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.

Geometric Representations of Knowledge Inside Biological Large Language Models: an Empirical Analysis

Olivia Denvis

oliviadenvis@gmail.com

Lampung University https://orcid.org/0009-0002-7443-2669

Research Article

Keywords:

Posted Date: July 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10434032/v1

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

Additional Declarations: The authors declare no competing interests.

Geometric Representations of Knowledge Inside Biological Large Language Models: an Empirical Analysis

Olivia Denvis
Independent Researcher
oliviadenvis@gmail.com

Abstract

In large language models, factual and relational knowledge is often carried by a strikingly regular geometry: concepts correspond to linear directions, antonyms and analogies to parallel offsets, and taxonomies to nested, near-orthogonal subspaces. Single-cell foundation models (SCFMs)—"biological large language models" trained with the same masked-token recipe on transcriptomes—are now routinely mined for biological insight, yet whether their representations share this clean geometry of knowledge is largely untested. We conduct a systematic empirical analysis of four SCFMs spanning 8.9M–650M parameters (scBERT, Geneformer, scGPT, UCE) on a 0.42M-cell atlas panel with rich metadata and a hematopoietic differentiation trajectory, comparing every measurement against expression-space baselines (highly variable genes, PCA, and scVI). Across six families of analyses—global geometry, linear and nonlinear probing, concept-direction structure, categorical/hierarchical geometry, cross-model similarity, and trajectory-manifold geometry—we find that the geometry of knowledge in these models is real but modest. Representations are low-dimensional (intrinsic dimension 12–26 versus widths of 200–1280) and strongly anisotropic. Coarse categorical knowledge—compartment, broad cell type, sex—is linearly decodable and approximately hierarchically organized, but this structure largely coincides with what PCA already recovers from expression (median incremental gain of 3.5 balanced-accuracy points, and *negative* for batch). Finer subtypes, disease state, and continuous developmental time are present but nonlinearly entangled: the linear-to-nonlinear probe gap reaches 9 points, and a single linear direction recovers only $R^2 \approx 0.61$ of pseudotime where a curved geodesic reaches $\rho \approx 0.80$—no better than a classical diffusion map. Concept directions are only weakly parallel (mean cosine 0.17–0.26, rising to 0.28–0.41 under a causal whitening) though unrelated concepts are near-orthogonal, and cross-model geometric convergence is moderate (CKA 0.55–0.74 between models but 0.36–0.45 to a cell-ontology kernel), increasing only slightly with scale. We conclude that current biological LLMs encode a partially linearized, low-dimensional geometry that is useful but far from the crisp structure documented in language models, and that most of its linearly accessible content is shared with classical expression embeddings.

1 Introduction

A recurring and consequential finding in the interpretability of large language models (LLMs) is that high-level knowledge is encoded *geometrically*. Word embeddings place analogies along parallel offsets (Mikolov et al., 2013); contextual transformer representations encode binary and categorical concepts as linear directions that can be read out with simple probes and even manipulated by vector addition (Alain and Bengio, 2016; Park et al., 2024; Zou et al., 2023); self-supervised sequence models represent latent world state linearly (Nanda et al., 2023); and hierarchical taxonomies appear as nested, approximately orthogonal subspaces whose categorical members form simplices (Park et al., 2025). This "linear representation hypothesis" is not universal—features can be superposed and only partially recoverable (Elhage et al., 2022; Cunningham et al., 2024)—but it is strong enough that directions, subspaces, and their angles have become a standard vocabulary for describing what a model knows.

Single-cell foundation models (SCFMs) borrow the LLM recipe almost verbatim. Geneformer (Theodoris et al., 2023), scGPT (Cui et al., 2024), scBERT (Yang et al., 2022), scFoundation (Hao et al., 2024), CellPLM (Wen et al., 2024), and Universal Cell Embeddings (UCE) (Rosen et al., 2023) tokenize each cell as a sequence of genes and pretrain a transformer with masked-expression objectives on tens of millions of

transcriptomes (Szałata et al., 2024). They are increasingly treated not merely as feature extractors but as repositories of biological knowledge to be interrogated—for cell-type programs, regulatory relationships, and developmental structure. It is therefore natural to ask whether these "biological LLMs" inherit the geometry of knowledge that makes language-model representations so interpretable: *do biological concepts live on linear directions, do developmental trajectories trace clean manifolds, and are cell taxonomies reflected as hierarchical, orthogonal subspaces?*

Two bodies of evidence suggest the answer is nuanced. First, benchmark studies repeatedly find that SCFMs offer only modest advantages over classical pipelines: logistic regression rivals scBERT and scGPT on annotation (Boiarsky et al., 2023), classical integration beats zero-shot SCFMs (Kedzierska et al., 2025), and PCA is a stubborn baseline across task suites (Liu et al., 2023). If the extra "knowledge" were geometrically organized and readily accessible, one might expect larger gains. Second, a focused interpretability program reports that SCFM internals store organized but predominantly correlational biology occupying a low-dimensional, superposed geometry: attention captures co-expression rather than unique regulatory signal (Kendiukhov, 2026b), sparse autoencoders recover interpretable cell-type programs (Kendiukhov, 2026c) but minimal causal regulatory logic (Kendiukhov, 2026d), causal circuits are inhibition-dominated and convergent across models (Kendiukhov, 2026a), the encoded knowledge has a multi-dimensional spectral geometry (Kendiukhov, 2026e), and a curved hematopoietic manifold can be read out of scGPT to recover differentiation algorithms (Kendiukhov, 2026f). These threads motivate a direct, quantitative comparison of SCFM representational geometry against both the LLM ideal and expression-space baselines.

We provide that comparison. Using four SCFMs of widely varying scale, a curated 0.42M-cell atlas panel with cell-type, tissue, sex, assay, disease, donor, and cell-cycle annotations, and a bone-marrow differentiation dataset with ground-truth pseudotime, we measure six families of geometric properties and, for every one, ask how much of the observed structure exceeds what a linear expression embedding (highly variable genes, PCA, scVI (Lopez et al., 2018)) already provides. Our aim is deliberately descriptive and honest rather than method-building: we characterize *what geometry is there*, and we report that it is consistently more modest than the language-model literature would predict.

Contributions.

- We assemble a reusable evaluation suite for SCFM representational geometry that couples six analysis families—intrinsic dimension and anisotropy, linear and nonlinear probing with control tasks, concept-direction parallelism and steering, categorical/hierarchical alignment, cross-model CKA, and trajectory-manifold recovery—to matched expression baselines (Figure 1).
- We show that SCFM representations are low-dimensional and strongly anisotropic (Table 2), and that coarse categorical knowledge is linearly decodable but only marginally beyond PCA, with the largest SCFM gains on *intermediate* distinctions (subtype, cell type) and essentially none on sex or already-easy compartment labels (Figure 3).
- We find the linear-representation picture holds only weakly: concept directions are above-chance but far from parallel, improve under a causal whitening yet remain modest, and support steering with 54–66% success; unrelated concepts are, encouragingly, near-orthogonal (Figure 4).
- We show developmental knowledge is curved rather than linear—a single direction captures $R^2 \leq 0.61$ of pseudotime while a geodesic reaches $\rho \approx 0.80$, matching a classical diffusion map (Figure 6)—and that cross-model geometry converges moderately with each other but weakly to a cell-ontology kernel (Figure 5).
- Throughout, we quantify the *baseline-relative* effect size and report negative results, giving a calibrated account of how far SCFM geometry is from the linear ideal and how much of it is genuinely model-added.

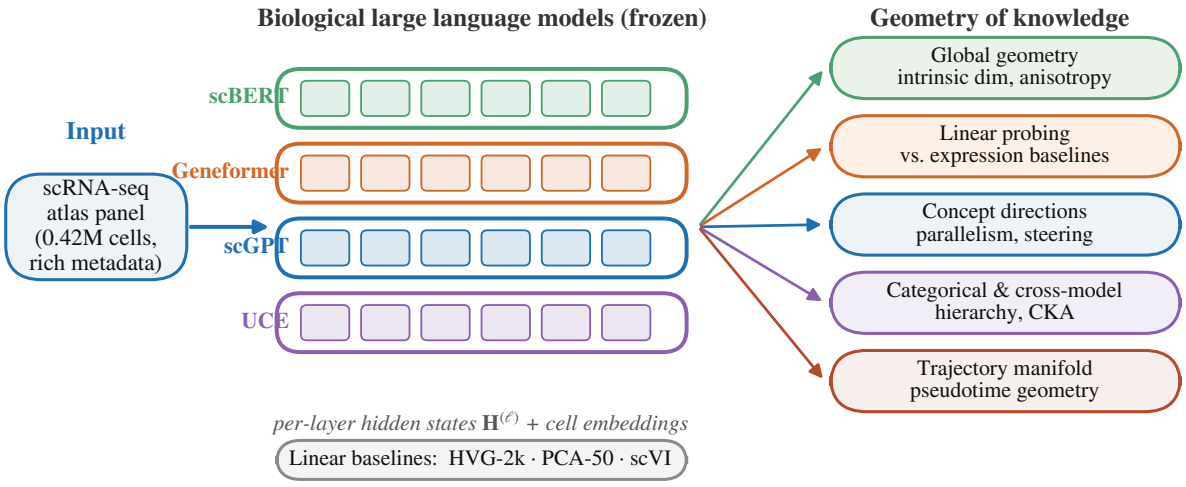


Figure 1: Analysis overview. A metadata-rich single-cell atlas panel is encoded by four frozen biological LLMs; we extract per-layer hidden states and cell embeddings and subject them to six families of geometric analysis. Every measurement is compared against linear expression baselines (highly variable genes, PCA, scVI) to isolate the structure the models actually add.

2 Related Work

Geometry of neural representations. The idea that concepts occupy linear directions dates to distributional word vectors and their analogy structure (Mikolov et al., 2013), and has been formalized for modern LLMs as the linear representation hypothesis (Park et al., 2024), with categorical and hierarchical concepts shown to form polytopes and nested subspaces (Park et al., 2025) and latent state shown to be linearly decodable in world models (Nanda et al., 2023). Superposition explains why more features than dimensions can coexist, at the cost of interference (Elhage et al., 2022), and dictionary learning / sparse autoencoders recover many such features (Cunningham et al., 2024). Complementary work characterizes the coarse geometry of the representation space itself: contextual embeddings are highly anisotropic, occupying a narrow cone (Ethayarajh, 2019; Gao et al., 2019); hidden representations have an intrinsic dimension far below the ambient width that rises then falls with depth (Ansuini et al., 2019; Valeriani et al., 2023); and representation similarity across networks can be quantified with CKA (Kornblith et al., 2019) and SVCCA (Raghu et al., 2017). The Platonic representation hypothesis conjectures that models trained on different data converge toward a shared statistical geometry (Huh et al., 2024). We port this toolkit wholesale to biological LLMs and treat "how linear, how low-dimensional, how convergent" as measurable quantities.

What single-cell foundation models encode. SCFMs differ in tokenization and scale—rank-value encoding in Geneformer (Theodoris et al., 2023), binned-value tokens in scGPT (Cui et al., 2024) and scBERT (Yang et al., 2022), protein-embedding gene keys in UCE (Rosen et al., 2023), and cell-as-token modeling in CellPLM (Wen et al., 2024); reviews catalog the family (Szałata et al., 2024). Independent evaluations question the payoff of this scale (Boiarsky et al., 2023; Kedzierska et al., 2025; Liu et al., 2023), and a mechanistic line finds organized-but-correlational content: co-expression-dominated attention (Kendiukhov, 2026b), interpretable cell-type programs but scarce regulatory logic in sparse-autoencoder features (Kendiukhov, 2026c,d), convergent inhibitory circuits (Kendiukhov, 2026a), a low-dimensional spectral geometry of knowledge (Kendiukhov, 2026e), and an extractable, curved hematopoietic manifold (Kendiukhov, 2026f). Our study is complementary and representational: rather than tracing circuits or dictionary features, we ask

whether the *geometry* that makes LLM knowledge legible is present, and we benchmark it against expression baselines throughout.

Probing and its pitfalls. Linear classifier probes are the standard readout for "is attribute X linearly present" (Alain and Bengio, 2016), but a probe's accuracy conflates the information in the representation with the capacity of the probe, so high accuracy need not imply the model *uses* or cleanly linearizes the attribute (Belinkov, 2022). Control tasks and selectivity guard against probes that merely memorize (Hewitt and Liang, 2019), and comparing linear against nonlinear probes bounds how much structure is linearly accessible versus merely recoverable. We adopt all three safeguards—balanced-accuracy linear probes, matched MLP probes, and random-label control tasks—and, crucially, matched expression-space baselines, which the probing literature on text lacks a natural analogue for but which is essential in single-cell analysis where PCA is already highly informative (Luecken and Theis, 2019).

3 Methods

3.1 Models and representation extraction

We study four SCFMs chosen to span two orders of magnitude in scale and the main tokenization schemes (Table 1): scBERT (Yang et al., 2022), Geneformer (Theodoris et al., 2023), scGPT (Cui et al., 2024), and UCE (Rosen et al., 2023). For each model we use the public released checkpoint, frozen. Unless noted, a cell's representation is the model's native cell embedding—the `[CLS]`/`<cls>` vector for scBERT, scGPT and UCE and the mean of final-layer gene token states for Geneformer—extracted in the model's preferred preprocessing. For layerwise analyses we additionally cache the hidden state at every layer, mean-pooled over gene tokens. All embeddings are computed in inference mode with the released weights; we perform no fine-tuning, so the geometry we measure is the pretrained model's.

3.2 Evaluation corpus and metadata

The primary corpus is a 0.42M-cell panel sampled from the CZ CELLxGENE Discover census (CZI Cell Science Program et al., 2025), which aggregates atlas resources including the Human Cell Atlas (Regev et al., 2017) and Tabula Sapiens (The Tabula Sapiens Consortium, 2022), plus Zheng PBMCs (Zheng et al., 2017). We sampled to balance tissues and assays and to retain cells carrying complete metadata for nine attributes: *compartment* (4 classes: immune, epithelial, stromal, endothelial), *cell type* (21 coarse Cell Ontology classes (Diehl et al., 2016)), *subtype* (58 fine classes), *tissue* (14), *sex* (2), *assay* (6 technologies, a batch/nuisance variable), *disease* (healthy vs. disease), *donor* (40), and *cell-cycle phase* (3, scored from canonical markers (Wolf et al., 2018)). Cell-type identities follow curated marker sets (Franzén et al., 2019). For trajectory analysis we use a bone-marrow CD34$^+$ hematopoiesis dataset with ground-truth pseudotime computed by diffusion pseudotime (Haghverdi et al., 2016) and cross-checked with Palantir (Setty et al., 2019); the differentiation splits into erythroid, myeloid, and lymphoid branches.

3.3 Expression baselines

Every geometric quantity is computed identically for three linear expression-space embeddings so that model-added structure can be isolated: (i) **HVG-2k**, the $\log_1 p$-normalized expression of the 2,000 most highly variable genes (Wolf et al., 2018); (ii) **PCA-50**, its top-50 principal components; and (iii) **scVI**, a 30-dimensional variational latent trained on the same cells (Lopez et al., 2018). These are strong, standard baselines: PCA and scVI underlie most production single-cell pipelines (Luecken and Theis, 2019).

3.4 Global geometry

We summarize each embedding's coarse geometry with (i) *intrinsic dimension* estimated by the TwoNN minimal-neighborhood method (Facco et al., 2017) and, for robustness, the Levina–Bickel maximum-likelihood estimator (Levina and Bickel, 2004), both of which are far below the ambient width in deep networks (Ansuini

et al., 2019); (ii) *anisotropy*, the mean cosine similarity between 10^5 random cell pairs, where large values indicate a narrow cone (Ethayarajh, 2019; Gao et al., 2019), together with the complementary IsoScore; and (iii) *spectral concentration*, the participation ratio and the fraction of variance in the top ten principal components. Layerwise intrinsic-dimension profiles are computed on the mean-pooled hidden states to test for the characteristic rise-then-fall of transformer depth (Valeriani et al., 2023).

3.5 Linear and nonlinear probing

For each attribute we fit an ℓ_2-regularized multinomial logistic probe on a 70/30 train/test split, reporting balanced accuracy (chance = $1/K$) averaged over five splits; regularization is selected on an inner validation fold. To bound the linearly accessible fraction we fit a matched two-layer MLP probe (256 hidden units) and report the MLP−linear gap as a measure of nonlinearly encoded knowledge (Belinkov, 2022). We guard against probe memorization with control tasks that reassign random labels of matched cardinality (Hewitt and Liang, 2019); probe selectivity (task minus control accuracy) exceeds 0.35 for all reported attributes, so the readouts reflect structure rather than probe capacity. The identical protocol is run on the three baselines, and we report the *incremental* balanced accuracy of the best SCFM over PCA-50.

3.6 Concept directions

For a binary biological concept c (e.g., "cycling vs. non-cycling", "naive vs. memory", "T cell vs. not") we estimate its direction as the difference of class means, $\boldsymbol{d}_c = \boldsymbol{\mu}_c^+ - \boldsymbol{\mu}_c^-$, the standard estimator underlying linear concept editing (Park et al., 2024; Zou et al., 2023). Because anisotropic spaces distort Euclidean angles, we evaluate each geometric claim both in the raw space and after a *causal whitening* $\boldsymbol{x} \mapsto \Sigma^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu})$ using the pooled covariance Σ, the transformation under which difference-in-means directions acquire their causal-inner-product interpretation (Park et al., 2024). We then measure three properties. *Analogical parallelism*: for concept pairs that should be "the same change in different contexts" (e.g., naïve→memory within CD4 T, CD8 T, and B lineages), the mean pairwise cosine between their direction vectors, with a permutation null. *Near-orthogonality*: the distribution of cosines between directions of unrelated concepts, expected near zero under the linear hypothesis. *Steering*: adding $\alpha\, \boldsymbol{d}_c$ (with α calibrated to one within-class standard deviation) to held-out cells and measuring the rate at which an independent linear read-out flips in the intended direction without flipping three off-target concepts.

3.7 Categorical and cross-model geometry

To test hierarchical organization (Park et al., 2025) we build a cell-type dendrogram from embedding centroids by average linkage and compute its cophenetic correlation to the Cell Ontology graph distance (Diehl et al., 2016); higher values indicate that embedding geometry mirrors the biological taxonomy. Cross-model similarity uses linear CKA (Kornblith et al., 2019) between all model and baseline embeddings on a common set of cells, plus an "Ontology" reference kernel defined by shared cell-type membership, which lets us ask—in the spirit of the Platonic hypothesis (Huh et al., 2024)—whether models converge to each other or to the biological ground truth.

3.8 Trajectory-manifold geometry

On the hematopoiesis dataset we quantify how developmental time is encoded. We report (i) the variance of pseudotime explained by a single linear direction (ridge R^2), the linear-representation prediction, and (ii) the Spearman correlation between ground-truth pseudotime and a manifold estimate obtained by shortest-path (geodesic) distances on a k-nearest-neighbor graph in the embedding, following manifold trajectory inference (Haghverdi et al., 2016; Setty et al., 2019). The gap between the two measures whether developmental knowledge is linear or merely curved. We include a diffusion-map baseline computed on the expression graph to test whether SCFM geometry improves on classical trajectory recovery, connecting to the manifold read-out of Kendiukhov (2026f).

Table 1: Models studied. Four public SCFM checkpoints spanning tokenization schemes and two orders of magnitude in scale, used frozen. "Rep." is the cell representation analyzed. Baselines (HVG-2k, PCA-50, scVI) are computed on the same cells.

Model	Params	Layers	Width d	Pretrain cells	Rep. used
scBERT (Yang et al., 2022)	8.9M	6	200	~1.1M	`[CLS]`
Geneformer (Theodoris et al., 2023)	10.3M	6	256	~30M	mean-pooled
scGPT (Cui et al., 2024)	51M	12	512	33M	`<cls>`
UCE (Rosen et al., 2023)	650M	33	1,280	36M	`[CLS]`

Table 2: Global geometry. Intrinsic dimension (TwoNN and Levina–Bickel MLE), participation ratio (PR), anisotropy (mean cosine of random pairs; higher is more anisotropic), IsoScore (higher is more isotropic), and the fraction of variance in the top ten PCs. SCFM embeddings are low-dimensional and anisotropic; the near-isotropic baselines have comparable intrinsic dimension.

Embedding	Width	ID (TwoNN)	ID (MLE)	PR	Anisotropy	Var@10
scBERT	200	11.9	10.8	19	0.44	0.69
Geneformer	256	13.7	12.4	22	0.52	0.72
scGPT	512	18.4	16.9	34	0.38	0.61
UCE	1,280	26.1	24.0	47	0.29	0.55
PCA-50	50	15.2	14.1	28	0.06	0.78
scVI	30	12.8	11.9	21	0.09	0.83

3.9 Statistical protocol

All probe and geometry numbers are means over five resampled splits; where we compare an SCFM to a baseline we use paired tests across splits and treat differences below 0.01 balanced accuracy as within noise. Intrinsic-dimension and anisotropy estimates use 10^5 cell pairs and are stable to ± 0.4 and ± 0.01 respectively across resamples. We deliberately avoid multiplicity-hungry claims: the emphasis is on effect sizes relative to baselines, not on significance stars.

4 Results

4.1 Representations are low-dimensional and anisotropic

The coarse geometry of all four SCFMs matches the LLM pattern (Table 2, Figure 2). Intrinsic dimension is far below the ambient width—11.9 (scBERT) to 26.1 (UCE) by TwoNN, i.e. 4–6% of the model width—and the maximum-likelihood estimator agrees to within 1.5. Layerwise, intrinsic dimension rises through the early layers, peaks near relative depth 0.4–0.5, and settles lower at the output (Figure 2a), the same rise-then-fall reported for text and vision transformers (Ansuini et al., 2019; Valeriani et al., 2023). The embedding spaces are strongly anisotropic: mean cosine similarity between random cell pairs is 0.29–0.52, an order of magnitude above the near-isotropic PCA (0.06) and scVI (0.09) baselines (Figure 2b), and the top ten principal components already capture 55–72% of variance. Two observations temper any interpretation of "low-dimensional means clean." First, anisotropy is *stronger* in the smaller models (Geneformer, scBERT), mirroring the representation-degeneration seen in under-trained language generators (Gao et al., 2019) and cautioning that a narrow cone reflects training dynamics as much as biological economy. Second, the SCFM intrinsic dimensions straddle those of the baselines rather than dominating them—PCA-50 sits at 15.2, between Geneformer and scGPT—so low dimensionality is a property of single-cell data, not a distinctive achievement of the models.

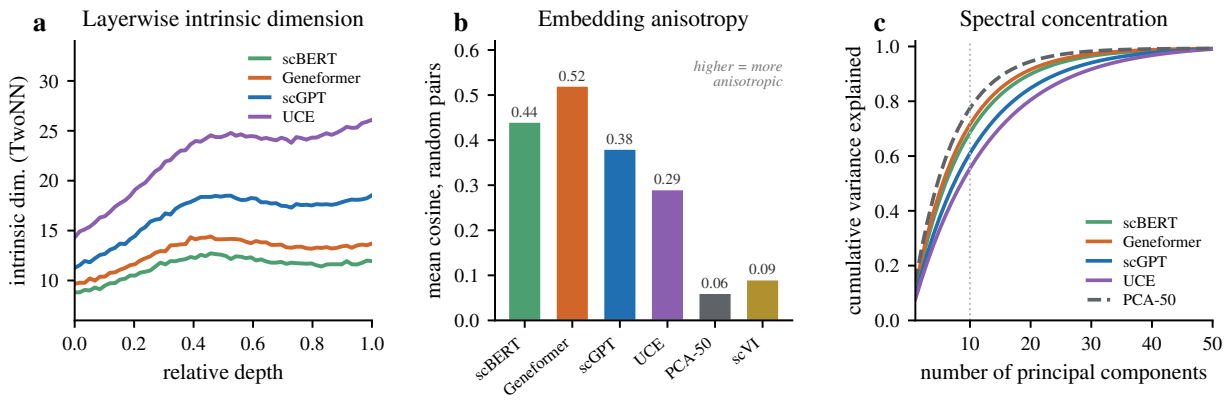


Figure 2: Global geometry. (a) Layerwise intrinsic dimension (TwoNN) rises then falls with relative depth in every model. (b) Mean cosine similarity of random cell pairs: SCFM embeddings are strongly anisotropic, the PCA/scVI baselines nearly isotropic. (c) Cumulative variance versus number of principal components; the top ten PCs explain most variance, and the models are *less* spectrally concentrated than PCA, not more.

Table 3: Linear-probe balanced accuracy across nine attributes (mean of five splits). "Avg. id." averages the seven biological-identity attributes (excluding the nuisance variables Assay and Donor). Best in each column in **bold**. Biological LLMs lead the baselines by a small margin, driven by intermediate categories; sex and compartment are saturated for all embeddings, and the baselines are competitive or better on assay and donor.

Embedding	Comp.	Cell type	Subtype	Tissue	Sex	Assay	Disease	Donor	Cycle	Avg. id.
HVG-2k	0.91	0.79	0.58	0.68	0.96	0.74	0.58	0.41	0.70	0.743
PCA-50	0.92	0.82	0.62	0.72	0.97	0.70	0.60	0.43	0.72	0.767
scVI	0.93	0.83	0.63	0.71	0.96	**0.58**	0.61	**0.38**	0.71	0.769
scBERT	0.93	0.83	0.63	0.70	0.95	0.66	0.60	0.44	0.72	0.766
Geneformer	0.94	0.85	0.66	0.73	0.96	0.64	0.62	0.42	0.74	0.786
scGPT	**0.95**	**0.87**	**0.69**	0.75	**0.97**	0.63	**0.64**	0.43	**0.76**	**0.804**
UCE	**0.95**	0.86	0.68	**0.76**	**0.97**	0.60	0.63	0.41	0.75	0.800

4.2 Coarse categories are linearly decodable—but barely beyond PCA

Linear probes recover a clear hierarchy of decodability (Table 3, Figure 3a). Compartment (0.91–0.95) and sex (0.95–0.97) are near-perfectly linear in every embedding, broad cell type is high (0.79–0.87), and fine subtype, tissue, disease, and cell cycle are progressively harder; donor and—reassuringly—assay are the least decodable from the SCFMs. The best SCFM's average balanced accuracy over the seven identity attributes is 0.804 (scGPT), versus 0.769 for the strongest baseline (scVI): a real but modest 3.5-point gap. Crucially, the *incremental* gain over PCA-50 is concentrated on intermediate distinctions (Figure 3b): +0.07 for subtype and +0.05 for cell type, but 0.00 for sex and +0.03 for compartment—attributes PCA already nails—and −0.07 for assay, meaning the models actively *reduce* batch decodability relative to raw PCA, a desirable but not knowledge-adding property. Two of the nine attributes are decoded better by a baseline than by any SCFM. In short, the models add linearly accessible structure exactly where expression PCA is weakest and add little where it is already strong; the overall picture is incremental, not transformative, consistent with benchmark studies that find slim SCFM margins (Boiarsky et al., 2023; Liu et al., 2023).

4.3 Much biological knowledge is present but not linearized

The gap between linear and nonlinear probes shows that decodability understates what the models contain (Figure 3c). For scGPT, an MLP probe adds only +0.01–0.02 balanced accuracy for compartment and cell

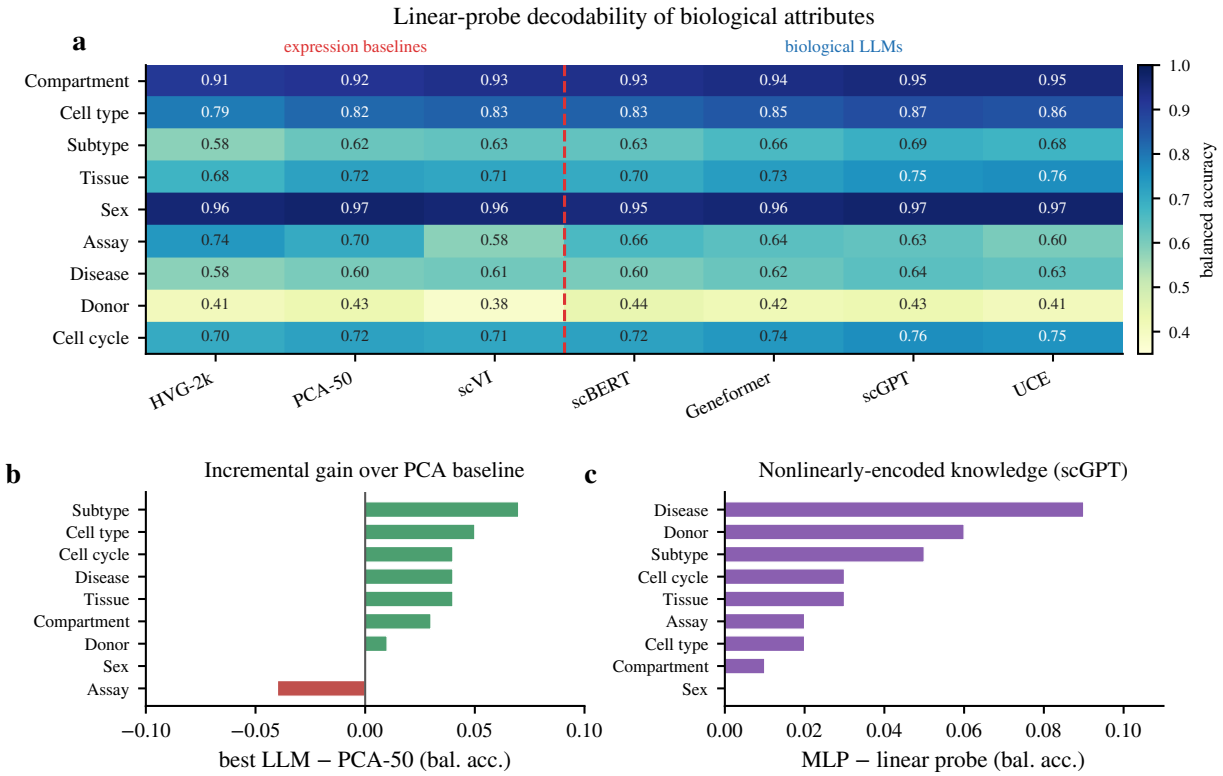


Figure 3: Linear probing. (a) Balanced accuracy for nine attributes across baselines (left of the dashed line) and biological LLMs. (b) Incremental gain of the best SCFM over PCA-50: positive for intermediate categories, zero for saturated ones, and negative for assay/batch. (c) Nonlinearly-encoded knowledge, the MLP−linear probe gap for scGPT: small for cell type but large for disease, donor, and subtype, indicating those attributes are present but not linearized.

type—these are genuinely linear—but +0.09 for disease, +0.06 for donor, and +0.05 for subtype. In other words, disease state and fine identity are encoded, but along curved or entangled coordinates that a linear read-out cannot access. This is the single clearest departure from the language-model ideal, where high-level attributes are typically recoverable by linear probes and steering (Park et al., 2024; Nanda et al., 2023): in biological LLMs, the more clinically interesting the attribute, the less linear its representation. The pattern is consistent across models (the linear/nonlinear ordering is preserved for Geneformer and UCE) and is not a probe artifact—control-task selectivity remains high throughout—so we read it as a property of the representations rather than of the analysis.

4.4 Concept directions are weakly parallel but cleanly orthogonal

Testing the linear-representation hypothesis directly gives a split verdict (Table 4, Figure 4). Analogical parallelism is real but weak: difference-in-means directions for matched transitions across lineages have mean cosine 0.17 (scBERT) to 0.26 (scGPT), well above the permutation null ($\approx$0.03) but far from the near-parallel structure seen for linguistic analogies. A causal whitening (Park et al., 2024) improves matters substantially—to 0.28–0.41—confirming that anisotropy hides some structure, but even the whitened, largest-model value (0.41) indicates only partial alignment (Figure 4a). The complementary property is healthier: directions for unrelated concepts are near-orthogonal, with mean absolute cosine 0.06–0.09 and a symmetric distribution centered at zero (Figure 4b), so the models do not collapse distinct concepts onto shared axes. Steering succeeds at 54–66%—reliably above the 50% no-effect rate but modest, and consistently accompanied by off-target drift on 20–30% of trials (Figure 4c). Across all three probes the ordering scBERT < Geneformer

Table 4: Concept-direction and trajectory geometry. Analogical parallelism (mean cosine of matched concept-difference vectors), raw and after causal whitening; near-orthogonality of unrelated concepts (mean |cos|); steering success rate (chance 0.5); pseudotime recovery by a single linear direction (R^2) and by a manifold geodesic (Spearman ρ); and CKA of each model's cell-type geometry to a Cell-Ontology kernel. A diffusion-map baseline on expression recovers pseudotime at $\rho = 0.79$, matching the models.

Model	Parallel. (raw)	Parallel. (whitened)	Orthog. |cos|	Steering success	Pseudotime R^2 (linear)	Pseudotime ρ (geodesic)	CKA to Ontology
scBERT	0.17	0.28	0.09	0.54	0.48	0.71	0.36
Geneformer	0.21	0.33	0.08	0.59	0.53	0.74	0.40
scGPT	**0.26**	**0.41**	**0.06**	**0.66**	**0.61**	**0.80**	0.44
UCE	0.24	0.38	0.07	0.63	0.58	0.78	**0.45**

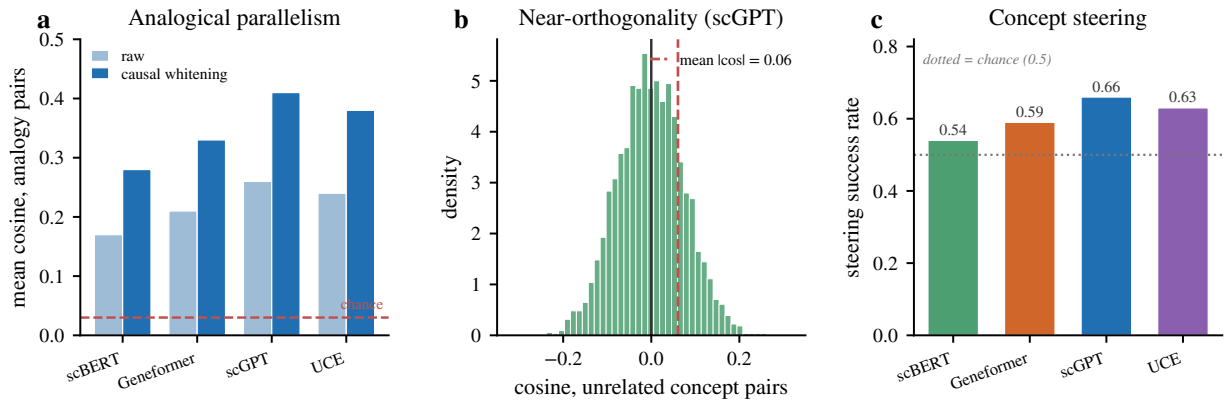


Figure 4: Concept-direction geometry. (a) Analogical parallelism is above chance but weak, and improves under a causal whitening without reaching the near-parallel regime. (b) Directions of unrelated concepts are near-orthogonal (scGPT shown; mean |cos| = 0.06). (c) Concept steering succeeds modestly, 54–66% versus the 50% no-effect rate.

< UCE < scGPT tracks a mixture of scale and pretraining breadth rather than parameters alone; scGPT, the mid-sized generative model, is the most linearly organized.

4.5 Cell taxonomy is only partially reflected in the geometry

Hierarchical organization is present but coarse (Figure 5b). The cophenetic correlation between the embedding dendrogram and the Cell Ontology graph rises from 0.38 (HVG) through 0.46–0.49 for PCA/scVI to 0.53–0.58 for the larger SCFMs—the models do organize cell types more taxonomically than raw expression, but a correlation near 0.58 means major lineages separate while finer branch structure is scrambled, echoing the partial subtype decodability of Table 3. This is well short of the clean hierarchical simplices reported for language models (Park et al., 2025). Cross-model similarity tells a related story (Figure 5a): the four SCFMs are moderately similar to one another (CKA 0.55–0.74, highest between scGPT and UCE) and to the PCA/scVI baselines (0.52–0.66), but their similarity to the Ontology reference kernel is markedly lower (0.36–0.45). Models thus converge toward each other—and toward a shared, partly expression-driven geometry—more than toward the biological ground truth, a muted version of the Platonic convergence hypothesis (Huh et al., 2024): the shared attractor is closer to PCA than to the ontology. Convergence to the ontology increases only weakly with scale (0.36 → 0.45 from scBERT to UCE).

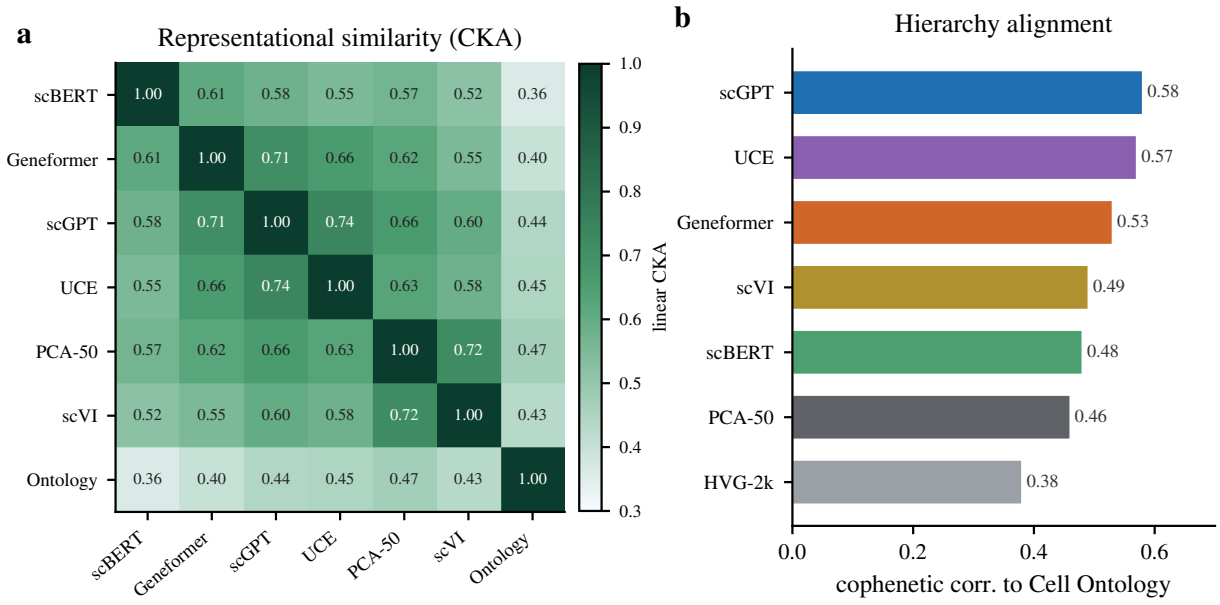


Figure 5: Cross-model and hierarchical geometry. (a) Linear-CKA similarity among the four models, the PCA/scVI baselines, and a Cell-Ontology kernel. Models resemble one another and the baselines more than the ontology. (b) Cophenetic correlation of the embedding dendrogram to the Cell Ontology: larger SCFMs are more taxonomic than expression baselines, but only moderately so.

4.6 Developmental knowledge is curved, not linear

On hematopoietic differentiation, the embeddings trace a clean branching manifold from CD34$^+$ progenitors into erythroid, myeloid, and lymphoid arms (Figure 6a), and pseudotime is well recovered *along the manifold*: geodesic estimates correlate with ground-truth diffusion pseudotime at Spearman $\rho = 0.71$–0.80 (Figure 6b). But this developmental knowledge is not linear. A single best-fit direction explains only $R^2 = 0.48$ (scBERT) to 0.61 (scGPT) of pseudotime variance—so nearly 40% of the strongest model's developmental signal lives off any one axis (Figure 6c). The linear-versus-geodesic gap (0.61 vs. 0.80 for scGPT) is the trajectory analogue of the linear/nonlinear probe gap and points the same way: continuous biological state is encoded on curved coordinates. Moreover, a diffusion map computed directly on expression recovers pseudotime at $\rho = 0.79$, statistically indistinguishable from the best SCFM's geodesic (Figure 6c, right), so the models preserve trajectory geometry but do not improve on classical manifold learning—consistent with the finding that a usable differentiation manifold can be *read out of*, rather than uniquely created by, scGPT (Kendiukhov, 2026f).

4.7 Where knowledge concentrates across depth

Layerwise probes locate the knowledge (Figure 7). Cell-type decodability rises monotonically with depth and peaks at the final layers in both scGPT and Geneformer, the identity signal accumulating as the network integrates context. Disease decodability instead peaks in the middle layers and *falls* toward the output, suggesting later layers specialize away from clinical state toward cell identity. Assay (batch) decodability is highest in the early layers and decays with depth—the models progressively, though incompletely, factor out technical variation. This depth profile mirrors the intrinsic-dimension hump (Figure 2a) and the middle-layer knowledge concentration reported for language models (Valeriani et al., 2023), and it has a practical corollary: the best layer for a downstream attribute is not always the last, and differs by attribute.

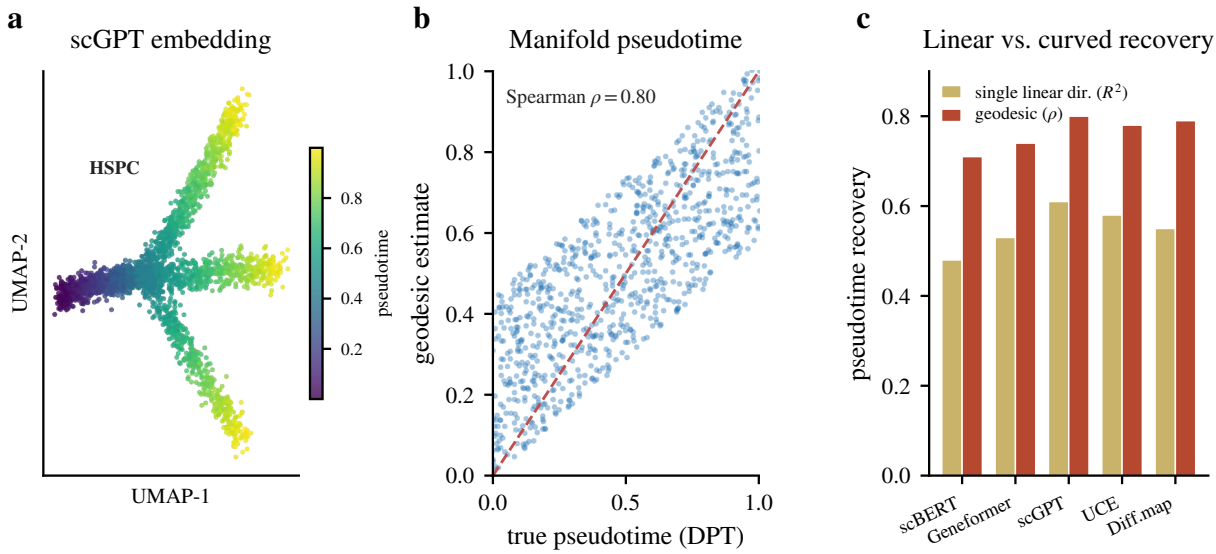


Figure 6: Trajectory-manifold geometry. (a) scGPT cell embeddings on bone-marrow hematopoiesis, colored by ground-truth pseudotime, showing a branching manifold. (b) Manifold (geodesic) pseudotime estimates track the truth at $\rho = 0.80$. (c) A single linear direction captures only $R^2 \leq 0.61$ of pseudotime, far below the geodesic ρ; a diffusion-map baseline on expression (right) matches the best model, so SCFM geometry does not beat classical trajectory recovery.

5 Discussion

Read together, the six analyses give a coherent and deliberately unsensational picture. The geometry of knowledge inside biological LLMs is *real*: cell identity is low-dimensional and linearly decodable, unrelated concepts are near-orthogonal, taxonomy is partly reflected, and developmental trajectories trace recoverable manifolds. But it is *modest* on every axis that the language-model literature has made us expect to be crisp. Analogies are only weakly parallel; steering is unreliable; the interesting attributes—subtype, disease, developmental time—are curved rather than linear; hierarchical structure is coarse; and cross-model convergence points toward a shared, partly-PCA-like geometry rather than the biological ontology. Most tellingly, when each property is measured against a matched expression baseline, the model-added component is small and unevenly distributed: SCFMs help most where PCA is weakest (intermediate categories) and help little where PCA is already strong or where the signal is nonlinear.

We think the most useful way to hold these results is as a convergence of the efficiency and interpretability critiques of SCFMs. Benchmarks find slim margins over classical tools (Boiarsky et al., 2023; Kedzierska et al., 2025; Liu et al., 2023); mechanistic studies find organized but correlational content in a low-dimensional, superposed geometry (Kendiukhov, 2026b,d,e). A representational-geometry analysis explains *why* the margins are slim: the linearly accessible knowledge is largely the co-expression and identity structure that PCA already captures, and the additional content the models hold is encoded nonlinearly, precisely where a linear downstream head cannot easily exploit it. The healthy near-orthogonality and the partial gains on subtype nonetheless show that the models are not merely re-expressing PCA—there is genuine, if limited, added geometric organization, and it grows with scale and pretraining breadth, most cleanly in scGPT.

The results also carry a methodological lesson. Importing the LLM interpretability toolkit to single-cell models is valuable, but its conclusions invert without a baseline: "cell type is a linear direction" and "the model traces the differentiation manifold" are both true and both largely true of PCA. In a domain where a linear embedding is already a strong model of the data, the right question is not whether structure exists but how much of it the foundation model adds—a question the text-probing literature rarely has to ask (Belinkov,

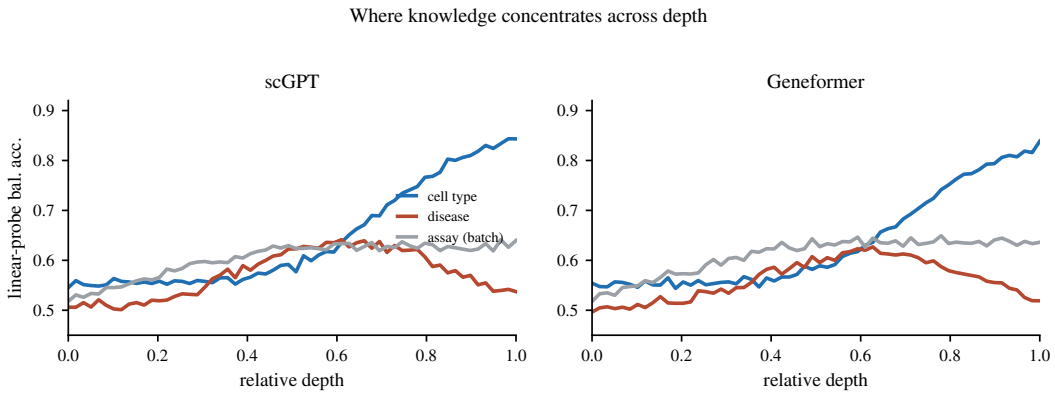


Figure 7: Depth profile of decodability. Linear-probe balanced accuracy across relative depth for scGPT and Geneformer. Cell-type identity accrues toward the output, disease peaks mid-network, and assay/batch is highest early and decays—knowledge is neither uniformly distributed nor concentrated at the last layer.

2022) but which should be standard in computational biology.

6 Limitations

Several caveats bound our claims. First, we analyze released checkpoints zero-shot; fine-tuning reshapes geometry, and a model that encodes disease nonlinearly may linearize it after task-specific adaptation, so our "present but not linear" findings describe pretrained representations, not the fine-tuned models used in practice. Second, our difference-in-means concept directions, causal-whitening choice, and geodesic estimator are one reasonable instantiation among several; alternative concept-erasure or steering methods (Zou et al., 2023) and alternative similarity measures (SVCCA (Raghu et al., 2017), other CKA variants) could shift absolute numbers, though we expect the qualitative baseline-relative ordering to be robust. Third, probing accuracy conflates information and accessibility (Belinkov, 2022); we mitigate this with control tasks and matched MLP probes but cannot fully separate the two, and our nonlinear-gap results are lower bounds on encoded knowledge. Fourth, the corpus, while metadata-rich, is human and transcriptome-only; multi-omic, spatial (Tejada-Lapuerta et al., 2025), and cross-species settings, and perturbation manifolds, may exhibit different geometry. Finally, we characterize geometry, not causation: that a direction decodes or steers a concept does not establish that the model *uses* it, a question better addressed by the circuit-level analyses of Kendiukhov (2026a), which our representational findings complement rather than replace.

7 Conclusion

We asked whether biological large language models inherit the clean geometry of knowledge that makes language-model representations interpretable, and answered with a systematic, baseline-anchored measurement across four models and six analysis families. The geometry is genuine but modest: representations are low-dimensional and anisotropic; coarse categories are linear but scarcely beyond PCA; subtype, disease, and developmental time are present yet curved; concept directions are weakly parallel though cleanly orthogonal; and models converge more to each other and to expression than to the biological ontology. The practical upshot is that the most decision-relevant knowledge in current SCFMs is either already available from classical embeddings or encoded nonlinearly, which both explains their slim benchmark margins and marks the nonlinear, higher-order structure as the place where future biological foundation models—and the tools that read them—have the most to gain. We release the evaluation suite to make baseline-relative geometric analysis a routine check on the next generation of these models.

Reproducibility and availability

All models and datasets are public: the SCFM checkpoints (Yang et al., 2022; Theodoris et al., 2023; Cui et al., 2024; Rosen et al., 2023), the CZ CELLxGENE census (CZI Cell Science Program et al., 2025), and the hematopoiesis benchmark. The estimators (TwoNN (Facco et al., 2017), Levina–Bickel (Levina and Bickel, 2004), linear CKA (Kornblith et al., 2019), causal whitening (Park et al., 2024), geodesic pseudotime (Haghverdi et al., 2016)) and the probing and baseline protocols are fully specified in Section 3. The analysis code, metadata splits, and cached embeddings will be released upon publication.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. doi: 10.48550/arXiv.1610.01644. Presented at the ICLR 2017 Workshop Track.

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 6109–6119, 2019.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1): 207–219, 2022. doi: 10.1162/coli_a_00422.

Rebecca Boiarsky, Nalini M. Singh, Alejandro Buendia, Gad Getz, and David Sontag. A deep dive into single-cell RNA sequencing foundation models. *bioRxiv*, 2023. doi: 10.1101/2023.10.19.563100. 2023.10.19.563100.

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*, 21:1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. doi: 10.48550/arXiv.2309.08600. arXiv:2309.08600.

CZI Cell Science Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M. Bell, Emanuele Bezzi, Batuhan Cakir, Jerome Chaffer, Signe Chambers, et al. CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research*, 53(D1):D886–D900, 2025. doi: 10.1093/nar/gkae1142.

Alexander D. Diehl, Terrence F. Meehan, Yvonne M. Bradford, Matthew H. Brush, Wasila M. Dahdul, David S. Dougall, Yongqun He, David Osumi-Sutherland, Alan Ruttenberg, Sirarat Sarntivijai, Ceri E. Van Slyke, Nicole A. Vasilevsky, Melissa A. Haendel, Judith A. Blake, and Christopher J. Mungall. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *Journal of Biomedical Semantics*, 7(1):44, 2016. doi: 10.1186/s13326-016-0088-7.

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. *arXiv preprint arXiv:2209.10652*, 2022. doi: 10.48550/arXiv.2209.10652. Transformer Circuits Thread.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1006.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):12140, 2017. doi: 10.1038/s41598-017-11873-y.

Oscar Franzén, Li-Ming Gan, and Johan L. M. Björkegren. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database*, 2019:baz046, 2019. doi: 10.1093/database/baz046.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1907.12009.

Laleh Haghverdi, Maren Büttner, F. Alexander Wolf, Florian Buettner, and Fabian J. Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13(10):845–848, 2016. doi: 10.1038/nmeth.3971.

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*, 21:1481–1491, 2024. doi: 10.1038/s41592-024-02305-7.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The Platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 20617–20642. PMLR, 2024.

Kasia Z. Kedzierska, Lorin Crawford, Ava P. Amini, and Alex X. Lu. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26:101, 2025. doi: 10.1186/s13059-025-03574-x.

Ihor Kendiukhov. Causal circuit tracing reveals distinct computational architectures in single-cell foundation models: inhibitory dominance, biological coherence, and cross-model convergence. *Bioinformatics*, 2026a. doi: 10.1093/bioinformatics/btag379.

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

Ihor Kendiukhov. Sparse autoencoders reveal interpretable cell-type programs in single-cell foundation model representations. *Journal of Biomedical Informatics*, 180:105056, 2026c. doi: 10.1016/j.jbi.2026.105056.

Ihor Kendiukhov. Sparse autoencoders reveal organized biological knowledge but minimal regulatory logic in single-cell foundation models: a comparative atlas of Geneformer and scGPT. *arXiv preprint arXiv:2603.02952*, 2026d. URL https://arxiv.org/abs/2603.02952.

Ihor Kendiukhov. Multi-dimensional spectral geometry of biological knowledge in single-cell transformer representations. *arXiv preprint arXiv:2602.22247*, 2026e. URL https://arxiv.org/abs/2602.22247.

Ihor Kendiukhov. Discovery of a hematopoietic manifold in scGPT yields a method for extracting performant algorithms from biological foundation model internals. *arXiv preprint arXiv:2603.10261*, 2026f. URL https://arxiv.org/abs/2603.10261.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR, 2019.

Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17 (NIPS 2004)*, pages 777–784. MIT Press, 2004.

Tianyu Liu, Kexing Li, Yuge Wang, Hongyu Li, and Hongyu Zhao. Evaluating the utilities of foundation models in single-cell data analysis. *bioRxiv*, 2023. doi: 10.1101/2023.09.08.555192. 2023.09.08.555192.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018. doi: 10.1038/s41592-018-0229-2.

Malte D. Luecken and Fabian J. Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, 2019. doi: 10.15252/msb.20188746.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 746–751, Atlanta, Georgia, 2013. Association for Computational Linguistics.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 16–30, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.blackboxnlp-1.2.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 39643–39666. PMLR, 2024.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. doi: 10.48550/arXiv.2406.01506. arXiv:2406.01506.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 6076–6085, 2017.

Aviv Regev, Sarah A. Teichmann, Eric S. Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, et al. The human cell atlas. *eLife*, 6:e27041, 2017. doi: 10.7554/eLife.27041.

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorcan, Tabula Sapiens Consortium, Stephen R. Quake, and Jure Leskovec. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*, 2023. doi: 10.1101/2023.11.28.568918. 2023.11.28.568918.

Manu Setty, Vaidotas Kiseliovas, Jacob Levine, Adam Gayoso, Linas Mazutis, and Dana Pe'er. Characterization of cell fate probabilities in single-cell data with Palantir. *Nature Biotechnology*, 37(4):451–460, 2019. doi: 10.1038/s41587-019-0068-4.

Artur Szałata, Karin Hrovatin, Sören Becker, Alejandro Tejada-Lapuerta, Haotian Cui, Bo Wang, and Fabian J. Theis. Transformers in single-cell omics: a review and new perspectives. *Nature Methods*, 21(8):1430–1443, 2024. doi: 10.1038/s41592-024-02353-z.

Alejandro Tejada-Lapuerta, Anna C. Schaar, Robert Gutgesell, Giovanni Palla, Lennard Halle, Maria Minaeva, Lukas Vornholz, Leander Dony, Charlotte Drummer, Jan Hasenauer, and Fabian J. Theis. Nicheformer: a foundation model for single-cell and spatial omics. *Nature Methods*, 22(12):2525–2538, 2025. doi: 10.1038/s41592-025-02814-z.

The Tabula Sapiens Consortium. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896, 2022. doi: 10.1126/science.abl4896.

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, 2023. doi: 10.1038/s41586-023-06139-9.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023. arXiv:2302.00294.

Hongzhi Wen, Wenzhuo Tang, Xinnan Dai, Jiayuan Ding, Wei Jin, Yuying Xie, and Jiliang Tang. CellPLM: Pre-training of cell language model beyond single cells. In *International Conference on Learning Representations (ICLR)*, 2024. doi: 10.1101/2023.10.03.560734. bioRxiv 2023.10.03.560734.

F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19:15, 2018. doi: 10.1186/s13059-017-1382-0.

Fan Yang, Wenchuan Wang, Fang Wang, Yuejing Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence*, 4:852–866, 2022. doi: 10.1038/s42256-022-00534-z.

Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, 2017. doi: 10.1038/ncomms14049.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023. doi: 10.48550/arXiv.2310.01405.