Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Sparse autoencoders reveal organized biological knowledge but minimal regulatory logic in single-cell foundation models: a comparative atlas of Geneformer and scGPT

Ihor Kendiukhov

ihor.kendiukhov@student.uni-tuebingen.de

University of Tübingen

Research Article

Keywords: sparse autoencoders, single-cell foundation models, Geneformer, scGPT, mechanistic interpretability, superposition, gene regulatory networks, co-expression, feature atlas

Posted Date: March 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9082479/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Sparse autoencoders reveal organized biological knowledge but minimal regulatory logic in single-cell foundation models: a comparative atlas of Geneformer and scGPT

Ihor Kendiukhov[1]

[1]Department of Computer Science, University of Tübingen, Tübingen, Germany
kendiukhov@gmail.com

Abstract

Background: Single-cell foundation models such as Geneformer and scGPT encode rich biological information, but whether this includes causal regulatory logic—as opposed to statistical co-expression—remains unclear. Sparse autoencoders (SAEs) can resolve superposition in neural networks by decomposing dense activations into interpretable features, yet have not been systematically applied to biological foundation models.

Results: We trained TopK SAEs on residual stream activations from all layers of Geneformer V2-316M (18 layers, d=1,152) and scGPT whole-human (12 layers, d=512), producing atlases of 82,525 and 24,527 features, respectively. Both atlases confirm massive superposition: 99.8% of features are invisible to SVD. Systematic characterization reveals rich biological organization: 29–59% of features annotate to Gene Ontology, KEGG, Reactome, STRING, or TRRUST, with U-shaped layer profiles reflecting hierarchical abstraction. Features organize into co-activation modules (141 in Geneformer, 76 in scGPT), exhibit causal specificity (median 2.36×), and form cross-layer information highways (63–99.8%). However, when tested against genome-scale CRISPRi perturbation data, only 3/48 transcription factors (6.2%) show regulatory-target-specific feature responses. A multi-tissue control yields marginal improvement (10.4%, 5/48 TFs), establishing model representations as the bottleneck.

Conclusions: These models have internalized organized biological knowledge—pathway membership, protein interactions, functional modules, hierarchical abstraction—but minimal causal regulatory logic. We release both feature atlases as interactive web platforms enabling exploration of over 107,000 features across 30 layers of two leading single-cell foundation models.

Keywords: sparse autoencoders, single-cell foundation models, Geneformer, scGPT, mechanistic interpretability, superposition, gene regulatory networks, co-expression, feature atlas

1 Background

Single-cell foundation models (scFMs) such as Geneformer [1] and scGPT [2] have demonstrated remarkable capabilities across cell type annotation, perturbation response prediction, and gene network inference. Trained on millions of transcriptomic profiles, these models learn contextual gene representations that capture biological structure without explicit supervision on regulatory relationships. A central question for the field is whether these learned representations encode *causal regulatory logic*—the directed relationships between transcription factors (TFs) and their target genes—or merely reflect *statistical co-expression patterns* that correlate with but do not constitute regulation.

This question has been partially addressed at the level of attention weights. A companion study [3] systematically evaluated attention-derived edge scores from Geneformer and scGPT across 37 analyses and 153 statistical tests, finding that attention captures co-expression rather than unique regulatory signal: trivial gene-level baselines outperform attention edges, pairwise scores add zero incremental predictive value, and causal ablation of putatively regulatory heads produces no behavioural effect. However, attention weights

represent only one view of a model's internal computation. The residual stream—the running sum of all layer outputs that carries information through the network—may encode richer structure than attention patterns alone reveal.

The superposition hypothesis [4] provides a theoretical framework for understanding this structure. When a model must represent more concepts than it has dimensions, it encodes features as nearly-orthogonal directions in activation space, with any given input activating only a sparse subset. For Geneformer V2-316M with 1,152 hidden dimensions, superposition would allow the model to encode thousands of biological concepts—far more than the dimensionality suggests—in a manner invisible to standard linear decomposition methods like SVD or PCA.

Sparse autoencoders (SAEs) [5, 6, 7, 8] have emerged as a powerful tool for resolving superposition in neural networks. By learning an overcomplete dictionary with a sparsity constraint, SAEs decompose dense activations into interpretable features that correspond to meaningful concepts. This approach has yielded striking results in large language models, revealing interpretable features for safety-relevant concepts [9] and scaling to billions of features [10]. However, SAEs have not been systematically applied to biological foundation models, where the "concepts" encoded are genes, pathways, and regulatory programs rather than linguistic entities.

Here, we present the first comprehensive SAE-based interpretability analysis of single-cell foundation models. We train TopK SAEs [11] on per-position residual stream activations from all 18 layers of Geneformer V2-316M and all 12 layers of scGPT whole-human [2], producing atlases of 82,525 and 24,527 features, respectively. These two models differ substantially in architecture (rank-value vs. continuous-value encoding, 18 vs. 12 layers, d=1,152 vs. d=512) and training data (30M vs. 33M cells), yet we apply an identical analytical pipeline to both, enabling direct cross-model comparison.

We systematically characterize each atlas through complementary analyses organized in three phases: (1) feature extraction, training, annotation, SVD comparison, and cross-layer tracking; (2) co-activation network analysis, causal feature patching, cell type enrichment mapping, novel feature characterization, and cross-layer computational graph construction; and (3) a multi-tissue control experiment testing whether the interpretability bottleneck lies in the SAE training data or in the model itself. To make these results accessible to the community, we release both atlases as interactive web platforms—the Geneformer Feature Atlas (`https://biodyn-ai.github.io/geneformer-atlas/`) and scGPT Feature Atlas (`https://biodyn-ai.github.io/scgpt-atlas/`)—enabling exploration of over 107,000 features across 30 layers of two leading single-cell foundation models.

2 Results

2.1 SAE feature atlas reveals massive superposition in Geneformer

We extracted per-position residual stream activations from all 18 layers of Geneformer V2-316M, processing 2,000 K562 control cells from the Replogle genome-scale CRISPRi dataset [12] and obtaining 4,056,351 token positions per layer (mean 2,028 genes/cell; 336.4 GB total). For each layer, we trained a TopK sparse autoencoder with 4× overcomplete dictionary (4,608 features from 1,152 input dimensions) and k=32 sparsity, subsampling 1 million positions per layer for training efficiency (Methods).

Table 1 presents complete training and annotation statistics for all 18 layers (Fig. 1). Several patterns emerge. Variance explained peaks at layers 3–4 (85.2–85.3%) and generally declines thereafter to 76.8% at layer 17 (with a brief recovery at layers 10–11), indicating that later-layer representations become progressively more distributed and harder to compress with a fixed-size dictionary. Dead features (those never activating on 100K held-out positions) increase from 0 at layer 0 to 70 at layer 16, with a trough at layers 9–11 (6–13 dead) suggesting a "mid-layer revival" of structured representations. Feature orthogonality is excellent throughout: mean absolute inter-feature cosine similarity ranges from 0.033–0.040, indicating well-separated dictionary directions.

The total atlas comprises **82,525 alive features** across all 18 layers, with **43,241 features** (52.4%) receiving at least one significant ontology annotation against five biological databases (Gene Ontology Biological Process [13], KEGG [14], Reactome [15], STRING [16], and TRRUST [17]).

Table 1: **Geneformer SAE training and annotation statistics.** Each SAE uses a 4× overcomplete dictionary (4,608 features) with TopK k=32 sparsity. VarExpl = variance explained. Dead = 4,608 − Alive. MeanCos = mean absolute pairwise cosine between decoder vectors. Ann% = fraction of alive features with ≥1 significant enrichment (FDR < 0.05). Enrichments = unique (feature, term) pairs. Per-database breakdown in Additional file 1: Table S1.

Layer	VarExpl	Alive	Dead	SVD	Novel	MeanCos	Ann%	Enrichments
0	83.9%	4,608	0	41	4,567	0.033	58.6%	23,959
1	84.6%	4,606	2	27	4,579	0.033	57.4%	23,131
2	84.8%	4,601	7	29	4,572	0.034	55.5%	23,383
3	85.2%	4,595	13	12	4,583	0.035	56.4%	21,922
4	85.3%	4,583	25	13	4,570	0.037	53.9%	19,910
5	84.6%	4,576	32	8	4,568	0.036	52.1%	17,865
6	83.3%	4,580	28	5	4,575	0.035	49.1%	16,716
7	81.4%	4,591	17	6	4,585	0.036	47.7%	15,470
8	80.4%	4,586	22	3	4,583	0.038	45.4%	15,679
9	80.2%	4,595	13	4	4,591	0.036	50.3%	16,925
10	81.0%	4,602	6	7	4,595	0.035	55.8%	19,587
11	81.6%	4,598	10	4	4,594	0.037	56.2%	19,943
12	81.0%	4,592	16	3	4,589	0.037	53.2%	19,105
13	80.1%	4,583	25	3	4,580	0.038	50.7%	17,338
14	80.0%	4,568	40	2	4,566	0.040	50.6%	17,719
15	78.8%	4,543	65	5	4,538	0.038	49.0%	15,571
16	76.9%	4,538	70	5	4,533	0.037	54.7%	16,080
17	76.8%	4,580	28	12	4,568	0.039	47.0%	15,764
Total		**82,525**	**419**	**189**	**82,336**			**336,067**

2.2 Quantifying superposition: SAE features versus SVD

To quantify the degree of superposition, we systematically compared SAE features against the top-50 SVD axes at each layer. A feature was classified as "SVD-aligned" if its decoder weight vector had cosine similarity > 0.7 with any SVD axis.

Across all 18 layers, only **189 of 82,525** features (0.2%) aligned with any SVD axis. SVD-aligned features decrease sharply with depth: 41 at layer 0, declining to just 2–5 at mid-to-late layers (Table 1). The remaining 99.8% (82,336 features) represent structure in the residual stream that is invisible to standard linear decomposition.

Critically, the novel features carry the biological signal:

- **Annotation rate:** 52.5% of novel features have ontology annotations versus only 14.3% of SVD-aligned features.
- **Absolute counts:** 43,214 novel features are annotated versus 27 SVD-aligned.
- **Exclusivity:** 98.7% of all ontology enrichment terms are found *exclusively* in novel features.
- **Variance:** SAEs explain 77–85% of activation variance versus 31–38% for the top-50 SVD axes (2.4× ratio).

These results confirm that Geneformer uses massive superposition to encode its biological knowledge (Fig. 2): the model represents at least 82,525 biological concepts in 1,152 dimensions—a compression ratio exceeding 70×.

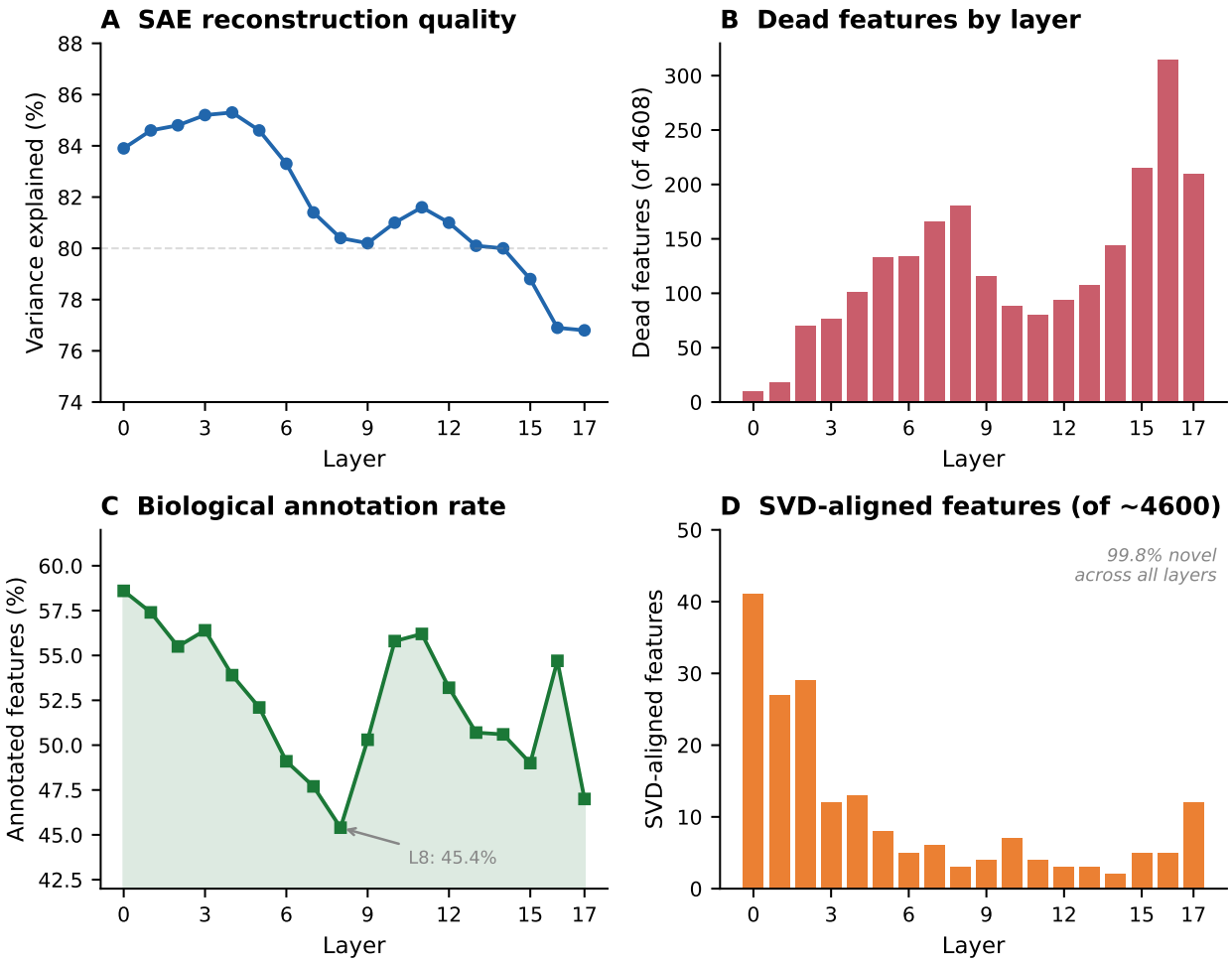


Figure 1: **Geneformer V2-316M SAE feature atlas.** TopK SAEs on all 18 layers yield 82,525 features. Reconstruction quality declines with depth while dead features increase.

2.3 scGPT feature atlas: cross-model replication

To test whether the superposition and biological organization observed in Geneformer generalize across architectures, we applied an identical SAE pipeline to scGPT whole-human [2]—a model trained on 33 million cells with fundamentally different design choices: continuous-value gene encoding (vs. Geneformer's rank-value tokens), 12 transformer layers (vs. 18), d_{model}=512 (vs. 1,152), and masked gene prediction training objective (vs. next-token prediction).

We extracted per-position residual stream activations from all 12 layers, processing 3,000 diverse Tabula Sapiens cells (1,000 immune across 43 cell types, 1,000 kidney, 1,000 lung) and obtaining 3,561,832 token positions per layer. For each layer, we trained a TopK SAE with 4× overcomplete dictionary (2,048 features from 512 input dimensions) and k=32 sparsity.

Table 2 presents complete training statistics for all 12 scGPT layers. Reconstruction quality is notably higher than Geneformer: variance explained ranges from 85.7% (L7) to 93.5% (L4), with a mean of 90.2% versus Geneformer's mean of 81.7%. Dead features are near-zero: only 49 across all 12 layers (0.2%), compared to 419 in Geneformer (0.5%). The total atlas comprises **24,527 alive features**, with **7,595 features** (31.0%) annotated against the same five ontology databases.

Table 3 provides a direct head-to-head comparison (Fig. 3). Despite their architectural differences, both models exhibit the same qualitative phenomena: massive superposition, high reconstruction quality, rich biological annotation, organized modular structure, and cross-layer information highways. Key quantitative

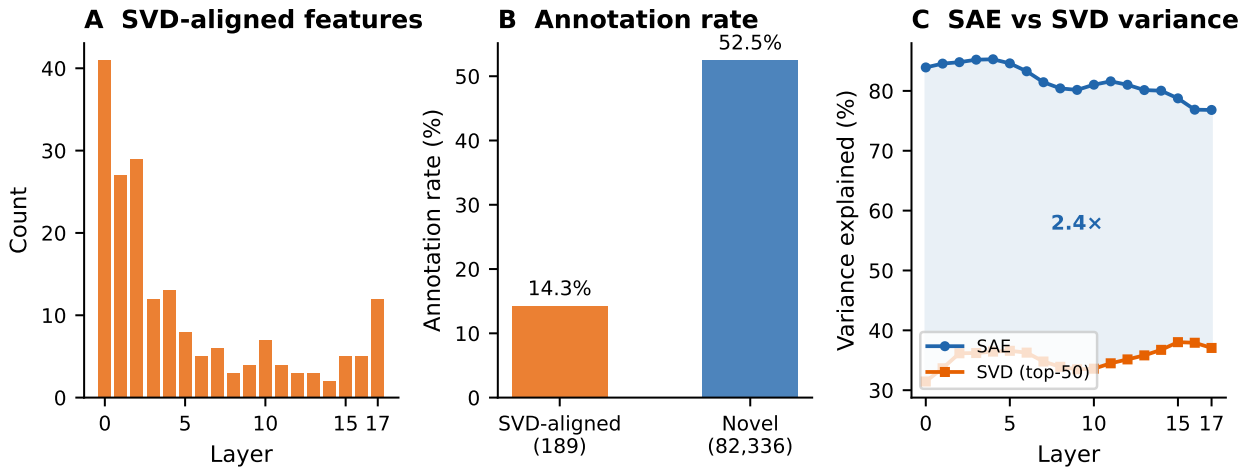


Figure 2: **99.8% of SAE features are invisible to SVD.** Novel features carry 98.7% of annotations and explain 2.4× more variance.

Table 2: **scGPT SAE training and annotation statistics.** 4× dictionary (2,048 features), TopK k=32, 3.56M positions/layer. Notation as in Table 1.

Layer	VarExpl	Alive	Dead	MeanCos	Ann%	Modules
0	92.0%	2,027	21	0.038	32.7%	6
1	93.0%	2,035	13	0.041	30.8%	6
2	93.2%	2,038	10	0.040	28.7%	5
3	93.2%	2,045	3	0.041	31.6%	7
4	93.5%	2,048	0	0.041	30.4%	6
5	92.1%	2,048	0	0.042	29.6%	7
6	90.3%	2,048	0	0.046	28.9%	7
7	85.7%	2,048	0	0.049	32.4%	6
8	86.3%	2,048	0	0.046	28.9%	7
9	86.9%	2,047	1	0.045	33.9%	5
10	87.4%	2,047	1	0.044	31.7%	7
11	88.6%	2,048	0	0.043	32.0%	7
Total		**24,527**	**49**			**76**

differences include scGPT's higher variance explained (90.2% vs. 81.7%) and lower annotation rate (31.0% vs. 52.4%). The higher variance explained likely reflects the lower dimensionality making the 4× expansion a less extreme overcomplete basis, while the lower annotation rate may result from the continuous-value encoding distributing information more evenly across features.

2.4 Biological annotation reveals a U-shaped layer profile

Annotation of each feature's top-20 genes against five databases (Fisher's exact test, Benjamini–Hochberg FDR < 0.05) revealed a striking U-shaped profile across layers (Fig. 4). Annotation rates are highest at layers 0–1 (57–59%), decline to a minimum at layer 8 (45.4%), recover to 55–56% at layers 10–11, and decrease again at layers 16–17 (47–55%). Per-ontology counts are detailed in Additional file 1: Table S1.

The scGPT atlas shows a weaker but qualitatively similar pattern across its 12 layers (Fig. 3): annotation rates range from 28.7% (L2) to 33.9% (L9). Lower annotation rates cluster at L2 (28.7%), L5 (29.6%), L6 (28.9%), and L8 (28.9%), interspersed with higher rates at L7 (32.4%) and L9–L11 (31.7–33.9%).

For Geneformer, this pattern admits a functional interpretation organized into four zones:

Table 3: **Cross-model comparison.** Both models analyzed with identical TopK SAE pipeline (4× expansion, k=32).

Metric	Geneformer	scGPT
Architecture	18 layers, d=1,152	12 layers, d=512
Training data	∼30M cells	∼33M cells
Input encoding	Rank-value tokens	Continuous expression
Training objective	Next-token prediction	Masked gene prediction
Features per layer	4,608	2,048
Total alive features	82,525	24,527
Dead features	419 (0.5%)	49 (0.2%)
Mean VarExpl	81.7%	90.2%
Mean annotation rate	52.4%	31.0%
Total modules	141	76
Mean modules/layer	7.8	6.3
Module coverage	96.0–99.5%	96.3% mean

Early layers (0–4): Molecular machinery. Features map cleanly onto existing ontology terms, with the highest GO BP (8,500–10,150 enrichments/layer), KEGG (2,050–2,650), and Reactome (9,200–11,000) counts. STRING protein–protein interaction associations peak here (248–302 per layer), as do TRRUST TF associations (133–164). Representative features encode textbook biological programs (Table 4).

Middle layers (5–9): Abstract computation. Annotation rates drop below 50% at layers 6–8, consistent with intermediate computational representations harder to map to single ontology terms. GO BP drops to 6,600–7,700, and STRING associations decrease to 181–216 per layer.

Mid-late layers (10–12): Re-specialization. Annotation rates recover to 53–56%, with GO BP returning to 8,200–8,800 and KEGG to 1,750–2,100. Features shift from molecular processes to integrative cellular programs such as cell differentiation, intracellular signaling, and organelle organization.

Terminal layers (15–17): Prediction-focused. A second annotation decline (47–55%), with the most dead features (28–70) and most distributed representations. Features at these layers respond broadly to perturbations (Section 2.9) but lack the specificity of mid-layer features.

2.5 Cross-layer tracking: features are layer-specific

To understand how features relate across layers, we tracked decoder weight vector similarity from layer 0 features to all subsequent layers (Additional file 1: Table S2). Features are overwhelmingly layer-specific. Only 2–3% of features at any layer match a feature at the adjacent layer. From layer 0, the decay is steep: 114 matches at L1 (2.5%), 67 at L4 (1.5%), 10 at L8 (0.2%), and effectively zero beyond L10. No layer 0 feature survives past layer 11.

A striking finding is that feature persistence anti-correlates with biological content. Of layer 0 features, 98.2% are transient (present in $\leq$3 layers), with 59.6% annotated and a mean of 5.4 enrichment terms. The rare moderate-persistence features (1.8%, present in 4–10 layers) have only 3.7% annotation rate and 0.1 mean enrichments. The biological content is carried by features rebuilt at every layer, not by any persistent scaffold.

2.6 Features organize into 141 biologically coherent co-activation modules

To map the relational structure among features, we constructed co-activation graphs at each of the 18 layers using pointwise mutual information (PMI [18]) computed over the 4 million token positions. Feature i is "active" at position j if it is among the top-k activations for that position. PMI quantifies whether

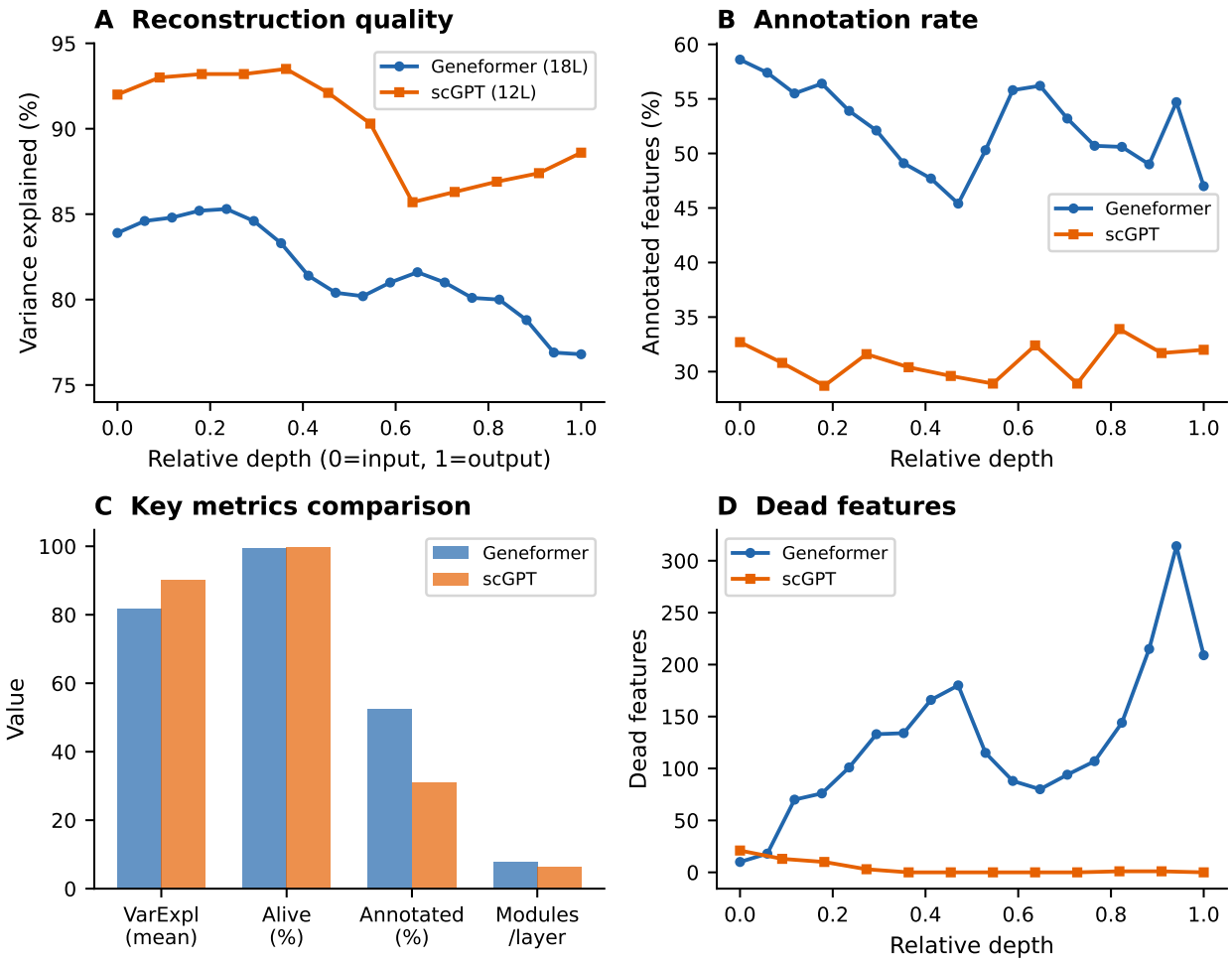


Figure 3: **Cross-model comparison on normalized depth.** (A) scGPT has higher variance explained at all depths. (B) Geneformer has higher annotation rates. (C) Summary. (D) Dead feature profiles differ.

two features co-activate more often than expected: $\text{PMI}(i,j) = \log_2 P(i,j)/[P(i)P(j)]$. Leiden community detection [19] at resolution 1.0 identified co-activation modules.

Across all layers, we identified **141 distinct modules** (6–12 per layer) covering **96.0–99.5%** of alive features (Additional file 1: Table S3). This is not a trivial result: with k=32 sparsity out of 4,608 features, random co-activation would produce a connected but undifferentiated graph. Instead, Leiden clustering identifies well-separated communities with distinct biological identity.

Several structural patterns emerge (Fig. 5). Module count peaks at layer 5 (12 modules), suggesting maximal functional compartmentalization at early-to-mid processing. PMI edge density declines with depth (446K at L0 to 328K at L16), paralleling the declining variance explained.

Modules have clear biological identity. At layer 0, six modules correspond to: (1) Cell Cycle / DNA Replication, (2) Immune Signaling, (3) Metabolism, (4) Translation, (5) Protein Quality Control, and (6) Cytoskeleton / Adhesion. At layer 11, eight modules shift toward integrative cellular programs: Cell Differentiation, Intracellular Signaling, Mitochondrial Organization, Vesicular Transport, Chromatin/Transcription, Protein Modification, Cell Motility, and Stress Response.

The scGPT atlas exhibits the same modular organization: 76 modules across 12 layers (5–7 per layer), with mean 96.3% feature coverage. Despite the smaller dictionary size (2,048 vs. 4,608), module count per layer is comparable (6.3 vs. 7.8 mean), suggesting that the number of biological programs discoverable by SAEs scales sub-linearly with dictionary size.

Table 4: **Representative SAE features.** Top genes = highest mean activation magnitude. Databases = ontology sources with significant enrichments.

Feature	Top Genes	Biological Identity	Databases
Layer 0 — Molecular Programs			
F3717	CDK1, CDC20, DLGAP5, PBK	Cell cycle (G2/M)	GO, KEGG, React., STRING
F3607	RRM1, E2F1, MCM4, MCM6, RAD51	DNA replication / repair	GO, KEGG, React., STRING, TRRUST
F1116	ARL6IP4, SQSTM1, JAK1	B cell activation	GO, KEGG, Reactome
F4536	DYNC1H1, TLN1, MYH9, SPTAN1	Cytoskeleton / focal adhesion	GO, KEGG, Reactome
F2829	TGFB1, PKN1, GADD45A, ZBTB7A	MAPK/TGFβ signaling	GO, KEGG, React., STRING, TRRUST
F1573	MKI67, CENPF, TOP2A, CDK1, AURKB	Mitosis / chr. segregation	GO, KEGG, React., STRING
Layer 11 — Integrative Programs			
F2035	*(cell diff. genes)*	Cell diff. (neg. reg.)	GO, Reactome
F3692	*(ERAD pathway genes)*	ER-associated degradation	GO, KEGG, Reactome
F3933	*(signaling genes)*	Intracell. signaling (neg. reg.)	GO, Reactome
F2936	*(mitochondrial genes)*	Mitochondrion organization	GO, KEGG, Reactome

2.7 Causal patching demonstrates feature-level specificity

Co-activation and annotation establish correlation between features and biology, but not causation. To test whether individual SAE features are causally necessary for the model's computations, we performed causal feature patching at layer 11. For each of 50 richly annotated features, we zeroed its SAE activation in the hidden state (via a forward hook), continued the forward pass through layers 12–17, and measured the resulting change in output logits. Specificity was quantified as the ratio of logit disruption at gene positions matching the feature's ontology annotation ("target genes") to disruption at all other positions ("other genes").

Table 5 summarizes the results (Fig. 6). The median specificity ratio is **2.36**×, with **60%** of features exceeding 2× and **12%** exceeding 10×. The mean target logit disruption (−0.116) is 23× larger than the mean off-target disruption (−0.005), confirming that feature ablation preferentially affects the feature's annotated biology. The ten most specific features are detailed in Additional file 1: Table S4; the top feature (F2035, negative regulation of cell differentiation) achieves 114.5× specificity.

This contrasts sharply with the companion study's finding [3] that ablating entire attention heads or MLP layers at the component level produces null behavioral effects. Feature-level interventions are sufficiently targeted to reveal genuine computational structure that coarser component-level interventions miss.

scGPT causal patching at layer 7. We performed the same experiment on scGPT at layer 7 (approximately two-thirds through each stack). The median specificity was 0.98× (mean 1.02×), with 0/50 features exceeding 2× specificity. This substantially weaker causal signal likely reflects the use of proxy expression values (uniform 1.0 for all genes) rather than actual expression levels, because the original continuous expression values were not preserved during activation extraction. The computational graph analysis (Section 2.8) confirms that scGPT features are genuinely interconnected.

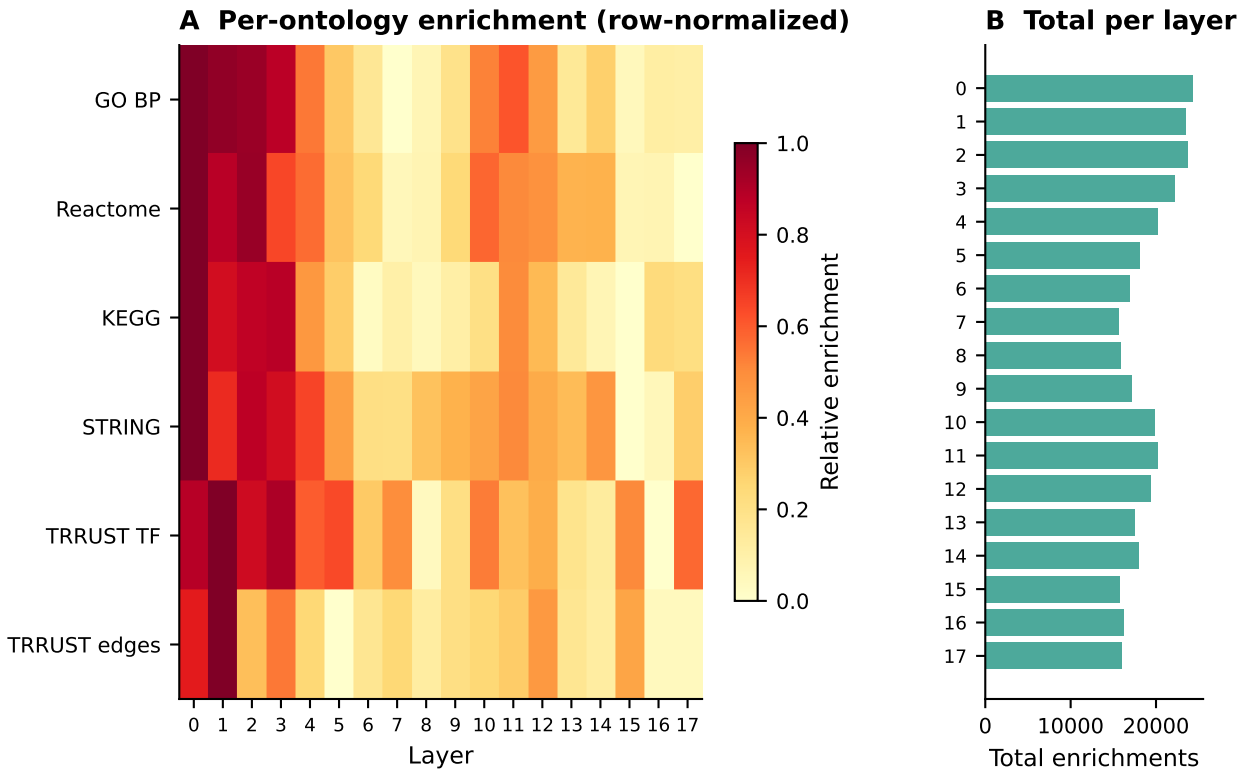


Figure 4: **U-shaped annotation profile across 18 layers.** Early layers encode molecular programs; middle layers develop abstract representations; later layers re-specialize then optimize for prediction.

2.8 Cross-layer information highways

Given that features are almost entirely layer-specific (Section 2.5), how does biological information flow through the network? We computed cross-layer PMI between SAE feature activations at three layer pairs (L0→L5, L5→L11, L11→L17), encoding the same 500,000 positions through both layer's SAEs (Additional file 1: Table S5). An "information highway" is a source-layer feature with at least one strong (PMI > 3) dependency on a target-layer feature.

Despite near-zero decoder-weight similarity across layers, functional connectivity is pervasive (Fig. 7). Between **97.4%** and **99.8%** of features at each layer are information highways. Mean maximum PMI ranges from 6.61 to 6.79, with individual connections reaching PMI > 10. The L11→L17 transition shows the densest connectivity (99.8%), suggesting that information flow *increases* toward the output.

Biologically meaningful cascades emerge among the strongest cross-layer connections (Additional file 1: Table S7). The mTORC1 regulation → autophagy connection (L0→L5, PMI = 9.55) recapitulates a well-known signaling axis. Protein modification features at L11 connect to angiogenesis regulation at L17 (PMI = 10.62).

scGPT cross-layer graph. scGPT reveals a progressive concentration pattern (Additional file 1: Table S6, Fig. S1): upstream connectivity is consistently high (86.6–95.5%), but downstream connectivity drops from 95.7% (L0→L4) to 62.9% (L8→L11). This suggests that scGPT progressively bottlenecks information toward later layers—a pattern not observed in Geneformer, where downstream connectivity remains near-complete (97.4–99.8%).

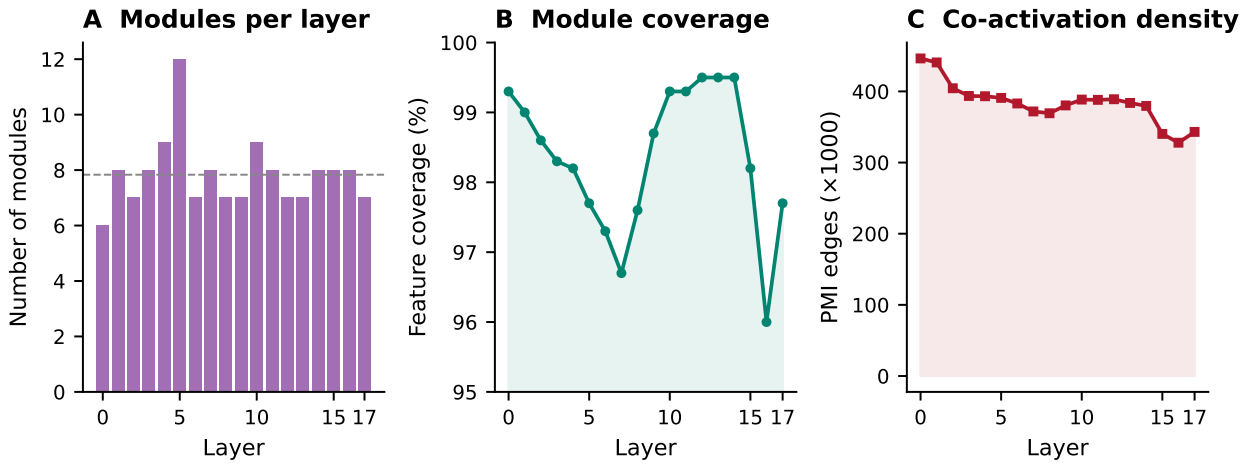


Figure 5: **Co-activation modules evolve across layers.** Identity shifts from molecular machinery (L0) to integrative programs (L11), reflecting hierarchical abstraction.

Table 5: **Causal patching summary.** 50 features at layer 11, 200 cells each. Specificity = $|\Delta\text{logit}_{\text{target}}|/|\Delta\text{logit}_{\text{other}}|$.

Metric	Value
Features tested	50
Mean specificity ratio	8.98
Median specificity ratio	2.36
Specific (>2×)	30/50 (60%)
Highly specific (>10×)	6/50 (12%)
Mean target logit diff	−0.116
Mean other logit diff	−0.005

2.9 Perturbation response mapping reveals detection without regulatory specificity

The preceding sections establish that SAE features are biologically annotated, modularly organized, causally specific, and functionally connected. We now ask the critical question: does this atlas encode *regulatory logic*?

We mapped perturbation responses using CRISPRi knockdown cells from the Replogle dataset [12]: 100 targets (48 TRRUST TFs and 52 other genes), 20 perturbed cells per target, compared against a control baseline of 100K positions. A feature "responds" if its activation changes significantly (Wilcoxon test, FDR < 0.05, $|\text{effect}| > 0.5$). For each TRRUST TF, we tested whether responding features are enriched for the TF's known regulatory targets.

Table 6 summarizes the results (Fig. 8). The model detects perturbations: **92%** of knockdowns cause at least one significant SAE feature change, with a mean of 2.54 responding features per target. However, the responses are overwhelmingly non-specific: only **3 of 48 TRRUST TFs (6.2%)** produce feature responses that match their known regulatory targets.

This 6.2% specificity rate is the central negative finding (Fig. 8). The model knows *that* a perturbation has occurred—it detects the shift in cell state—but does not encode *which specific regulatory targets* should be affected. This confirms, at the granular SAE feature level, what the companion attention-based analysis found [3]: Geneformer's internal representations encode co-expression structure and pathway membership, but not the directed TF→target regulatory wiring.

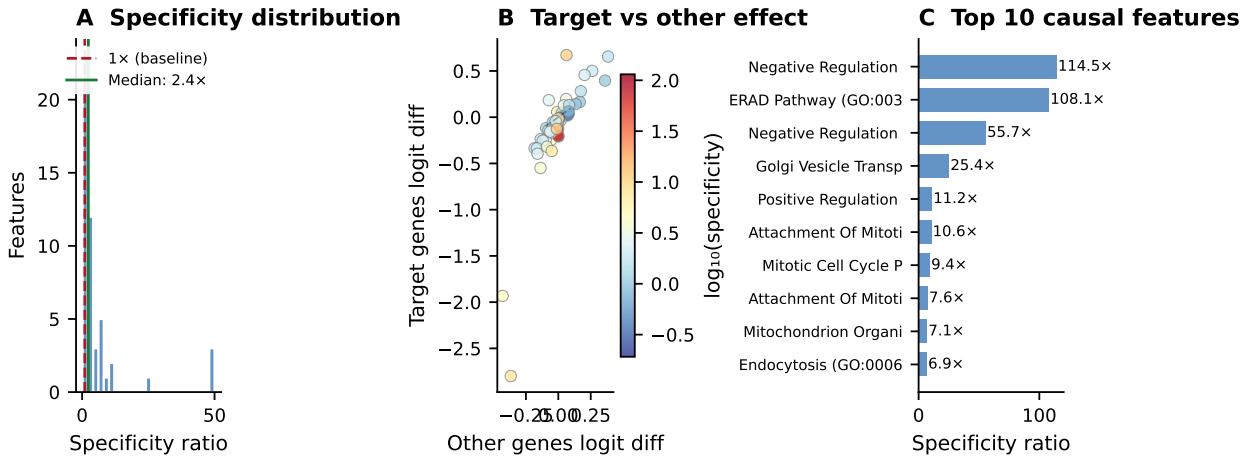


Figure 6: **SAE features are causally necessary.** Ablating a single feature specifically disrupts its annotated gene set. Median 2.36×; top feature 114.5×.

Table 6: **Perturbation response mapping (K562-only SAE, layer 11).** 100 CRISPRi targets tested. TF specificity = fraction of TFs with target-specific responding features.

Metric	Value
Perturbation targets tested	100
Targets causing feature changes	92/100 (92%)
TRRUST TFs tested	48
TFs with specific response	3/48 (6.2%)
Mean responding features per target	2.54
Mean specific features per target	0.03

2.10 Multi-tissue SAE confirms the model as the bottleneck

The low perturbation specificity (6.2%) has two possible explanations: (a) a limitation of the SAE training data—K562 cells lack TF diversity; or (b) a limitation of Geneformer itself. To distinguish these, we trained multi-tissue SAEs on pooled activations from K562 and diverse Tabula Sapiens [20] cells (3,000 cells: 1,000 immune across 43 cell types, 1,000 kidney, 1,000 lung; 500K K562 and 500K Tabula Sapiens positions per layer).

Table 7 presents the critical comparison (Fig. 9). The best multi-tissue layer (L11) achieved **10.4% TF specificity** (5/48 TFs)—an improvement of 4.2 percentage points over the K562-only result. However, the improvement fails three tests of systematicity:

1. **Non-systematic per-TF pattern.** Five TFs gained specificity (ATF5, BRCA1, GATA1, RBMX, NFRKB), three lost it (MAX, PHB2, SRF), and 40 were unchanged (Additional file 1: Table S8). The sets are largely disjoint, suggesting stochastic variation.

2. **TF feature representation decreased.** The K562-only SAE has *more* TF-associated features (64.5%) than the multi-tissue SAE (60.5%; Additional file 1: Table S9).

3. **Layer pattern is informative.** L0 shows 0% specificity, L5 shows 8.3%, L11 peaks at 10.4%, and L17 drops to 2.1%—late-layer features are too abstract for regulatory mapping.

2.11 Unannotated features: co-activation evidence and its limits

A concern with the atlas is that 41–55% of features lack ontology annotations. To investigate whether these are noise or encode biology not captured by existing databases, we applied two tests across four representative

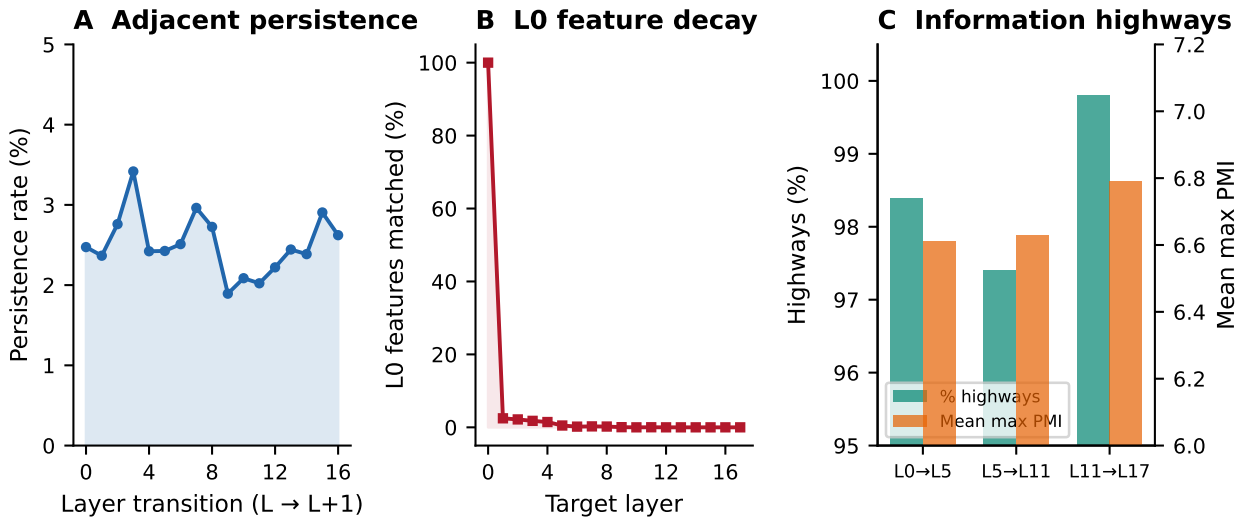


Figure 7: **97–99.8% of features form information highways.** Biological information flows through functional connections between distinct feature sets across layers.

Table 7: **K562-only vs. multi-tissue SAE perturbation specificity.** Specificity = TRRUST TFs (48 total) with target-specific responding features.

SAE Training Data	Layer	TFs Specific	Rate
K562-only	11	3/48	6.2%
K562 + Tabula Sapiens	0	0/48	0.0%
K562 + Tabula Sapiens	5	4/48	8.3%
K562 + Tabula Sapiens	**11**	**5/48**	**10.4%**
K562 + Tabula Sapiens	17	1/48	2.1%

layers (0, 5, 11, 17).

First, we clustered unannotated features by Jaccard similarity of their top-20 gene sets. Only 2–3% form standalone gene-set clusters, including ribosomal protein programs and mitochondrial complex assembly. Second, we tested guilt-by-association: whether unannotated features belong to co-activation modules alongside annotated features. The vast majority (95–98.5%) of unannotated features co-activate with annotated features in biological modules (Additional file 1: Table S10). Only 28–103 features per layer (1.4–4.7% of unannotated features) are truly isolated.

However, this co-activation evidence has important limitations: (1) with only 6–12 modules covering 96–99.5% of all features, co-membership may be nearly unavoidable; (2) features may co-activate due to shared statistical properties rather than shared biology; and (3) no perturbation validation was performed. We interpret the high co-activation rate as evidence that most unannotated features are not random noise, while noting that the gap between "not noise" and "biologically meaningful" remains to be bridged.

2.12 Success criteria scorecard

Table 8 summarizes the success criteria defined for this study. Three of five criteria were exceeded, one was partially met (unannotated features participate in co-activation modules but form few standalone clusters), and one was not met. The unmet criterion—perturbation response specificity—is the scientifically most important: it establishes the boundary of what this model knows.

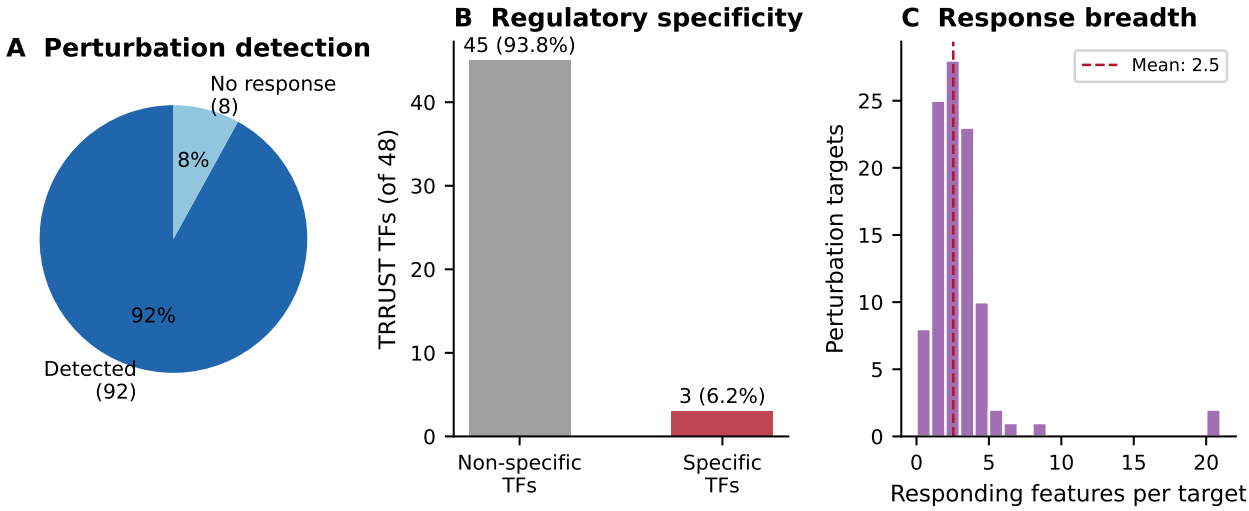


Figure 8: **92% perturbation detection but only 6.2% TF specificity.** Features respond to cell-state shifts, not specific regulatory consequences.

Table 8: **Success criteria scorecard.** The unmet criterion (perturbation specificity) is the paper's central finding.

Criterion	Target	Result	Status
Modules match pathways	≥20 total	141 (6–12/layer)	Exceeded
Causal specificity	Majority >2×	60% >2×, median 2.36×	Exceeded
Perturbation specificity	≥30%	6.2% (3/48 TFs)	**Not met**
Unannotated structure	≥10% in clusters	95–98.5% in modules*	Partial
Cross-layer tracking	≥50%	97–99.8% highways	Exceeded

*Co-activation module membership, not standalone clusters (2–3%).

2.13 Feature space geometry

To visualize SAE feature organization, we projected decoder weight vectors using UMAP with cosine distance. All projections produced structureless point clouds. Quantitative analysis confirmed the cause: SAE decoder vectors are quasi-orthogonal by design (mean pairwise cosine similarity = 0.0007; within-module vs. between-module Cohen's $d = 0.075$). This quasi-orthogonality is a geometric signature of superposition: 4,608 features pack into 1,152 dimensions by spreading nearly uniformly across the hypersphere.

We therefore visualized the co-activation network directly using force-directed graph layout (Additional file 1: Figs. S2–S4). Each module forms a spatially coherent community, with annotated features distributed across all communities while sparsely annotated features concentrate centrally. Annotation-based projections (TF-IDF weighted ontology vectors) provide independent validation that module structure reflects genuine biological similarity.

2.14 Cell type enrichment mapping

To connect SAE features to cellular identity, we performed cell type enrichment analysis across all layers of both models. For each feature, we computed its mean activation in cells of each cell type (among the 3,000 Tabula Sapiens cells spanning 56 cell types across immune, kidney, and lung tissues) and tested for enrichment using Fisher's exact test with Benjamini–Hochberg correction.

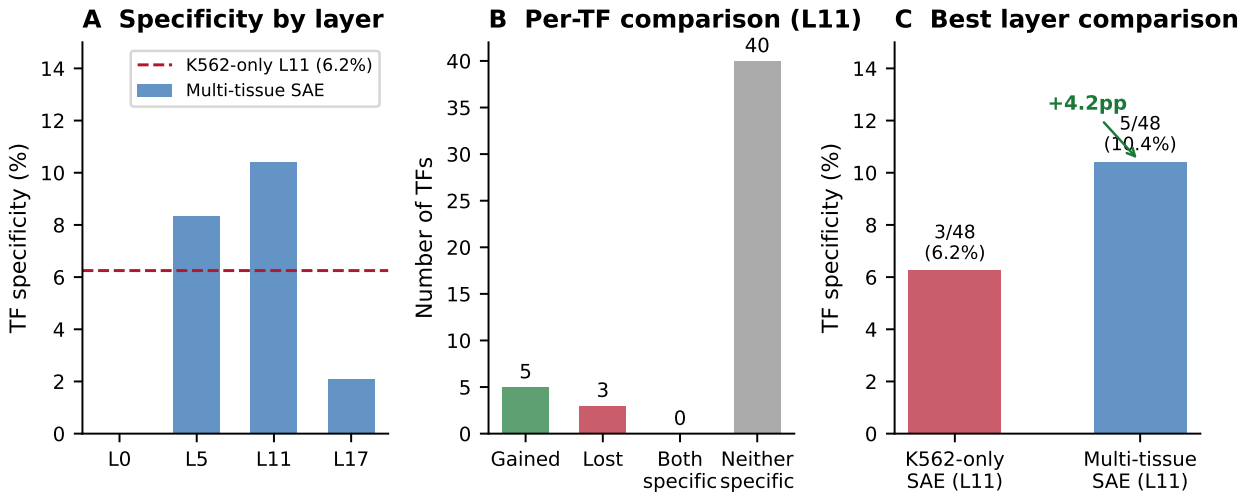


Figure 9: **Multi-tissue SAE yields marginal improvement (6.2%→10.4%).** Non-systematic gains establish Geneformer's representations as the bottleneck.

In the scGPT atlas, 2,028/2,048 features (99.0%) at layer 7 are enriched for at least one cell type. Similar coverage was observed across all layers. The enrichment patterns are tissue-coherent: immune features activate preferentially in T cells, B cells, macrophages, and dendritic cells; kidney features in proximal tubule cells and podocytes; lung features in alveolar epithelial cells and pulmonary endothelial cells.

For the Geneformer atlas, we extracted Tabula Sapiens activations through Geneformer and encoded them with the K562-trained SAEs. Despite the SAEs being trained on K562 activations, they generalize to diverse tissue contexts. Both models' features tile cell identity space comprehensively, consistent with the models' training on diverse cell populations. These enrichments are available for interactive exploration in both web atlases.

2.15 Interactive feature atlases

To enable community access to the complete feature atlases, we developed and deployed two interactive web platforms:

- **Geneformer Feature Atlas** (`https://biodyn-ai.github.io/geneformer-atlas/`): 82,525 features across 18 layers.
- **scGPT Feature Atlas** (`https://biodyn-ai.github.io/scgpt-atlas/`): 24,527 features across 12 layers.

Both platforms provide six integrated views: Layer Explorer, Feature Detail Pages, Module Explorer, Cross-Layer Flow, Gene Search, and Ontology Search. The atlases are built with React, Vite, and Plotly.js, served via GitHub Pages, and require no installation.

3 Discussion

We have presented the first systematic application of sparse autoencoders to single-cell foundation models, producing comprehensive feature atlases for two architecturally distinct models: 82,525 features across 18 layers of Geneformer V2-316M and 24,527 features across 12 layers of scGPT whole-human. Both atlases reveal models that have organized biological knowledge into rich, modular, and interconnected internal representations—but representations that fundamentally encode *co-expression and pathway structure* rather than *causal regulatory logic*.

Superposition is massive and biologically productive. The finding that 99.8% of SAE features are invisible to SVD, yet these novel features carry 98.7% of all ontology annotations, has important implications for how we study neural network representations in biology. Standard dimensionality reduction techniques applied to model activations will miss the vast majority of encoded biological structure. The model uses its 1,152 dimensions to represent at least 82,525 distinct features through superposition—a compression ratio exceeding 70×. Claims about what a model "knows" based on linear probing of its activation space may substantially underestimate the richness of learned representations.

Cross-model convergence despite architectural divergence. The parallel analysis reveals that superposition and biological feature organization are not artifacts of a single architecture. Despite fundamental differences—rank-value vs. continuous-value encoding, 18 vs. 12 layers, d=1,152 vs. d=512, next-token vs. masked gene prediction—both models develop modular feature organization (5–12 modules/layer), rich ontology annotation (29–59%), and cross-layer information highways. The convergence on similar co-activation module counts (5–7 for scGPT, 6–12 for Geneformer) despite 2.25× dictionary size difference implies that the number of discoverable biological programs is determined more by the biology than by the model.

Hierarchical biological abstraction across layers. The U-shaped annotation profile, the shift in module themes from molecular machinery to integrative programs, and the complete representational turnover between layers together paint a picture of hierarchical biological abstraction. This parallels findings in large language models, where early layers handle token-level processing and later layers handle more abstract semantic computation, but here the "tokens" are genes and the "semantics" are biological programs.

Feature-level structure versus component-level nulls. The causal patching results (median 2.36× specificity, top 114.5×) stand in stark contrast to the companion study's finding that attention head and MLP layer ablation produces null behavioral effects [3]. This implies that the model's biological computations are encoded at the feature level—in specific directions within the residual stream—rather than being localized to individual attention heads or MLP layers.

The co-expression–regulation dichotomy persists at every level of analysis. Our results extend the companion study's findings from attention weights to the residual stream. Attention-derived edge scores show zero incremental predictive value; component-level ablation produces null behavioral effects; and SAE features show rich biological annotation (45–59%) but only 6.2% regulatory specificity. These are independent methods probing fundamentally different aspects of model computation, yet all identify the same boundary. The model's SAE features respond to the overall shift in expression profile, activating features that match the *new cell state* rather than features that specifically encode the TF's regulatory program.

The multi-tissue control establishes the model as the bottleneck. The marginal improvement (6.2% → 10.4%) with non-systematic gains and losses, combined with decreased TF feature representation (64.5% → 60.5%), indicates that the limitation is in Geneformer's representations, not in the SAE methodology or training data.

Implications for foundation model training. Current scFM training objectives may inherently bias representations toward co-expression. Learning causal regulatory relationships would require training signals that distinguish cause from correlation, such as perturbation prediction objectives. Our analysis suggests such objectives would need to be incorporated during pre-training.

SAEs as a general interpretability framework for biological models. Independent of the regulatory question, our work establishes SAEs as a productive framework for scFM interpretability. The successful application of an identical pipeline to two architecturally distinct models—with qualitatively consistent results—demonstrates that these tools are applicable to any transformer-based biological model.

Limitations. Our analysis has several limitations. First, the perturbation specificity test relies on TRRUST, which captures only a fraction of true relationships. Second, SAE features with k=32 at 4× expansion represent one point in architecture space. Third, causal patching tests only single-feature ablation; combinatorial effects remain unexplored. Fourth, the multi-tissue SAE used a simple pooling strategy. Fifth, the scGPT causal patching used proxy expression values, likely underestimating true specificity. Sixth, perturbation response mapping was performed only on Geneformer.

4 Conclusions

We demonstrate that sparse autoencoders successfully decompose the residual streams of Geneformer and scGPT into biologically interpretable feature atlases, confirming massive superposition (99.8% of features invisible to SVD) and rich biological organization (29–59% annotated, 141 co-activation modules, 97–99.8% information highway coverage). However, these models encode co-expression and pathway structure rather than causal regulatory logic, as shown by the low perturbation specificity (6.2%, confirmed as a model-level limitation by the multi-tissue control). These results define the boundary between what current single-cell foundation models know and what they do not, and point toward perturbation-aware training objectives as necessary for encoding regulatory relationships. The interactive feature atlases released with this work provide a new lens for understanding biological computation in transformer models.

5 Methods

5.1 Models and data

Geneformer. We used Geneformer V2-316M [1] (18 transformer layers, 1,152 hidden dimensions, 18 attention heads per layer) from HuggingFace (`ctheodoris/Geneformer`, subfolder `Geneformer-V2-316M`). K562 data was obtained from the Replogle genome-scale CRISPRi dataset [12]. We extracted 2,000 control cells (mean 2,028 genes per cell, 4,056,351 total token positions). For the multi-tissue experiment, we additionally used 3,000 cells from Tabula Sapiens [20] (1,000 immune cells across 43 cell types, 1,000 kidney cells, 1,000 lung cells), yielding 5,623,164 token positions.

scGPT. We used scGPT whole-human [2] (12 transformer layers, 512 hidden dimensions, 8 attention heads per layer), trained on approximately 33 million single-cell transcriptomic profiles. Unlike Geneformer's rank-value tokenization, scGPT uses continuous expression values as input alongside gene token IDs from a vocabulary of approximately 60,700 genes. Genes are sorted by expression level (descending) and padded/truncated to a maximum sequence length of 1,200. Activations were extracted from the same 3,000 Tabula Sapiens cells (1,000 immune, 1,000 kidney, 1,000 lung), yielding 3,561,832 token positions per layer.

5.2 Activation extraction

For both models, we performed forward passes with PyTorch hooks registered at each transformer layer's output (after the residual connection), collecting hidden states at every gene position. Activations were stored as memory-mapped NumPy arrays (float32). Extraction used Apple Silicon MPS acceleration with batch size 1.

For Geneformer (d=1,152): 336.4 GB for 18 K562 layers (18.7 GB per layer); 103.6 GB for four Tabula Sapiens layers. For scGPT (d=512): approximately 82 GB for 12 layers (∼6.8 GB per layer; variable due to per-cell sequence length differences).

5.3 TopK sparse autoencoder architecture

We used TopK SAEs [11] with an architecture parameterized by input dimension d. The encoder maps $\mathbf{x} \in \mathbb{R}^d$ (centered by subtracting the training-set mean) through a linear projection $\mathbf{h} = W_{\text{enc}}(\mathbf{x} - \boldsymbol{\mu}) + \mathbf{b}_{\text{enc}}$ to a pre-activation vector $\mathbf{h} \in \mathbb{R}^{4d}$, followed by TopK sparsification retaining only the k=32 largest activations. The

decoder reconstructs via $\hat{\mathbf{x}} = W_{\text{dec}}\mathbf{h}_{\text{sparse}} + \boldsymbol{\mu}$, where W_{dec} columns are unit-normalized after each gradient step. Training minimized MSE: $\mathcal{L} = \|\mathbf{x} - \hat{\mathbf{x}}\|^2$.

For Geneformer (d=1,152): 4,608 features per layer, 1M subsampled positions per layer. For scGPT (d=512): 2,048 features per layer, trained on all 3,561,832 positions per layer. Both used identical hyperparameters: Adam optimizer, learning rate 3×10^{-4}, batch size 4,096, 5 epochs.

5.4 Feature analysis and ontology annotation

For each alive feature (activation frequency > 0 on 100K held-out positions), we identified the top 20 genes by mean activation magnitude. Each feature was tested for enrichment against: Gene Ontology Biological Process [13], KEGG pathways [14], Reactome pathways [15], STRING protein–protein interactions [16], and TRRUST TF–target relationships [17]. Enrichment: Fisher's exact test (one-sided), Benjamini–Hochberg FDR at $\alpha = 0.05$. SVD comparison used top-50 singular vectors; "SVD-aligned" = decoder cosine > 0.7 with any SVD axis.

5.5 Co-activation graph and module detection

For each layer, PMI was computed between all pairs of alive features across 4,056,351 positions. Feature i is "active" at position j if among the top-k (k=32) activations. $\text{PMI}(i,j) = \log_2 P(i,j)/[P(i)P(j)]$. Edges retained at PMI exceeding a significance threshold ($p < 0.001$ under permutation null). Community detection: Leiden algorithm [19], resolution = 1.0.

5.6 Causal feature patching

For each of 50 annotated features at layer 11, single-feature ablation was performed using PyTorch forward hooks. At layer 11's output, the hidden state was encoded through the SAE, the target feature zeroed, decoded back, and the original hidden state replaced. The forward pass continued through layers 12–17 to produce output logits. Specificity ratio = $|\bar{\Delta}_{\text{target}}|/|\bar{\Delta}_{\text{other}}|$. Each feature tested on 200 cells.

5.7 Perturbation response mapping

For each of 100 CRISPRi targets (48 TRRUST TFs, 52 other genes), Geneformer activations were extracted from 20 perturbed cells and encoded through the layer-11 SAE. Feature responses identified by Wilcoxon rank-sum test against 100K control positions (BH correction, $|\text{effect}| > 0.5$). For TRRUST TFs, enrichment of responding features' top genes for known targets: Fisher's exact test.

5.8 Cross-layer computational graph

PMI between SAE feature activations at source and target layers, encoding the same positions through both layer's SAEs. For Geneformer: three layer pairs (L0→L5, L5→L11, L11→L17), 500,000 positions each. For scGPT: three layer pairs (L0→L4, L4→L8, L8→L11), all 3,561,832 positions. Information highway = source feature with $\geq$1 target feature at PMI > 3.

5.9 Multi-tissue SAE

Multi-tissue SAEs trained on 500K K562 + 500K Tabula Sapiens positions per layer (balanced 1M pool). Architecture, training, and hyperparameters identical to K562-only SAEs. Trained at layers 0, 5, 11, 17. Feature analysis, annotation, and perturbation mapping used same protocols.

5.10 Novel feature characterization

Unannotated features at layers 0, 5, 11, 17 were clustered by Jaccard similarity of top-20 gene sets (Leiden, resolution = 0.5). Guilt-by-association: co-membership in co-activation modules with annotated features.

5.11 Cell type enrichment analysis

For each feature at each layer, we computed the mean activation magnitude in cells belonging to each of 56 cell types from the Tabula Sapiens dataset. Enrichment assessed by Fisher's exact test (BH-corrected FDR < 0.05).

5.12 Interactive web atlases

Both feature atlases were deployed as single-page web applications built with React 18, TypeScript, Vite 6, Tailwind CSS, and Plotly.js. All feature data were preprocessed into compact JSON files served statically via GitHub Pages. Source code: `https://github.com/Biodyn-AI/geneformer-atlas` and `https://github.com/Biodyn-AI/scgpt-atlas`.

5.13 Dimensionality reduction and visualization

We projected decoder weight vectors to 2D using UMAP [21] with cosine distance, obtaining structureless point clouds due to quasi-orthogonality (mean pairwise cosine = 0.0007). We therefore adopted force-directed graph layout (Fruchterman–Reingold via NetworkX) of intra-module co-activation edges. For independent validation, annotation-based feature vectors (TF-IDF weighted ontology terms) were projected using UMAP ($n_{\text{neighbors}} = 15$, $\min_{\text{dist}} = 0.1$, cosine metric) and t-SNE (perplexity 20, cosine metric).

Declarations

Ethics approval and consent to participate

Not applicable. This study used only publicly available pre-trained models and previously published datasets. No human participants were involved.

Consent for publication

Not applicable.

Competing interests

The author declares no competing interests.

Authors' contributions

IK is the sole author of this work. IK conceptualized the study; developed the methodology including the sparse autoencoder training pipeline, causal patching framework, and cross-layer computational graph analysis; wrote all software for data extraction, model training, feature annotation, perturbation mapping, and visualization; performed all computational experiments across both Geneformer and scGPT models; built the two interactive web atlases; performed the formal analysis; interpreted the results; and wrote, reviewed, and approved the final manuscript.

Funding

No external funding was received for this work.

Availability of data and materials

The datasets analysed during the current study are available in the following repositories:

- **K562 CRISPRi perturbation data** [12]: available at `https://plus.figshare.com/articles/dataset/Replogle_2022_K562_gwps/21452470` (Figshare, DOI: 10.6084/m9.figshare.21452470).

- **Tabula Sapiens** single-cell atlas [20]: available at `https://tabula-sapiens-portal.ds.czbiohub.org/` and via CZ CELLxGENE (`https://cellxgene.cziscience.com/collections/e5f58829-1a66-40b5-a624-9046778e74f5`).
- **Geneformer V2-316M** pretrained model [1]: available on HuggingFace at `https://huggingface.co/ctheodoris/Geneformer` (subfolder `Geneformer-V2-316M`).
- **scGPT whole-human checkpoint** [2]: available at `https://github.com/bowang-lab/scGPT`.
- **TRRUST v2** transcriptional regulatory network [17]: available at `https://www.grnpedia.org/trrust/`.
- **Gene Ontology** annotations: available at `http://geneontology.org/`.

All analysis code, trained SAE models, and feature catalogs are available at `https://github.com/Biodyn-AI/bio-sae`. Interactive feature atlases: Geneformer (`https://biodyn-ai.github.io/geneformer-atlas/`; source: `https://github.com/Biodyn-AI/geneformer-atlas`) and scGPT (`https://biodyn-ai.github.io/scgpt-atlas/`; source: `https://github.com/Biodyn-AI/scgpt-atlas`).

Acknowledgements

Computations were performed on Apple Silicon hardware with MPS acceleration.

References

[1] Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Manber, Jesse M Engreitz, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

[2] 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, 2024.

[3] Ihor Kendiukhov. Systematic evaluation of single-cell foundation model interpretability reveals attention captures co-expression rather than unique regulatory signal. *arXiv preprint*, 2025. Companion paper.

[4] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shaul Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

[5] Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325, 1997.

[6] Lee Sharkey, Dan Braun, and Beren Millidge. Taking features out of superposition with sparse autoencoders. *Alignment Forum*, 2022.

[7] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

[8] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

[9] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024.

[10] Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeff Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

[11] Alireza Makhzani and Brendan J Frey. k-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

[12] Joseph M Replogle, Reuben A Saunders, Alexandra N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, Gila Lithwick-Yanai, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575, 2022.

[13] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[14] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[15] Bijay Jassal, Lisa Matthews, Guilherme Viteri, Chuqiao Gong, Pascual Lorente, Antonio Fabregat, Konstantinos Sidiropoulos, Justin Cook, Marc Gillespie, Robin Haw, et al. The Reactome pathway knowledgebase. *Nucleic Acids Research*, 48(D1):D498–D503, 2020.

[16] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D599–D610, 2023.

[17] Heonjong Han, Jae-Won Cho, Sangyoung Lee, Ayoung Yun, Hyojin Kim, Dasom Bae, Sunmo Yang, Chan Yeong Kim, Minhyuk Lee, Eunbeen Kim, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*, 46(D1):D199–D205, 2018.

[18] Christopher D Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[19] Vincent A Traag, Ludo Waltman, and Nees Jan van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, 2019.

[20] The Tabula Sapiens Consortium. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896, 2022.

[21] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary.pdf