Additional file 1

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

Ihor Kendiukhov

This file contains supplementary tables and figures for the main manuscript.

Contents

1 Supplementary Tables

Table S1: **Per-ontology enrichment counts across all 18 Geneformer layers.** Each entry is the number of significant enrichments (FDR < 0.05) for features at that layer. GO BP = Gene Ontology Biological Process. TRRUST columns show enrichment for TF target sets (TF) and individual TF→target edges (Edges).

Layer	GO BP	KEGG	Reactome	STRING	TRRUST TF	TRRUST Edges
0	10,153	2,650	11,001	302	155	42
1	10,022	2,433	10,512	258	164	48
2	9,948	2,495	10,790	283	150	32
3	9,726	2,514	9,525	273	157	37
4	8,537	2,045	9,195	248	133	30
5	7,695	1,845	8,189	216	136	24
6	7,180	1,555	7,871	182	110	28
7	6,628	1,637	7,080	181	125	30
8	6,850	1,570	7,169	199	90	27
9	7,299	1,643	7,880	207	103	29
10	8,461	1,751	9,247	214	128	30
11	8,785	2,089	8,957	227	112	31
12	8,217	1,915	8,856	210	117	35
13	7,158	1,686	8,393	202	101	28
14	7,615	1,595	8,412	221	97	27
15	6,790	1,520	7,135	150	126	34
16	7,040	1,781	7,172	158	87	25
17	7,002	1,762	6,869	193	131	25
Total	**145,106**	**34,486**	**154,253**	**3,924**	**2,222**	**562**

Table S2: **Cross-layer feature persistence from layer 0.** Matches = features at L0 with cosine similarity > 0.7 to any feature at the target layer. The model undergoes radical representational transformation: by layer 6, essentially all features are novel with no L0 ancestry.

L0 → Target	Matches (cos > 0.7)	Rate
L0 → L1	114	2.5%
L0 → L2	93	2.0%
L0 → L4	67	1.5%
L0 → L6	25	0.5%
L0 → L8	10	0.2%
L0 → L10	1	∼0%
L0 → L12+	0	0%

Table S3: **Co-activation module statistics across all 18 layers.** PMI-based graphs with Leiden clustering (resolution = 1.0). Modules = number of distinct communities. Coverage = fraction of alive features in at least one module.

Layer	Modules	Feats in Modules	PMI Edges	Coverage
0	6	4,577	446,324	99.3%
1	8	4,562	440,681	99.0%
2	7	4,536	404,403	98.6%
3	8	4,518	393,574	98.3%
4	9	4,502	393,194	98.2%
5	12	4,472	390,845	97.7%
6	7	4,458	383,033	97.3%
7	8	4,439	371,832	96.7%
8	7	4,478	369,280	97.6%
9	7	4,535	380,304	98.7%
10	9	4,571	388,498	99.3%
11	8	4,565	388,103	99.3%
12	7	4,567	388,977	99.5%
13	7	4,561	383,779	99.5%
14	8	4,543	379,595	99.5%
15	8	4,461	340,269	98.2%
16	8	4,358	327,895	96.0%
17	7	4,474	343,059	97.7%
Total	**141**			

Table S4: **Top 10 causally specific SAE features at layer 11.** ΔTarget and ΔOther = mean logit change at target and off-target gene positions, respectively, upon zeroing the feature. Specificity ratios were computed from unrounded values; displayed Δ values are rounded to three decimal places.

Feature	Annotation	Specificity	ΔTarget	ΔOther
F2035	Cell Differentiation (neg. reg.)	114.5×	−0.208	+0.002
F3692	ERAD Pathway	108.1×	−0.129	−0.001
F3933	Intracellular Signaling (neg. reg.)	55.7×	−0.196	−0.004
F157	Golgi Vesicle Transport	25.4×	−0.056	−0.002
F3532	Protein Metabolic Process (pos. reg.)	11.2×	−0.127	−0.011
F4516	Mitotic Spindle Microtubules	10.6×	+0.672	+0.063
F1337	Cell Cycle Phase Transition	9.4×	−0.058	−0.006
F1023	Mitotic Spindle Microtubules	7.6×	−2.799	−0.367
F2936	Mitochondrion Organization	7.1×	−0.366	−0.051
F3962	Endocytosis	6.9×	−0.099	−0.014

Table S5: **Geneformer cross-layer information highways.** PMI between SAE feature activations at source and target layers (500K positions each). A highway = source feature with ≥1 target-layer feature at PMI > 3.

Layer Pair	Feats w/ Deps	Mean Max PMI	Median Max PMI	Max PMI	Highways
L0 → L5	4,604	6.61	6.72	11.10	4,530 (98.4%)
L5 → L11	4,518	6.63	6.71	10.87	4,401 (97.4%)
L11 → L17	4,555	6.79	6.86	10.66	4,544 (99.8%)

Table S6: **scGPT cross-layer information highways.** Same methodology as Table S5. Note the progressive drop in downstream connectivity.

Layer Pair	PMI Edges	Upstream	Downstream	Max PMI
L0 → L4	75,305	1,935/2,027 (95.5%)	1,960/2,048 (95.7%)	9.15
L4 → L8	61,263	1,955/2,048 (95.5%)	1,723/2,048 (84.1%)	9.26
L8 → L11	45,258	1,773/2,048 (86.6%)	1,289/2,048 (62.9%)	10.78

Table S7: **Top cross-layer biological cascades.** Strongest annotated PMI connections between layer pairs. "Unlabeled" = the target feature lacks direct ontology annotation.

Pair	Source Feature	Target Feature	Biological Logic	PMI
L0→L5	Protein Processing in ER	*unlabeled*	ER stress cascade	11.10
L0→L5	mTORC1 Regulation	Autophagy	mTORC1→autophagy	9.55
L0→L5	Wnt Signaling	*unlabeled*	Wnt pathway processing	9.48
L5→L11	Protein Polyubiq.	*unlabeled*	Protein quality control	10.87
L5→L11	Translation	*unlabeled*	Translational regulation	10.35
L5→L11	RNA Splicing (neg. reg.)	*unlabeled*	Post-transcriptional	10.21
L11→L17	Protein Modification	Angiogenesis (pos. reg.)	PTM→phenotype	10.62
L11→L17	COPII Vesicle Budding	Thermogenesis	Secretory→metabolic	10.29
L11→L17	Actomyosin Org.	Cell Locomotion (neg. reg.)	Structure→motility	10.14

Table S8: **Per-TF head-to-head comparison at layer 11.** Only TFs with changed specificity are shown; 40 additional TFs had no specific features in either condition.

TF	K562-SAE Specific	MT-SAE Specific	Change
ATF5	0	1	gained
BRCA1	0	1	gained
GATA1	0	1	gained
RBMX	0	1	gained
NFRKB	0	1	gained
MAX	1	0	lost
PHB2	1	0	lost
SRF	1	0	lost

Table S9: **TF feature diagnostics.** Features with known TFs in top-20 genes and TF-dominant features ($\geq$3 TFs in top genes). Denominators = features with non-empty top-20 gene lists (may differ from alive counts because some features that are "dead" on the 100K held-out sample still produce gene lists from the full training data). K562-only SAE has more TF-associated features than the multi-tissue SAE.

SAE	Features with TFs in top genes	TF-dominant ($\geq$3 TFs)
K562-only L11	2,967/4,598 (64.5%)	424
Multi-tissue L0	2,796/4,608 (60.7%)	452
Multi-tissue L5	2,777/4,568 (60.8%)	337
Multi-tissue L11	2,782/4,601 (60.5%)	343
Multi-tissue L17	2,680/4,603 (58.2%)	346

Table S10: **Unannotated feature analysis.** Co-activate = unannotated feature belongs to a co-activation module containing annotated features. Isolated = no module membership. A small number of features (3 at L11, 17 at L17) belong to modules containing only unannotated features and are excluded from both columns. Clusters = standalone gene-set clusters among unannotated features.

Layer	Annotated	Unannotated	Clusters	Co-activate w/ Annotated	Isolated
0	2,702	1,906	15 (48 feats)	1,876 (98.4%)	30 (1.6%)
5	2,383	2,193	19 (69 feats)	2,090 (95.3%)	103 (4.7%)
11	2,583	2,015	11 (47 feats)	1,984 (98.5%)	28 (1.4%)
17	2,154	2,426	12 (58 feats)	2,334 (96.2%)	75 (3.1%)

2 Supplementary Figures

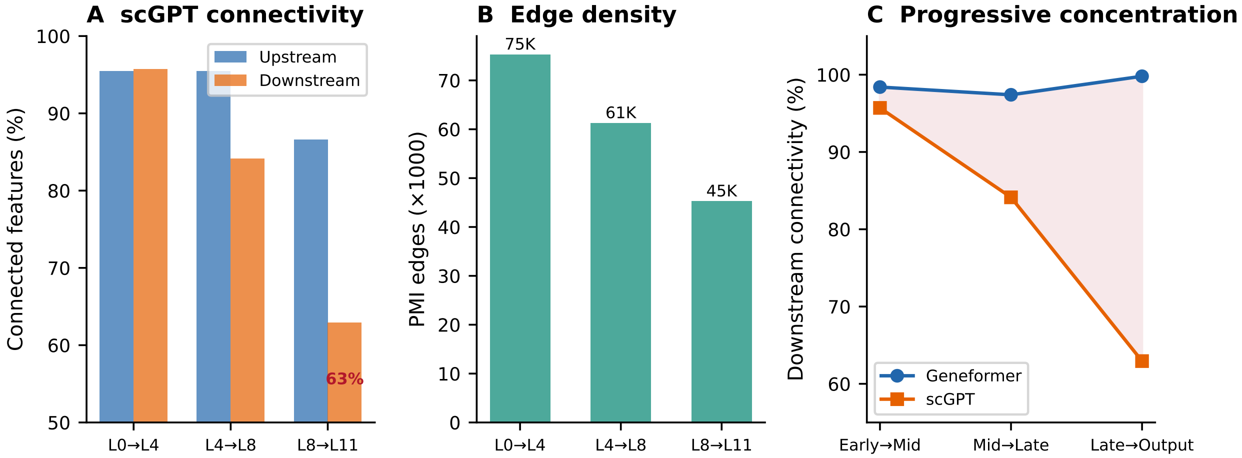

Figure S1: **scGPT cross-layer connectivity reveals progressive information concentration.** **(A)** Upstream connectivity remains high (86–96%) but downstream connectivity drops sharply from 96% to 63% across layer pairs, indicating progressive bottlenecking. **(B)** PMI edge density decreases from 75K to 45K edges across the three layer pairs. **(C)** Comparison with Geneformer: Geneformer maintains near-complete downstream connectivity (97–100%) while scGPT drops to 63%, suggesting fundamentally different information flow architectures.

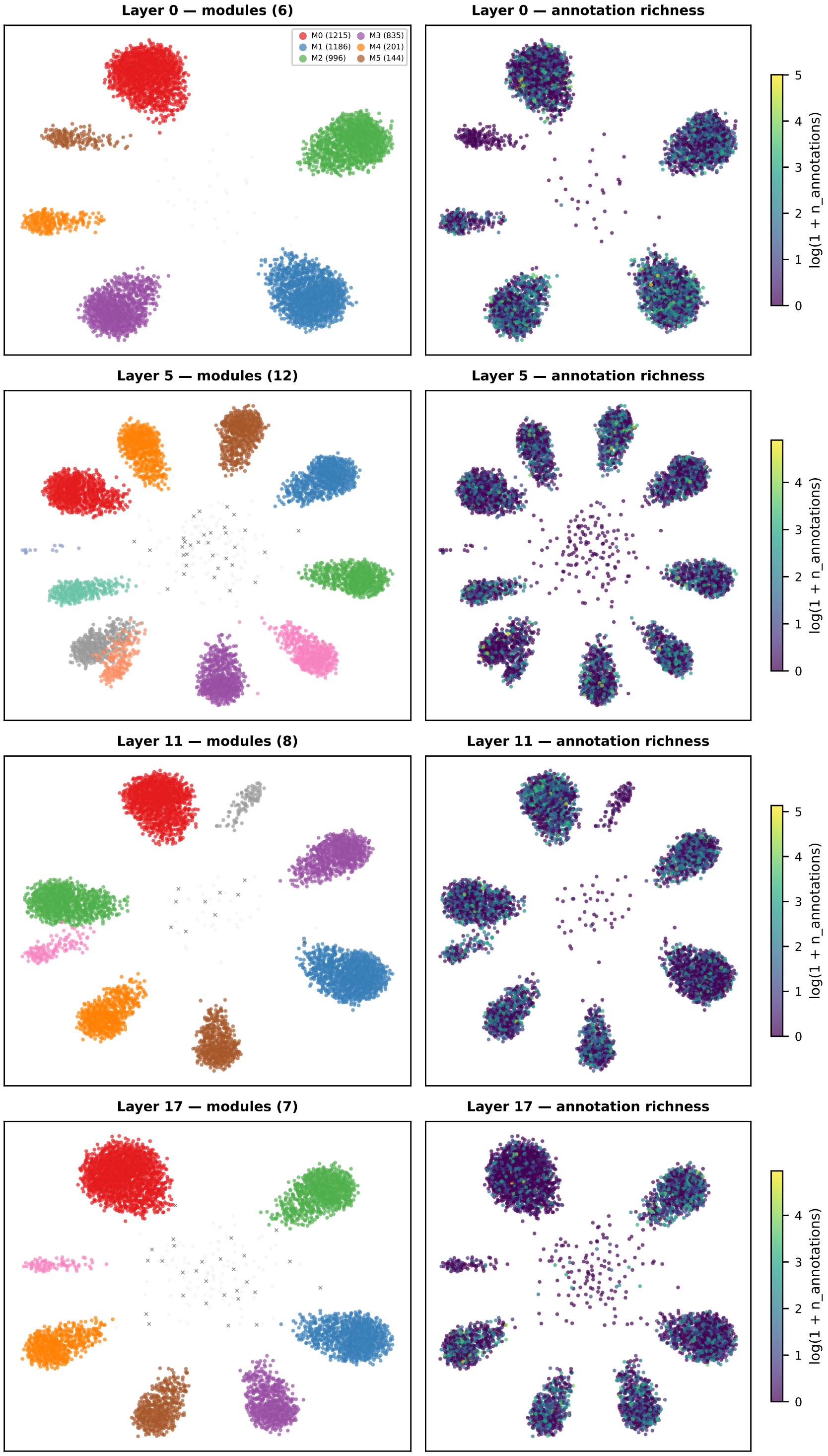

Figure S2: **Co-activation network layout of SAE features across layers.** Left column: force-directed graph layout of intra-module co-activation edges, colored by Leiden module membership. Each module forms a spatially distinct community. Right column: same layouts colored by annotation richness (log-transformed number of significant enrichment terms). Unassigned features (gray) cluster centrally. Module count varies from 6 (L0) to 12 (L5), reflecting the complexity of co-activation patterns at different layers.

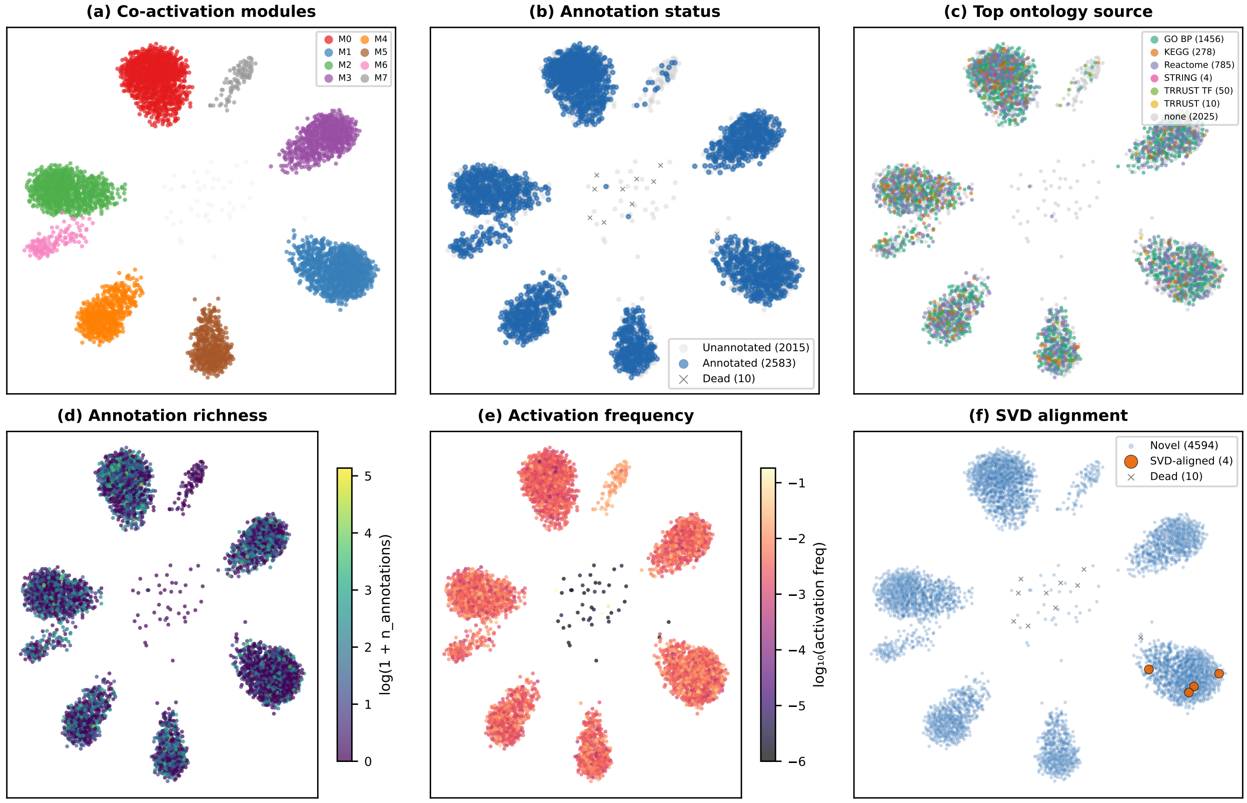

Figure S3: **Six-panel co-activation layout of layer 11 SAE features.** **(a)** Eight Leiden modules form spatially distinct communities. **(b)** Annotated features (blue) distribute across all module clusters; unannotated features (gray) concentrate centrally. **(c)** Top ontology source reveals module-specific enrichment patterns: certain modules are dominated by GO BP (green), others by STRING interactions (pink) or Reactome pathways (purple). **(d)** Annotation richness gradient across modules. **(e)** Activation frequency varies systematically across modules. **(f)** SVD-aligned features (orange, $n = 4$) are scattered across different modules, while 4,594 novel features (blue) fill the landscape.

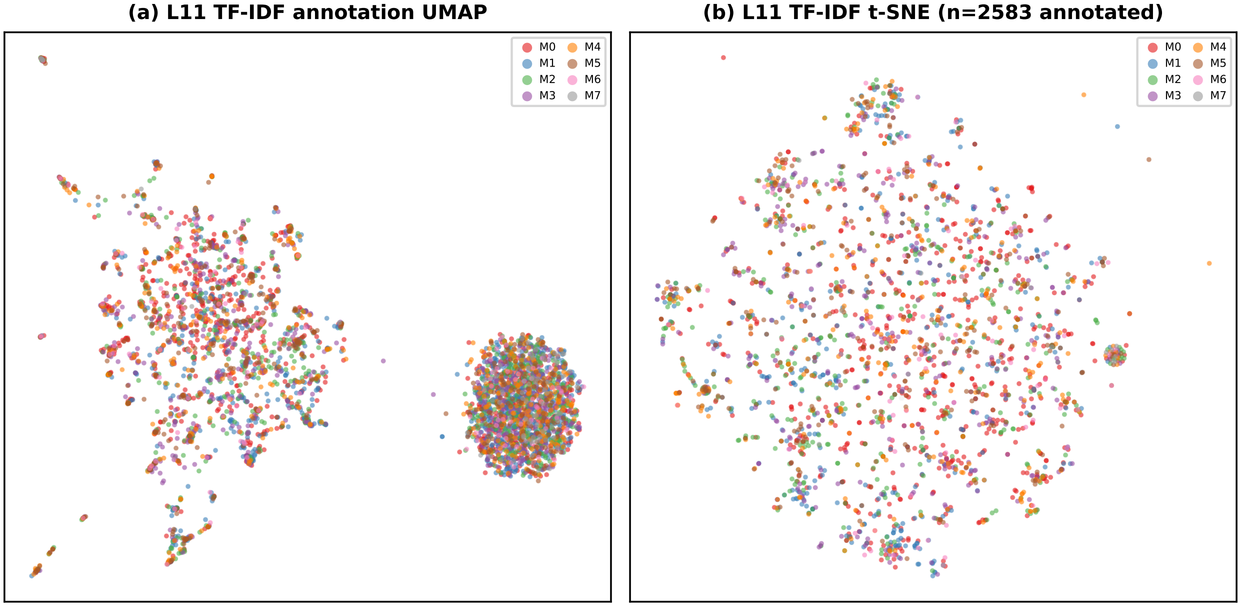

Figure S4: **Annotation-based projections provide independent validation. (a)** UMAP of TF-IDF weighted ontology term vectors for layer 11 (4,608 features). Annotated features (left cluster) separate from unannotated features (right blob), with internal structure reflecting shared biological annotations. **(b)** t-SNE of TF-IDF annotation vectors for annotated features only (n = 2,583). Fine-grained subclusters partially correspond to co-activation modules, confirming that module structure reflects genuine biological similarity.