λ_z give similar explanations.

On 13 of the 76 `t014` queries, SLI returns the empty set. For those patients the feature-free counterfactual, in which every gene token is collapsed to $\mathbf{b}_{\mathrm{pre}}$ and only the `cls` embedding is unchanged, already reproduces the prediction with an error $\overline{|\Delta|} = 0.124$ on average. Retaining any feature

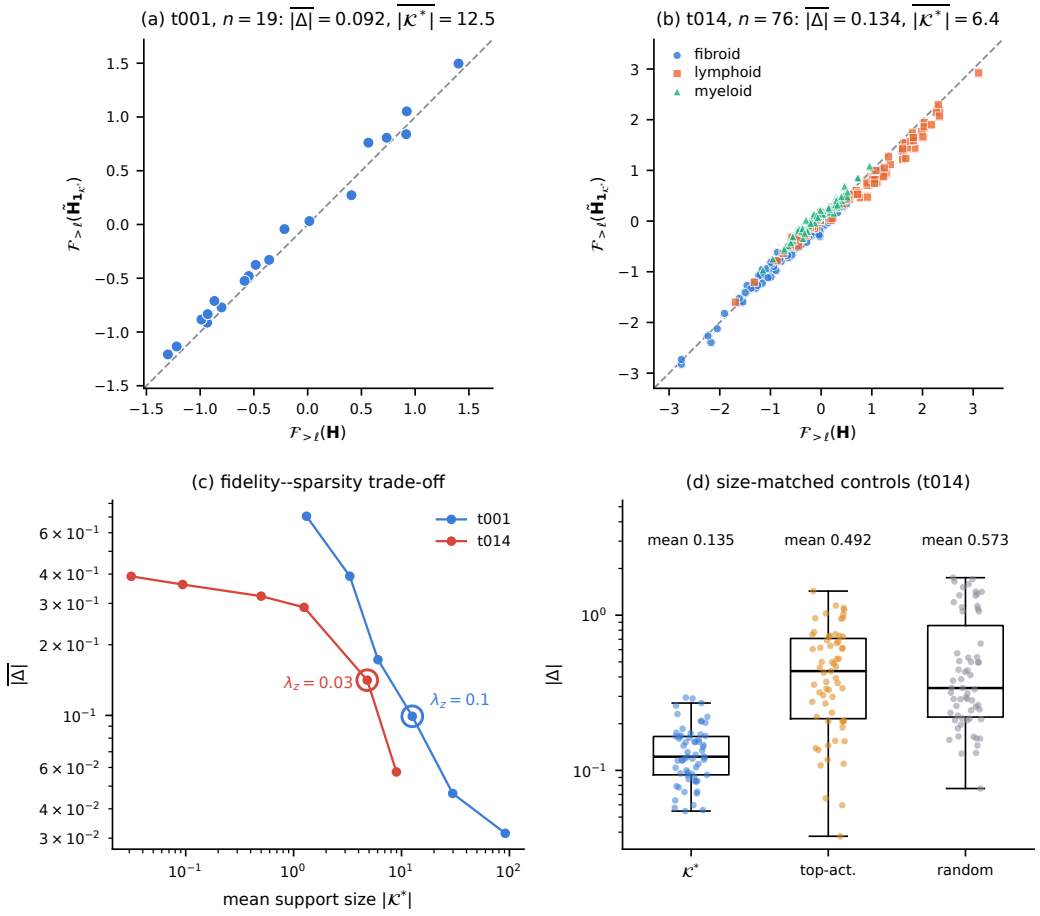

Figure 2: **SLI on EVA-RNA (layer 15, TopK SAE with K=2048, k=64). (a)** `t001`: the logit predicted from the selected SAE features against the model's original logit, over all 19 explanation-pool patients. **(b)** `t014`: the same over all 76 patients, one point per class logit. **(c)** Fidelity–sparsity trade-off traced by sweeping $\lambda_z \in \{0.01, 0.03, 0.1, 0.3, 1, 3\}$ (both axes logarithmic); circles mark the adopted λ_z. The sweeps are separate runs, on 19 (`t001`) and 32 (`t014`) queries. **(d)** Size-matched controls on the 63 `t014` queries with a non-empty support: $\mathcal{K}^*$ against the $|\mathcal{K}^*|$ most active latents and against 10 random draws of $|\mathcal{K}^*|$ latents per query, all evaluated at the binary read-out (Section 5.2). Boxes show the median and quartiles.

would add λ_z to the objective equation 3 while reducing its discrepancy term by less than that amount, so the optimum retains none. On the other 63, removing all features costs $\overline{|\Delta|} = 0.577$ on average, and the selected set brings this down to $\overline{|\Delta|} = 0.134$, recovering a median 61% of each patient's gap. On `t001` no selected feature set is empty: removing all features costs $\overline{|\Delta|} = 1.12$ on average, and the selected set recovers a median 92% of this gap for most patients. The two exceptions are patients whose prediction the feature-free counterfactual already reproduces well ($|\Delta| = 0.010$ and $|\Delta| = 0.126$); there selecting features sets results in slightly worse performance ($|\Delta| = 0.084$ and $|\Delta| = 0.156$).

A shared core of mucosal programmes on `t001`. We name each latent by the three genes on whose tokens it has the highest mean activation over the queries that selected it. Across the 19 patients, 69 distinct latents appear in at least one explanation, and the supports are not disjoint: two latents are selected on every patient, five on at least half, and 28 on exactly one (Fig. 3a). Of the seven latents selected on more than a third of the cohort, five correspond to known transcriptional programmes of intestinal mucosa: inflammation and antimicrobial defence (643: *MMP12*,

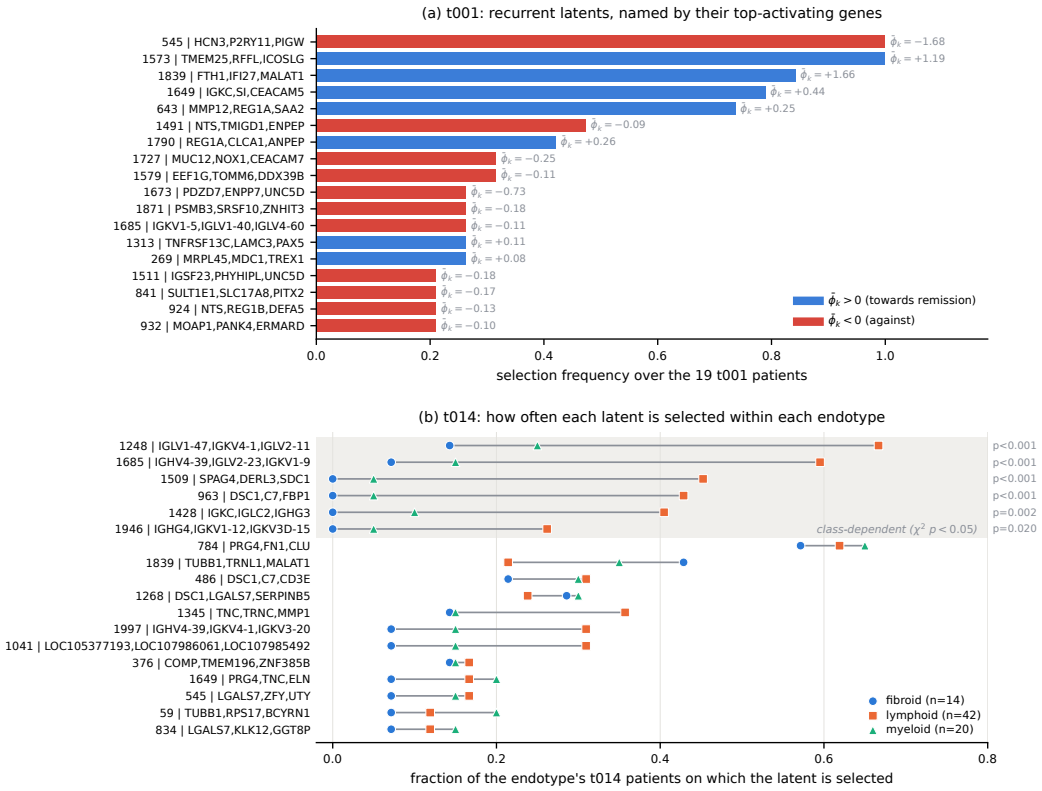

Figure 3: **Gene composition of selected latents. (a)** `t001`: the 18 latents selected on at least four of the 19 patients, each named by the three genes on whose tokens it has the highest mean activation over the queries that selected it. Bars are coloured by the sign of the mean Shapley value $\bar{\phi}_k$ of the latent under the same intervention (Section 5.2), annotated to their right. **(b)** `t014`: the 18 latents that appear in the most selected sets across the 76 patients. Each marker gives the fraction of an endotype's patients whose selected set contains the latent; the grey segment spans the three fractions. The shaded rows are the latents whose selection frequency differs across endotypes (χ^2, p-value on the right); labels as in (a).

REG1A, *SAA2*), the secretory and absorptive epithelium (1790: *REG1A*, *CLCA1*, *ANPEP*; 1491: *NTS*, *TMIGD1*, *ENPEP*; 1649: *IGKC*, *SI*, *CEACAM5*, which pairs a plasma-cell transcript with epithelial ones) and the interferon response (1839: *FTH1*, *IFI27*, *MALAT1*). These programmes overlap with the inflammatory and plasma-cell axes along which mucosal signatures of anti-TNF response have been described (Arijs et al., 2009; Martin et al., 2019). The remaining two (545: *HCN3*, *P2RY11*, *PIGW*; 1573: *TMEM25*, *RFFL*, *ICOSLG*) are selected on every patient and carry two of the three largest mean Shapley values, of opposite sign, yet their genes share no function. We interpret them not as biological programmes but as compensating for a shift common to all patients, introduced when every feature is removed.

Class-specific latents on `t014`. Of the 67 latents selected at least once, seven have a selection frequency that differs across endotypes at χ^2 $p < 0.05$, and six of the seven are enriched in the lymphoid patients (Fig. 3b). Five of these six name B-cell and plasma-cell products: three immunoglobulin variable regions (1248: *IGLV1-47*, *IGKV4-1*, *IGLV2-11*; 1685: *IGHV4-39*, *IGLV2-23*, *IGKV1-9*; 1946: *IGHG4*, *IGKV1-12*, *IGKV3D-15*), immunoglobulin constant regions (1428: *IGKC*, *IGLC2*, *IGHG3*) and the plasma-cell secretory programme (1509: *SPAG4*, *DERL3*, *SDC1*). The sixth (963: *DSC1*, *C7*, *FBP1*) names no single cell population. The six are selected on 26–67% of the lymphoid patients and on 0–14% of the fibroid patients. The lymphoid pathotype is defined histologically by B-cell and plasma-cell aggregates in the synovium (Lewis et al., 2019), the cell populations these latents name. The latent 784 (*PRG4*, *FN1*, *CLU*), naming synovial lining fibroblasts, is selected on 47

of 76 patients at indistinguishable rates in all three endotypes (χ^2 p=0.90). We therefore interpret it as tissue context that the probe relies on for every patient, rather than as a feature that distinguishes the endotypes. Per-class frequencies and tests are reported in Appendix A.5.

5.2 Comparison with Baseline Attribution Methods

SLI returns a set of features; attribution methods return a ranking. We ask whether a set of the same size chosen otherwise reproduces the prediction as closely, and whether attribution methods rank SLI's latents highest. All methods use SLI's intervention—a latent is either kept at its observed activations or removed—so they differ only in how they choose or score latents. For a query with $m = |\mathcal{K}^*|$ we use:

- *top-activation*: the m latents with the largest activation mass $\sum_i A_{ik}$ over the gene tokens;
- *random*: 10 draws of m latents, uniformly among the 2048;
- *Shapley values* (Lundberg & Lee, 2017), with the value of a coalition $\mathcal{S}$ being the prediction $v(\mathcal{S}) = \mathcal{F}_{>\ell}\big(\tilde{\mathbf{H}}_{\mathbf{1}_\mathcal{S}}\big)$ when the latents in $\mathcal{S}$ are retained and all others removed;
- *Integrated Gradients* (Sundararajan et al., 2017) with respect to the gates, along the straight path from $\mathbf{z} = \mathbf{0}$ to $\mathbf{z} = \mathbf{1}$.

Shapley values and Integrated Gradients are computed over a candidate pool of the 64 most active latents of the query together with $\mathcal{K}^*$, on the queries with a non-empty support (19 on `t001`, 63 on `t014`); estimation details are in Appendix A.4. Both need a scalar output, and on `t014` we report them for the lymphoid logit, whereas SLI preserves all three; the other two logits give the same picture (Appendix A.4).

Fidelity at matched size. On the `t014` queries with a non-empty support, SLI leaves a mean error $\overline{|\Delta|} = 0.135$, against $\overline{|\Delta|} = 0.492$ for the top-activation set and $\overline{|\Delta|} = 0.573$ for random sets of the same size (mean over 10 draws per patient) (Fig. 2d). SLI's set has a lower error than the top-activation set for 59 of the 63 patients. Only 3 of the 630 random sets (10 per patient) have a lower error than SLI's, all for the same patient. The two criteria also disagree on which latents qualify at all: 19% (`t001`) and 26% (`t014`) of the latents SLI selects are not among the 64 most active latents of the query.

Table 1: **Agreement between SLI's selected sets and attribution rankings on EVA-RNA.** Latents in the candidate pool are ranked by $|\phi_k|$ or $|\mathrm{IG}_k|$, with SLI's selection as the positive class and $m = |\mathcal{K}^*|$. Means over queries, except the median rank of the selected latents, which is a median over queries. On `t014` both methods attribute the lymphoid logit.

Task	Method	mean m	precision@m	AUROC	median rank
`t001` (19 queries)	Shapley	12.5	0.67	0.91	8.5
	Integrated Gradients	12.5	0.62	0.86	8
`t014` (63 queries)	Shapley	7.8	0.41	0.87	7.5
	Integrated Gradients	7.8	0.38	0.82	9

Agreement with attribution rankings. The latents SLI selects have large attributions: their median $|\phi_k|$ is about five times that of the unselected latents in the pool (Appendix A.4, Fig. 5a), and ranking the pool by $|\phi_k|$ places selected latents above unselected ones with an AUROC of 0.91 on `t001` and 0.87 on `t014` (Table 1; 0.86 and 0.82 with Integrated Gradients). The rankings nevertheless do not reduce to SLI's sets: of the m latents ranked highest by $|\phi_k|$, SLI selects 67% on `t001` and 41% on `t014` (Table 1; 62% and 38% with Integrated Gradients). Shapley values and Integrated Gradients are in higher mutual agreement than either is with SLI (Appendix A.4), so the gap reflects what each method asks for rather than how it is estimated. An attribution ranking rewards every latent that moves the prediction, even when several carry the same information; because every retained latent is penalised, SLI tends to keep only one of them. Consistently, the high-attribution latents that SLI leaves out are correlated with the ones it keeps (Appendix A.4). On `t014`, agreement is lowered further because the attributions target one logit while SLI preserves all three.

Within the selected set, Shapley values grade the latents that SLI leaves ungraded: after convergence the gates are binary, and their logits reflect how confidently a latent was selected, not how much it moves the prediction. Unlike the gates, ϕ_k is signed, so it separates latents that push the prediction towards the model's output from those that offset them. The values $\bar{\phi}_k$ in Fig. 3(a) are averages of ϕ_k over the queries that selected latent k.

6 Discussion and Limitations

We here introduced SLI, a novel mechanistic interpretability framework which identifies, for a frozen model and a single input, a sparse set of SAE features that closely reproduces the model's prediction (Section 4). The features are selected jointly over the whole dictionary by hard-concrete gates under an expected-ℓ_0 penalty (Section 4.3), and the counterfactual embedding is decoded from the retained features in closed form, which makes the selection end-to-end differentiable (Section 4.2).

We applied SLI to EVA-RNA on two PRIMO tasks, endoscopic remission under adalimumab in inflammatory bowel disease and the synovial endotype in rheumatoid arthritis (Section 5.1). SLI closely reproduces the model's prediction with 12.5 and 6.4 latents on average, less than 1% of the dictionary, and preserves the predicted class for all 19 patients on the first task and for 72 of the 76 patients on the second (Fig. 2a,b). The selected latents recur across patients and correspond to known disease programmes: mucosal inflammation and epithelial programmes on `t001`, and B-cell and plasma-cell programmes selected almost exclusively for lymphoid patients on `t014` (Fig. 3). These features are extracted from the embeddings of a self-supervised encoder that has never seen the clinical labels. The comparison with baselines (Section 5.2) shows that SLI answers a different question from the usual ways of reading SAE features. Activation shows what a layer encodes, but the most active latents reproduce the prediction barely better than randomly chosen ones (Fig. 2d). Feature attribution methods score each latent separately and rank SLI's latents highly (Table 1), yet they assign a high score to every latent that carries the same information, whereas SLI, which penalises every retained latent, tends to keep only one of them.

SLI nonetheless has limitations. Because the objective equation 3 is non-convex and optimised through a relaxation, $\mathcal{K}^*$ is not guaranteed to be the smallest set that reproduces the prediction: its size is set by λ_z, and when latents are redundant it is one of several sets that reproduce the prediction equally well. The optimisation can also stop at a local optimum, as for two `t001` patients for whom the empty set gives a lower error than the returned set (Section 5.1). The counterfactual $\psi(\mathbf{A}_{\mathbf{z}})$ keeps only a subset of the features, and the layers after the intervened one never received such embeddings during training, so a change in the prediction may reflect the tail model reacting to an out-of-distribution input rather than the removed features being relevant (Hooker et al., 2019; Hase et al., 2021). The explanation is also bound to the dictionary: SLI selects only among the features the SAE has learned and does not model its reconstruction error, which shifts the prediction even when every feature is kept, so the layer and the SAE hyper-parameters must be validated for each model. Finally, our evaluation covers a single transcriptomic foundation model, at a single layer, on two tasks with modest explanation cohorts (19 and 76 patients), and SLI explains single predictions rather than circuits across layers; whether the selected sets remain compact and interpretable on larger models, dictionaries and cohorts is left to future work.

7 Conclusion

Explaining a prediction requires knowing not only which features a model represents, but which of them the prediction actually depends on. SLI answers this question by searching for a sparse set of SAE features that closely reproduces the prediction. Our results show that such sets are small, recur across patients and correspond to known disease programmes, and that at the same size they reproduce the prediction far more closely than the most active or the most attributed features. Asking *which* features jointly reproduce a prediction, rather than ranking features by individual scores, is thus a useful criterion for explaining predictions with the interpretable model's representations. The same formulation opens a path towards steering: replacing the model's own prediction with a designed target would let SLI search for a sparse set of interpretable features that moves the model towards that output.

AI USE STATEMENT

In this work, we used generative AI tools for help develop theoretical models or conceptual frameworks, design or provide feedback on research methodology or experiments, implement methods, interpret results. We have not used generative AI tools for generate synthetic data sets, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, and generate synthetic data sets, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, are not applicable to this work. Additionally, we used generative AI tools for create or modify scientific figures or images, create or edit software code, creation of artifacts, draft parts of a research paper, summarize or analyse existing literature, discover research topics or identify gaps, brainstorming, sourcing/searching for information, edit a research paper to improve readability, identify relevant literature, propose a title or keywords for a research paper. We have reviewed all AI-assisted work. All LLM-generated code used to implement the method and conduct the experiments was reviewed, tested, and verified for correctness by the authors. All results and their interpretations were verified by the authors against the original annotation resources and scientific literature. All figures were regenerated from our results and verified against the underlying data. All texts generated or modified by LLMs were checked and revised by the authors, and all authors read and approved the final manuscript. Every bibliographic reference suggested by an LLM was verified on the publisher's website to confirm its existence, accuracy, and relevance, and was cited based on the original source. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only previously published, de-identified human transcriptomic data. No new data were collected, and no experiments involving human participants or animals were conducted by the authors.

REPRODUCIBILITY STATEMENT

Model. We use the pretrained EVA-RNA encoder (Bandasack et al., 2026) without any fine-tuning. The weights, tokenizer and inference code are publicly available on Hugging Face (`http://huggingface.co/ScientaLab/eva-rna`, commit 2407734), under the Scienta Lab EVA Model License; access requires accepting its conditions. This is the checkpoint benchmarked as EVA 60M in PRIMO (Kanbi et al., 2026). Its gene-level embeddings, on which SLI operates at layer 15, are exposed by the public API.

Data. Both tasks are PRIMO tasks, whose expression inputs are publicly available on Hugging Face (`http://huggingface.co/datasets/PRIMOmics/primo`) as raw counts indexed by NCBI Gene IDs. Task t001, endoscopic remission under adalimumab in inflammatory bowel disease (63 patients), is built from mucosal biopsies deposited in ArrayExpress (E-MTAB-7914 and E-MTAB-7604, Verstockt et al. (2019)). Task t014, rheumatoid arthritis synovial pathotype (251 patients), is built from synovial biopsies of the PEAC (E-MTAB-6141; Lewis et al. (2019)) and R4RA (E-MTAB-11611; Rivellese et al. (2022)) cohorts.

REFERENCES

Etowah Adams, Liam Bai, Minji Lee, Yiyang Yu, and Mohammed AlQuraishi. From mechanistic interpretability to mechanistic biology: Training, evaluating, and interpreting sparse autoencoders on protein language models. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=zdOGBRQEbz`.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31.

Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf.

I Arijs, K Li, G Toedter, R Quintens, L Van Lommel, K Van Steen, P Leemans, G De Hertogh, K Lemaire, M Ferrante, F Schnitzler, L Thorrez, K Ma, X-Y R Song, C Marano, G Van Assche, S Vermeire, K Geboes, F Schuit, F Baribaud, and P Rutgeerts. Mucosal gene signatures to predict response to infliximab in patients with ulcerative colitis. *Gut*, 58(12):1612–1619, 2009. ISSN 0017-5749. doi: 10.1136/gut.2009.178665. URL https://gut.bmj.com/content/58/12/1612.

Ethan Bandasack, Vincent Bouget, Apolline Bruley, Yannis Cattan, Charlotte Claye, Matthew Corney, Julien Duquesne, Karim El Kanbi, Aziz Fouché, Pierre Marschall, and Francesco Strozzi. Eva: Towards a universal model of the immune system, 2026. URL https://arxiv.org/abs/2602.10168.

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

Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 883–892. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/chen18j.html.

Charlotte Claye, Pierre Marschall, Wassila Ouerdane, Céline Hudelot, and Julien Duquesne. Discovering interpretable biological concepts in single-cell rna-seq foundation models, 2025. URL https://arxiv.org/abs/2510.25807.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=89ia77nZ8u.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, August 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02201-0. URL https://doi.org/10.1038/s41592-024-02201-0.

Piotr Dabkowski and Yarin Gal. Real Time Image Saliency for Black Box Classifiers. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/0060ef47b12160b9198302ebdb144dcf-Paper.pdf.

Nicola De Cao, Michael Sejr Schlichtkrull, Wilker Aziz, and Ivan Titov. How do decisions emerge across layers in neural models? interpretation with differentiable masking. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3243–3255, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.262. URL https://aclanthology.org/2020.emnlp-main.262/.

Lucas Paulo de Lima Camillo, Louis R. Lapierre, and Ritambhara Singh. A pan-tissue DNA-methylation epigenetic clock based on deep learning. *npj Aging*, 8(1):4, April 2022. ISSN 2731-6068. doi: 10.1038/s41514-022-00085-y. URL https://doi.org/10.1038/s41514-022-00085-y.

Lucas Paulo de Lima Camillo, Raghav Sehgal, Jenel Armstrong, Max Melnikas, Henry E. Miller, Jun Ding, Luigi Ferruci, Jessica A. Lasky-Su, Albert T. Higgins-Chen, Steve Horvath, and Bo Wang. CpGPT: a Foundation Model for DNA Methylation. *bioRxiv*, 2026. doi: 10.1101/2024.10.24.619766. URL https://www.biorxiv.org/content/early/2026/08/12/2024.10.24.619766.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017. URL https://arxiv.org/abs/1702.08608.

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. https://transformer-circuits.pub/2022/toy_model/index.html.

Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3449–3457, 2017. doi: 10.1109/ICCV.2017.371.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tcsZt9ZNKD.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21(8):1481–1491, August 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02305-7. URL https://doi.org/10.1038/s41592-024-02305-7.

Peter Hase, Harry Xie, and Mohit Bansal. The Out-of-Distribution Problem in Explainability and Search Methods for Feature Importance Explanations. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 3650–3666. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/1def1713ebf17722cbe300cfc1c88558-Paper.pdf.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching, 2024. URL https://arxiv.org/abs/2404.15255.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A Benchmark for Interpretability Methods in Deep Neural Networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/fe4b8556000d0f0cae99daa5c5c5a410-Paper.pdf.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Karim El Kanbi, Yannis Cattan, Pierre Marschall, Matthew Corney, Aziz Fouché, Vincent Bouget, and Julien Duquesne. A transcriptomic benchmark for foundation models in immunology and inflammation drug development. In *Learning Meaningful Representations of Life (LMRL) Workshop at ICLR 2026*, 2026. URL https://openreview.net/forum?id=v2SA8gHwqo.

Ihor Kendiukhov. Systematic evaluation of single-cell foundation model interpretability reveals attention captures co-expression rather than unique regulatory signal, 2026. URL https://arxiv.org/abs/2602.17532.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), 2018. URL https://arxiv.org/abs/1711.11279.

Myles J. Lewis, Michael R. Barnes, Kevin Blighe, Katriona Goldmann, Sharmila Rana, Jason A. Hackney, Nandhini Ramamoorthi, Christopher R. John, David S. Watson, Sarah K. Kummerfeld, Rebecca Hands, Sudeh Riahi, Vidalba Rocher-Ros, Felice Rivellese, Frances Humby, Stephen Kelly, Michele Bombardieri, Nora Ng, Maria DiCicco, Désirée van der Heijde, Robert Landewé, Annette van der Helm-van Mil, Alberto Cauli, Iain B. McInnes, Christopher D. Buckley, Ernest Choy, Peter C. Taylor, Michael J. Townsend, and Costantino Pitzalis. Molecular Portraits of Early Rheumatoid Arthritis Identify Clinical and Treatment Response Phenotypes. *Cell Reports*, 28(9):2455–2470.e5, August 2019. ISSN 2211-1247. doi: 10.1016/j.celrep.2019.07.091. URL https://doi.org/10.1016/j.celrep.2019.07.091.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen (eds.), *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.19. URL https://aclanthology.org/2024.blackboxnlp-1.19/.

Christos Louizos, Max Welling, and Diederik P. Kingma. Learning sparse neural networks through L_0 regularization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1Y8hhg0b.

Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=S1jE5L5gl.

Alireza Makhzani and Brendan Frey. k-sparse autoencoders, 2014. URL https://arxiv.org/abs/1312.5663.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=I4e82CIDxv.

Jerome C. Martin, Christie Chang, Gilles Boschetti, Ryan Ungaro, Mamta Giri, John A. Grout, Kyle Gettler, Ling-shiang Chuang, Shikha Nayar, Alexander J. Greenstein, Marla Dubinsky, Laura Walker, Andrew Leader, Jay S. Fine, Charles E. Whitehurst, M Lamine Mbow, Subra Kugathasan, Lee A. Denson, Jeffrey S. Hyams, Joshua R. Friedman, Prerak T. Desai, Huaibin M. Ko, Ilaria Laface, Guray Akturk, Eric E. Schadt, Helene Salmon, Sacha Gnjatic, Adeeb H. Rahman, Miriam Merad, Judy H. Cho, and Ephraim Kenigsberg. Single-Cell Analysis of Crohn's Disease Lesions Identifies a Pathogenic Cellular Module Associated with Resistance to Anti-TNF Therapy. *Cell*, 178(6):1493–1508.e20, September 2019. ISSN 0092-8674. doi: 10.1016/j.cell.2019.08.008. URL https://doi.org/10.1016/j.cell.2019.08.008.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=-h6WAS6eE4.

Gherman Novakovsky, Nick Dexter, Maxwell W. Libbrecht, Wyeth W. Wasserman, and Sara Mostafavi. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics*, 24(2):125–137, February 2023. ISSN 1471-0064. doi: 10.1038/s41576-022-00532-2. URL https://doi.org/10.1038/s41576-022-00532-2.

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. https://distill.pub/2020/circuits/zoom-in.

Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, June 1996. ISSN 1476-4687. doi: 10.1038/381607a0. URL https://doi.org/10.1038/381607a0.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, Janos Kramar, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumpreLU sparse autoencoders, 2025. URL https://openreview.net/forum?id=mMPaQzgzAN.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

Felice Rivellese, Anna E. A. Surace, Katriona Goldmann, Elisabetta Sciacca, Cankut Çubuk, Giovanni Giorli, Christopher R. John, Alessandra Nerviani, Liliane Fossati-Jimack, Georgina Thorborn, Manzoor Ahmed, Edoardo Prediletto, Sarah E. Church, Briana M. Hudson, Sarah E. Warren, Paul M. McKeigue, Frances Humby, Michele Bombardieri, Michael R. Barnes, Myles J. Lewis, Costantino Pitzalis, Felice Rivellese, Giovanni Giorli, Alessandra Nerviani, Liliane Fossati-Jimack, Georgina Thorborn, Frances Humby, Michele Bombardieri, Myles J. Lewis, Patrick Durez, Maya H. Buch, Hasan Rizvi, Arti Mahto, Carlomaurizio Montecucco, Bernard Lauwerys, Nora Ng, Pauline Ho, Vasco C. Romão, João Eurico Cabral da Fonseca, Patrick Verschueren, Stephen Kelly, Pier Paolo Sainaghi, Nagui Gendi, Bhaskar Dasgupta, Alberto Cauli, Piero Reynolds, Juan D. Cañete, Julio Ramirez, Raquel Celis, Robert Moots, Peter C. Taylor, Chris J. Edwards, John Isaacs, Peter Sasieni, Ernest Choy, Charlotte Thompson, Serena Bugatti, Mattia Bellan, Mattia Congia, Christopher Holroyd, Arthur Pratt, Laura White, Louise Warren, Joanna Peel, Rebecca Hands, Gaye Hadfield, Costantino Pitzalis, and the R4RA collaborative group. Rituximab versus tocilizumab in rheumatoid arthritis: synovial biopsy-based biomarker analysis of the phase 4 R4RA randomized trial. *Nature Medicine*, 28(6):1256–1268, June 2022. ISSN 1546-170X. doi: 10.1038/s41591-022-01789-0. URL https://doi.org/10.1038/s41591-022-01789-0.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x. URL https://doi.org/10.1038/s42256-019-0048-x.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeffrey Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Isaac Bloom, Stella Biderman, Adrià Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Mary Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, William Saunders, Eric J Michaud, Stephen Casper, Max Tegmark, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Thomas McGrath. Open problems in mechanistic interpretability. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=91H76m9Z94. Survey Certification.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3145–3153. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/shrikumar17a.html.

Elana Simon and James Zou. InterPLM: discovering interpretable features in protein language models via sparse autoencoders. *Nature Methods*, 22(10):2107–2117, October 2025. ISSN 1548-7105. doi: 10.1038/s41592-025-02836-7. URL https://doi.org/10.1038/s41592-025-02836-7.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. URL https://arxiv.org/abs/1312.6034.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

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, 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.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, June 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06139-9. URL https://doi.org/10.1038/s41586-023-06139-9.

Bram Verstockt, Erwin Dreesen, Maja Noman, An Outtier, Nathalie Van den Berghe, Isolde Aerden, Griet Compernolle, Gert Van Assche, Ann Gils, Séverine Vermeire, and Marc Ferrante. Ustekinumab Exposure-outcome Analysis in Crohn's Disease Only in Part Explains Limited Endoscopic Remission Rates. *Journal of Crohn's and Colitis*, 13(7):864–872, July 2019. ISSN 1873-9946. doi: 10.1093/ecco-jcc/jjz008. URL https://doi.org/10.1093/ecco-jcc/jjz008.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Kevin Ro 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 *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Kejun Ying, Jinyeop Song, Haotian Cui, Yikun Zhang, Siyuan Li, Xingyu Chen, Hanna Liu, Alec Eames, Daniel L McCartney, Riccardo E. Marioni, Jesse R. Poganik, Mahdi Moqri, Bo Wang, and Vadim N. Gladyshev. MethylGPT: a foundation model for the DNA methylome. *bioRxiv*, 2024. doi: 10.1101/2024.10.30.621013. URL https://www.biorxiv.org/content/early/2024/11/04/2024.10.30.621013.

A Appendix

A.1 Experimental Details for EVA-RNA

Predictor. PRIMO's own probe (Kanbi et al., 2026) consists of standardisation of the last-layer <cls> embedding, followed by ℓ_2-regularised logistic regression, with the benchmark's hyperparameter search. It is cross-fit over five stratified folds of the whole cohort, and each query is explained through the probe of the fold from which it was held out. All five heads divide their logits by the same value: the standard deviation of all patients' logits, each computed by the probe not trained on that patient. This rescaling changes no decision, but places the fidelity term on a common scale across tasks, so that a given λ_z imposes a comparable sparsity–fidelity trade-off on both; all errors in Section 5 are in units of these rescaled logits.

Data. The intervention acts on the SAE features of each gene token (T=1196 for `t001`, T=1195 for `t014`); the `<cls>` embedding is kept at its original value. Each cohort is split once, stratified, into an SAE fitting pool and an explanation pool (70/30): the SAE is trained only on the tokens of the first, and queries are drawn only from the second. The pools contain 44 and 19 patients on `t001`, 175 and 76 on `t014`, and never overlap.

Optimisation. The gates use β=2/3 and (γ, ζ)=$(-0.1, 1.1)$ and are initialised at $\log \alpha_k$=3, so that every gate starts open. We run 2000 Adam steps at learning rate 10^{-1}, drawing one gate sample per step, with the squared error on the rescaled logits as discrepancy d between original and counterfactual prediction.

A.2 TopK SAE Selection for EVA-RNA

The TopK SAE has two free hyper-parameters: the expansion factor E, which sets the dictionary width $K = E \cdot d_\ell$ with d_ℓ=256, and the per-token budget k. They could be selected by the relative ℓ_1 reconstruction error on held-out token embeddings. This criterion, however, measures distances in embedding space and ignores the tail model: a dictionary may reconstruct the embeddings accurately in ℓ_1 yet displace them along a direction to which $\mathcal{F}_{>\ell}$ is sensitive, in which case every counterfactual carries a prediction error unrelated to the selected features. For EVA-RNA we therefore select (E, k) by the *splice error*, the prediction shift in the limiting case $\mathbf{z} = \mathbf{1}$ of the counterfactual (Section 4.2). In this case every feature is retained: the SAE reconstruction replaces the gene token embeddings at layer ℓ, the `<cls>` embedding is kept at its original value, and we measure

$$\varepsilon_{\text{splice}}(E, k) = \mathbb{E}_{\mathbf{x}}\Big[\, \tfrac{1}{C} \left\| \mathcal{F}_{>\ell}\big(\tilde{\mathbf{H}}_{\mathbf{1}}\big) - \mathcal{F}_{>\ell}(\mathbf{H}) \right\|_1 \Big], \tag{4}$$

where the expectation is over held-out samples and C is the number of output components (C=1 for `t001`, C=3 for `t014`). Since $\mathbf{z} = \mathbf{1}$ is the starting point of SLI, $\varepsilon_{\text{splice}}$ is the fidelity error already present before any feature is removed.

We sweep $E \in \{4, 8, 16, 32\}$ and $k \in \{16, 32, 64\}$ with five-fold cross-validation over the samples of the fitting pool. Each SAE is trained with Adam (learning rate 10^{-3}, batch size 256, 50 epochs); the decoder rows are projected onto the unit sphere after every update, and $\mathbf{b}_{\text{pre}}$ is initialised at the mean training embedding. Configurations (E, k) for which more than 10% of the dictionary's features never activate on the held-out tokens are discarded. Among the remaining ones, we take the smallest dictionary whose splice error is within one standard error of the lowest. This selects $(E, k) = (8, 64)$, i.e. K=2048, on both tasks (Fig. 4). For this configuration, $\varepsilon_{\text{splice}} = 0.160 \pm 0.017$ on `t001` and 0.025 ± 0.001 on `t014` (mean $\pm$ standard error over the five folds).

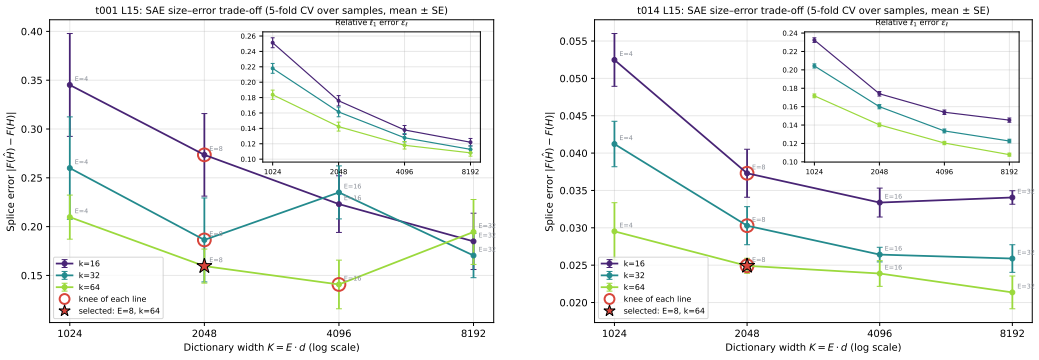

Figure 4: **TopK SAE size–error trade-off for EVA-RNA at layer** 15**.** Left: `t001` (44 fitting-pool samples). Right: `t014` (175). Each line fixes the per-token budget k and varies the expansion factor E, so that the dictionary width is $K = E \cdot d_\ell$. The main axes show the splice error equation 4, the criterion used for selection, with the knee of each line circled; the inset shows the relative ℓ_1 reconstruction error on held-out token embeddings for the same configurations. All values are means over five sample folds $\pm$ their standard error. The star marks the one-standard-error pick, $(E, k) = (8, 64)$, adopted on both tasks.

A.3 Sparsity Weight Selection for EVA-RNA

We freeze the SAE selected in Appendix A.2 and sweep the sparsity weight $\lambda_z \in \{0.01, 0.03, 0.1, 0.3, 1, 3\}$, solving equation 3 with the optimisation settings of Appendix A.1. The results are reported in Table 2 and plotted in Fig. 2c. We adopt λ_z=0.1 on `t001` and λ_z=0.03 on `t014`, which give comparable errors ($\overline{|\Delta|} = 0.099$ and $\overline{|\Delta|} = 0.141$) with a few to a dozen latents per query (Table 2). The two tasks sit at different weights despite the logit rescaling because, on `t014`, removing all features already changes the prediction less: by $\overline{|\Delta|} = 0.50$ against $\overline{|\Delta|} = 1.12$ on `t001`, averaged over all explanation-pool patients. Each retained feature can therefore reduce the error by less, and λ_z=0.1 that keeps about a dozen features on `t001` keeps about one on `t014`.

Table 2: **Fidelity–sparsity trade-off on EVA-RNA at layer** 15 (K=2048, k=64, 2000 Adam steps). Errors are binary fidelities, mean absolute discrepancies between the counterfactual and original predictions, averaged over output components and queries; the `t001` sweep uses all 19 explanation-pool queries, the `t014` sweep 32 of 76. Adopted rows in bold.

	`t001` (n=19)		`t014` (n=32)	
λ_z	$\overline{\lvert\mathcal{K}^*\rvert}$	$\overline{\lvert\Delta\rvert}$	$\overline{\lvert\mathcal{K}^*\rvert}$	$\overline{\lvert\Delta\rvert}$
0.01	91.5	0.031	9.0	0.057
0.03	29.8	0.047	**4.8**	**0.141**
0.1	**12.6**	**0.099**	1.3	0.289
0.3	6.1	0.173	0.5	0.322
1	3.3	0.392	0.09	0.361
3	1.3	0.706	0.03	0.391

A.4 Attribution Baselines

Shapley values. Computing Shapley values exactly would require evaluating every subset of the candidate pool, so we estimate them by permutation sampling. We start from the empty set and add the pool latents one at a time in a random order; the contribution of a latent is the change in the prediction when it is added. The Shapley value ϕ_k is estimated as the average contribution of latent k over 20 orders per query. These are 10 random orders and their reverses, which reduces the variance of the estimate. Latents outside the pool are removed throughout. In Fig. 3(a), $\bar{\phi}_k$ is the average of ϕ_k over the queries whose selected set contains latent k.

Integrated Gradients. Integrated Gradients attributes to each gate the gradient of the prediction integrated along the straight path from the baseline $\mathbf{z}' = \mathbf{0}$ (all latents removed) to $\mathbf{z} = \mathbf{1}$ (all latents retained):

$$\mathrm{IG}_k = \int_0^1 \frac{\partial}{\partial z_k} \mathcal{F}_{>\ell}\big(\tilde{\mathbf{H}}_{\mathbf{z}}\big)\Big|_{\mathbf{z}=t\mathbf{1}} \,\mathrm{d}t.$$

The integral is estimated with the midpoint rule over 64 steps.

Agreement between methods and redundancy. Table 3 gives the statistics behind the comparison of Section 5.2. Shapley values and Integrated Gradients are in higher mutual agreement than either is with SLI. Of the m latents that one attribution method ranks highest, 58–78% are also among the m highest under the other, whereas only 38–67% are in SLI's selected set (precision@m, Table 1). Over the whole pool, the rankings by $|\phi_k|$ and $|\mathrm{IG}_k|$ have a Spearman correlation of about 0.7 per query. To test redundancy, we take for each unselected latent its highest per-token activation correlation with a selected latent of the same query. For the unselected latents ranked in the top m by $|\phi_k|$, this correlation is markedly higher than for unselected latents in general.

Other logits on `t014`. Table 1 attributes the lymphoid logit. Attributing the fibroid and myeloid logits instead gives a Shapley precision@m of 0.35 and 0.42 and an AUROC of 0.86 and 0.84, and an Integrated Gradients precision@m of 0.28 and 0.42 and an AUROC of 0.80 and 0.80. The high-attribution latents SLI leaves out correlate with a selected one at a median of 0.42 in both cases.

Table 3: **Agreement between attribution methods, and redundancy of the latents SLI leaves out.** Spearman and top-m overlap compare $|\phi_k|$ with $|\mathrm{IG}_k|$ per query (mean over queries). Redundancy is the median, over unselected latents, of their highest per-token activation correlation with a selected latent. `t014`: lymphoid logit.

	Shapley vs IG		redundancy with $\mathcal{K}^*$	
Task	Spearman	top-m overlap	top-m by $\lvert\phi_k\rvert$	all unselected
`t001` (19 queries)	0.71	78%	0.45	0.26
`t014` (63 queries)	0.67	58%	0.40	0.27

Per-query results. Figures 5 and 6 show, for each task, the distribution of absolute scores for selected and unselected latents, and the agreement of each query's ranking with its selected set.

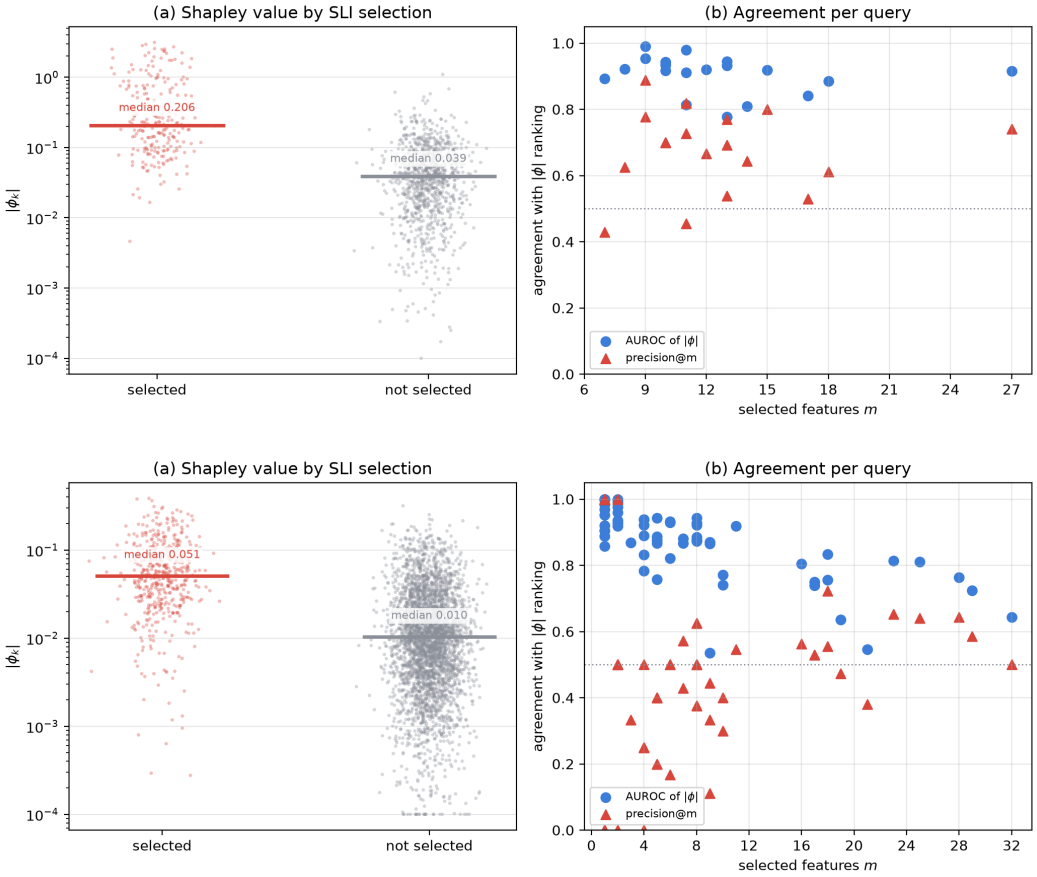

Figure 5: **Shapley values against SLI's selection.** Top: `t001` (19 queries). Bottom: `t014` (63 queries with a non-empty support; lymphoid logit). **(a)** $|\phi_k|$ of selected and unselected latents in the candidate pool, all queries pooled. **(b)** Per query, the AUROC of $|\phi_k|$ as a predictor of selection and the precision@m, against the size m of the selected set.

A.5 Class-Conditional Latent Enrichment on `t014`

Table 4 reports, for every latent selected on at least eight of the 76 `t014` queries, the fraction of each endotype's patients on which it was selected and a χ^2 test of independence between selection and endotype, together with the three genes on whose tokens the latent has the highest mean activation over the queries that selected it.

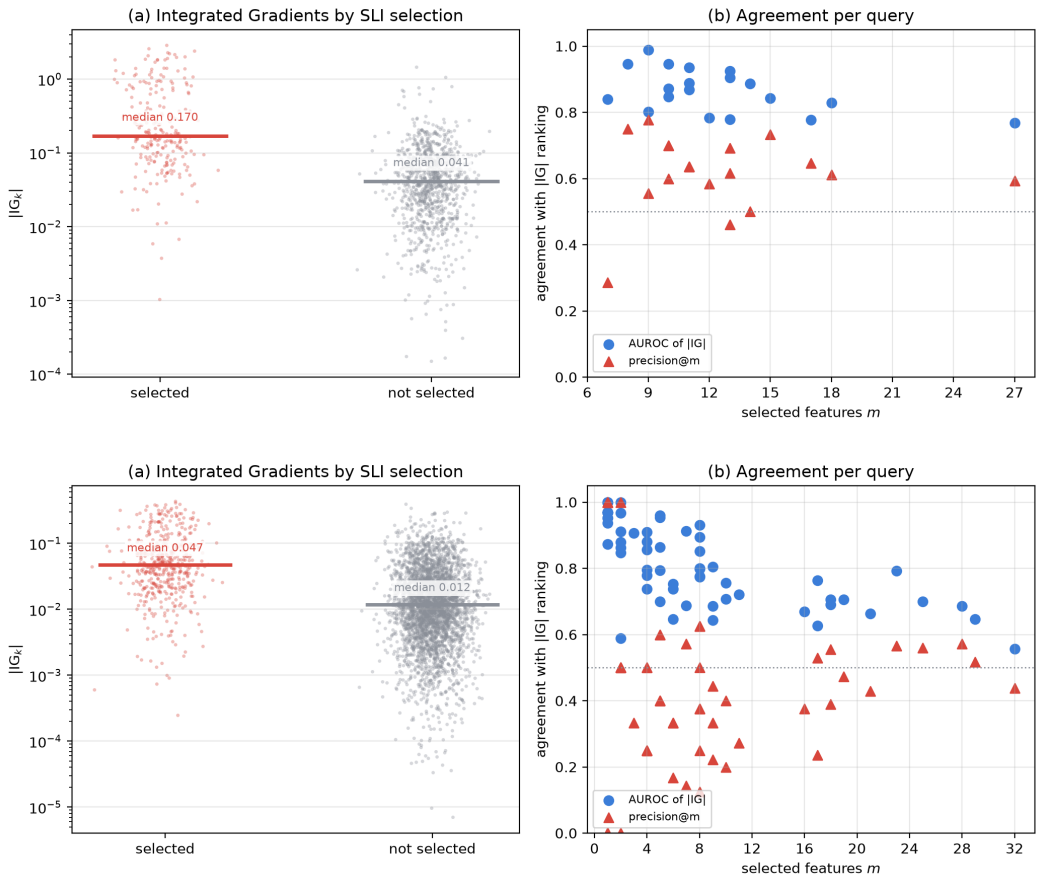

Figure 6: **Integrated Gradients against SLI's selection.** Panels as in Fig. 5, with $|\mathrm{IG}_k|$ in place of $|\phi_k|$.

A.6 Size-Matched Controls

The controls of Section 5.2 and Fig. 2(d) are constructed per query, on the 63 `t014` queries whose support is non-empty. Writing $m = |\mathcal{K}^*|$ for that query, the top-activation set is the m latents with the largest activation mass $\sum_i A_{ik}$ over the gene tokens, and the random sets are 10 draws of m latents, uniformly among the 2048, drawn independently per query. Every set is evaluated under exactly SLI's intervention—its latents at their observed activations, all others removed, the classification row pinned—and at the binary read-out, so the three differ only in which latents they name.

Table 4: **Selection frequency by endotype on `t014`** (n=14 fibroid, 42 lymphoid, 20 myeloid; latents selected on at least eight patients, ordered by total selections).

latent	n	fibroid	lymphoid	myeloid	χ^2 p	top-activating genes
784	47	0.57	0.62	0.65	0.898	*PRG4, FN1, CLU*
1248	35	0.14	0.67	0.25	<0.001	*IGLV1-47, IGKV4-1, IGLV2-11*
1685	29	0.07	0.60	0.15	<0.001	*IGHV4-39, IGLV2-23, IGKV1-9*
1839	22	0.43	0.21	0.35	0.243	*TUBB1, TRNL1, MALAT1*
486	22	0.21	0.31	0.30	0.788	*DSC1, C7, CD3E*
1509	20	0.00	0.45	0.05	<0.001	*SPAG4, DERL3, SDC1*
1268	20	0.29	0.24	0.30	0.855	*DSC1, LGALS7, SERPINB5*
1345	20	0.14	0.36	0.15	0.118	*TNC, TRNC, MMP1*
963	19	0.00	0.43	0.05	<0.001	*DSC1, C7, FBP1*
1428	19	0.00	0.40	0.10	0.002	*IGKC, IGLC2, IGHG3*
1997	17	0.07	0.31	0.15	0.118	*IGHV4-39, IGKV4-1, IGKV3-20*
1041	17	0.07	0.31	0.15	0.118	*LOC105377193, LOC107986061, LOC107985492*
376	12	0.14	0.17	0.15	0.972	*COMP, TMEM196, ZNF385B*
1649	12	0.07	0.17	0.20	0.583	*PRG4, TNC, ELN*
1946	12	0.00	0.26	0.05	0.020	*IGHG4, IGKV1-12, IGKV3D-15*
545	11	0.07	0.17	0.15	0.679	*LGALS7, ZFY, UTY*
59	10	0.07	0.12	0.20	0.517	*TUBB1, RPS17, BCYRN1*
834	9	0.07	0.12	0.15	0.784	*LGALS7, KLK12, GGT8P*
1987	9	0.07	0.10	0.20	0.409	*HBA1, HBA2, ALAS2*
341	9	0.07	0.12	0.15	0.784	*TUBB1, IGHV4-39, CDH1*
179	8	0.07	0.14	0.05	0.485	*IGKV3-20, IGKV2D-28, IGKV1-39*
1863	8	0.07	0.10	0.15	0.726	*ALAS2, SLC4A1, HBA1*
286	8	0.07	0.14	0.05	0.485	*TRNT, TRNN, TRNC*