Do Geometric Outliers Identify Important Genes in Single-Cell Foundation Models?

Justin P. Whalley

Center for Cancer Cell Biology, Immunology and Infection,
The Chicago Medical School, Rosalind Franklin University of Medicine and Science,
North Chicago, IL 60064, USA
E-mail: justin.whalley@rosalindfranklin.edu

Single-cell foundation models (scFMs) encode genes in static input embeddings, but it is unclear whether the same genes stand out geometrically across models, or what those extremes imply for cell-type annotation and disease association. The same four-metric screen was applied to Geneformer, scGPT and scFoundation, which differ in architecture, training objective and expression encoding. The models agreed weakly on individual outlier genes: Geneformer and scGPT overlapped above chance, scGPT and scFoundation did not, and outlier-set stability differed between models. Class-level convergence was nevertheless structured: ribosomal enrichment occurred in all three, although weaker and sensitive to the outlier definition in scFoundation, while strong mitochondrial enrichment was confined to Geneformer and scGPT. Most outliers were not extreme in ESM-2 protein-sequence space, indicating limited concordance between scFM and protein-sequence geometry. These patterns did not translate into the tested notions of gene importance: matched input-deletion tests of the highest-anomaly Geneformer genes did not detect an excess loss of cell-type annotation performance, and outlier status showed no detectable association with ClinVar membership in any model after covariate adjustment. Gene-embedding geometry therefore characterises model-specific structure, but does not provide standalone evidence of downstream leverage or disease relevance.

Keywords: Single-cell foundation models; embedding geometry; model auditing; matched controls; gene prioritisation; model comparison.

1. Introduction

Single-cell foundation models (scFMs) are increasingly being used to represent transcriptomic data and to support biological inference.[1–4] These models are commonly assessed through downstream tasks, including cell-type annotation, perturbation prediction and data integration. However, the models also provide gene-level representations that are used to study gene relationships and to nominate genes for experimental follow-up.[1,5,6] This use assumes that properties of a learned representation carry biological information, rather than reflecting how a particular model was constructed and trained.

The gene representations used by scFMs are produced in different ways. Geneformer uses a BERT-style encoder over rank-ordered gene-token sequences, whereas scGPT combines gene-

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

token and expression-value embeddings within a generative transformer.[1,2,7] scFoundation uses the xTrimoGene asymmetric encoder–decoder, in which unmasked non-zero positions pass through a larger encoder and continuous expression values are projected through learned auto-discretisation rather than fixed expression bins.[4] As such, the models differ in their architecture, expression encoding, vocabulary, pretraining objective and training corpus.

Most comparative evaluations of scFMs concentrate on downstream performance. Several recent studies have found that foundation models do not consistently outperform simpler approaches in zero-shot evaluation, cell-type annotation or perturbation prediction.[8–11] Recent work has examined internal scFM representations: sparse autoencoders trained on hidden activations have recovered biological and technical features in Geneformer, scGPT and scFoundation,[5,6] and a large automated hypothesis screen of Geneformer and scGPT found that the two models agree on the broad shape of gene space while placing individual genes at different coordinates within it.[12] Those studies read contextual representations produced by forward passes. The complementary question asked here is whether a common screen of the static input embeddings nominates the same genes across models, and whether those nominations carry information beyond matched and covariate-adjusted controls.

Embedding geometry provides a direct way to make this comparison. Work in natural language processing has described anisotropy (embeddings crowding into a narrow cone around a shared direction), variation in embedding norm and neighbourhood hubness (a few vectors recurring as nearest neighbours of many others), and has linked some geometrically unusual tokens to unreliable model behaviour.[13–18] Related geometric analyses have now been proposed for transcriptomic models.[19] However, geometric extremeness does not by itself demonstrate biological importance or functional consequence. Genes selected using an expression-correlated score may also differ in expression breadth, gene length and functional class, each of which can produce apparently meaningful associations.

The screen used here is deliberately weight-only: it operates on a model's static gene-embedding matrix without a forward pass or a shared expression dataset. This permits the same geometric calculation to be applied despite incompatible model input encodings and, once protein representations have been computed, to ESM-2 as a non-transcriptomic control. Outcome independence makes the screen useful for model characterisation, not evidence of biological or functional relevance; those claims require the matched and covariate-adjusted tests that follow. The corresponding limitation is that the screen describes input embeddings, not the contextual representations produced for particular cells.

Here, the same four-metric geometric screen is applied to Geneformer, scGPT and scFoundation. The resulting outlier sets are tested for robustness to alternative outlier definitions, for recurrence in ESM-2 protein-sequence space,[20] and for association with ClinVar after adjustment for expression, breadth, length, constraint and gene class. The 50 highest-anomaly Geneformer genes, whose ranking is stable under those alternatives, are then carried into matched deletion experiments in PBMC3k and Tabula Sapiens.

The three models produce substantially different outlier profiles, and most outliers do not recur in ESM-2 sequence space. In Geneformer, matched input-deletion tests of the highest-anomaly genes did not detect an excess loss of cell-type annotation performance, and no

model's outlier set shows a detectable ClinVar association after covariate adjustment. Embedding geometry is therefore useful for comparing how scFMs represent their vocabularies, but is not stand-alone evidence of functional or disease relevance.

2. Methods

2.1. *Models, embeddings and the geometry screen*

Geneformer V2-104M,[1,7] scGPT[2] and scFoundation[4] are audited. The screen is weight-only: it reads a static input gene-embedding matrix and its gene identifiers, and requires no forward pass. Each gene is scored on four metrics: embedding norm, Euclidean distance to the vocabulary centroid, cosine similarity to the centroid, and mean cosine distance to its $k = 10$ nearest neighbours (isolation). Each is converted to a z-score across the vocabulary; a gene is called an outlier if $|z| > 3$ on any metric, equivalently if its anomaly score $A_g = \max_j |z_{gj}|$, which also ranks genes, exceeds 3. The metrics are the transcriptomic analogues of quantities studied in NLP embedding spaces: norm and its relation to token statistics,[15] anisotropy and centroid structure,[13,14] and neighbourhood irregularity.[16] Informally, norm asks whether a gene vector is unusually long or short; centroid distance asks how far it lies from the vocabulary centre; cosine asks how closely it aligns with the direction shared by most genes; and isolation asks whether it lacks close neighbours. Both tails are flagged, so an unusually short vector or a crowded neighbourhood is also called. Calls are combined by union because a gene can be unusual on one axis alone; the screen is deliberately inclusive, since the matched and covariate-adjusted tests that follow decide whether a nomination carries information. In Geneformer, special tokens (`<pad>`, `<mask>`, `<cls>`, `<eos>`) are scored but excluded from the reported sets; removing them before scoring returns the same outlier set. The screen was also calibrated on simulated 20,000-gene matrices with 256, 512 or 768 dimensions, under an isotropic Gaussian null and an anisotropic null calibrated to the median of the three models' mean cosine similarities. In each planted matrix, 100 genes were given an inflated norm (1.05–1.2 times) or a direction rotated 5–20 degrees away from the shared axis, and all four statistics were recomputed; each condition was run over 20 seeds (code release).

2.2. *Robustness to alternative outlier definitions*

Outlier identity is re-derived using three robust alternatives to the original $|z| > 3$ call, each replacing the mean and standard deviation with the median and the median absolute deviation (MAD): MAD-based $|z| > 3$, MAD-based $|z| > 3.5$, and a rank-matched composite that takes the same number of genes as the original call, ranked by the largest MAD-based $|z|$ across metrics. The MAD is scaled by 1.4826, so its thresholds are on the same scale as the original call. Three further rules, a two-component Gaussian mixture, percentile 1/99 and Tukey's IQR rule ($k = 3$), are degenerate on these matrices (assigning up to half of all genes to the minority component, flagging 5–7% after union, or returning outliers for Geneformer alone) and are reported as diagnostics but excluded from stability assessment, with the supporting tables in the code release.

2.3. *Non-transcriptomic embedding control*

The identical screen is applied to ESM-2 (`esm2_t33_650M_UR50D`) protein embeddings to test whether outliers recur in sequence space.[20] Genes are mapped to MANE Select v1.3 proteins,[21] with reviewed UniProt entries as fallback.[22] Representations are final-layer residue means excluding BOS and EOS; proteins exceeding the 1,022-residue context are embedded in 50%-overlapping windows, averaged first within each window and then, unweighted, across windows. Comparisons use each model's protein-mappable vocabulary (19,017 genes for Geneformer, 18,748 for scGPT and 18,788 for scFoundation). This control tests recurrence attributable to primary sequence, not functional centrality or biology absent from protein sequence.

2.4. *Matched-control deletion test*

Treatment and control sets. The treatment is Geneformer's 50 highest-anomaly genes, whose ranking was stable across outlier definitions (Section 3.2). Two hundred control sets are matched on $\log_{1p}$ mean expression, expression breadth and $\log_{1p}$ gene length after candidate-pool z-scaling,[23] and exactly on mutually exclusive gene class (Section 2.5). Each control gene is drawn uniformly from the five nearest eligible genes of the same class in that z-scaled space. Draws are independent and deduplicated within, but not across, sets. Within each dataset, mean expression was the mean raw count across cells and breadth was the fraction of cells with non-zero expression. The contrast therefore estimates excess deletion cost relative to comparable genes, not arbitrary or low-expression genes.

Sensitivity treatment. The procedure is repeated on the 36 genes remaining after removing ribosomal and mitochondrial genes, using a separate 200-draw null matched to that subset.

Deletion mechanics. Genes are removed from each cell's rank-value encoding and sequences are re-padded. This removes a measurement from the model input, approximating an incomplete assay rather than a biological knockout. Attention masks are derived from token content so ablated sequences cannot attend to padding.

Outcome. Geneformer CLS embeddings[7] are evaluated by macro-F_1, the F_1 score averaged over cell types so that rare types count equally, from a multinomial logistic probe with balanced class weights, retrained under five-fold stratified cross-validation.[24] The treatment is displayed against the central 95% band of the matched-control distribution; inference is one-sided on the lower tail, with empirical $p = (b+1)/(m+1)$ where b of m controls are at least as damaging.[25] Fixed-probe, ARI and NMI results are diagnostic only; the former is in-sample, and the clustering nulls are too wide to resolve an effect of this size.

Datasets. The primary test uses PBMC3k (2,638 cells; 8 cell types).[26,27] Replication uses a stratified Tabula Sapiens Immune subset:[28,29] 19,984 cells were subsampled to 4,000 and types with fewer than 10 cells removed, leaving 3,919 cells across 22 of 43 types and 100 matched draws. Each dataset was run under its own pinned environment; both environment fingerprints are recorded.

2.5. *Annotation panels and class schemes*

Ribosomal genes are defined by a pinned 171-gene panel from HGNC groups 728, 729 and 646, matched by Ensembl ID; this includes `RPLP0`–`RPLP2` and `RPSA` while excluding similarly named kinases. Other panels are mitochondrial (`MT-`), loss-of-function constrained, and disease-associated (ClinVar[30]). Constraint uses gnomAD v4.1 values with pLI > 0.9 or LOEUF < 0.35.[31] Applying the LOEUF < 0.45 threshold now recommended for v4.1.1[32] to those values gives the same set, because all genes between the thresholds already meet the pLI criterion. Panels overlap: 1,784 of 8,285 ClinVar genes (22%) are also constrained.

Every association is therefore evaluated under two schemes. The mutually exclusive scheme assigns one class per gene in the precedence order mitochondrial → ribosomal → constrained → disease → other. The overlapping scheme tests direct membership, allowing a gene to belong to several panels. Reporting both is not redundancy: the two can disagree in direction (Section 3.5).

2.6. *Statistics, environment and reproducibility*

Unadjusted odds ratios use two-sided Fisher tests. Cross-model overlap p values are one-sided hypergeometric, Bonferroni-adjusted over the three pairwise comparisons. Enrichments in the robustness analysis are Bonferroni-adjusted over four panels; other class-association p values are raw and descriptive. ClinVar membership is regressed on outlier status, constraint, $\log_{1p}$ expression, breadth, $\log_{1p}$ gene length, and ribosomal and mitochondrial indicators. Regression covariates used maximum Human Protein Atlas consensus nTPM across 51 tissues,[33] the fraction of tissues with nTPM ≥ 1, and the longest Ensembl BioMart gene span. ClinVar membership required at least one pathogenic or likely pathogenic allele in the gene-specific summary dated 4 April 2026. Perfect separation by the mitochondrial indicator requires primary Firth fits with profile-likelihood intervals.[34] The same universe, covariates and estimator are used for all three models. Matching balance uses standardised mean differences.[35] Null bands are control 2.5th–97.5th percentiles; empirical p values are one-sided. Treatment and controls share a pinned environment; ΔF_1 uses a same-device baseline, whereas $z = (F_1^{\text{treat}} - \overline{F_1}^{\text{ctrl}})/s^{\text{ctrl}}$ is baseline-invariant.

3. Results

3.1. *Geometric outlier profiles differ substantially among models*

Applying the same geometric screen to all three models produced substantially different outlier profiles (Figures 1 and 2). Geneformer yielded 410 outliers across a 20,271-gene vocabulary, scGPT yielded 188 across 60,694 genes, and scFoundation yielded 164 across 19,264 genes. Ribosomal enrichment occurred in all three models, although it was weaker and sensitive to the outlier definition in scFoundation; strong mitochondrial enrichment was confined to Geneformer and scGPT.

Counts alone do not establish disagreement, so the outlier sets were compared directly on the 18,915 genes shared by all three vocabularies, where the sets contain 388, 102 and 161 outliers (Table 1). Geneformer and scGPT share 72 outliers against 2.1 expected by chance,

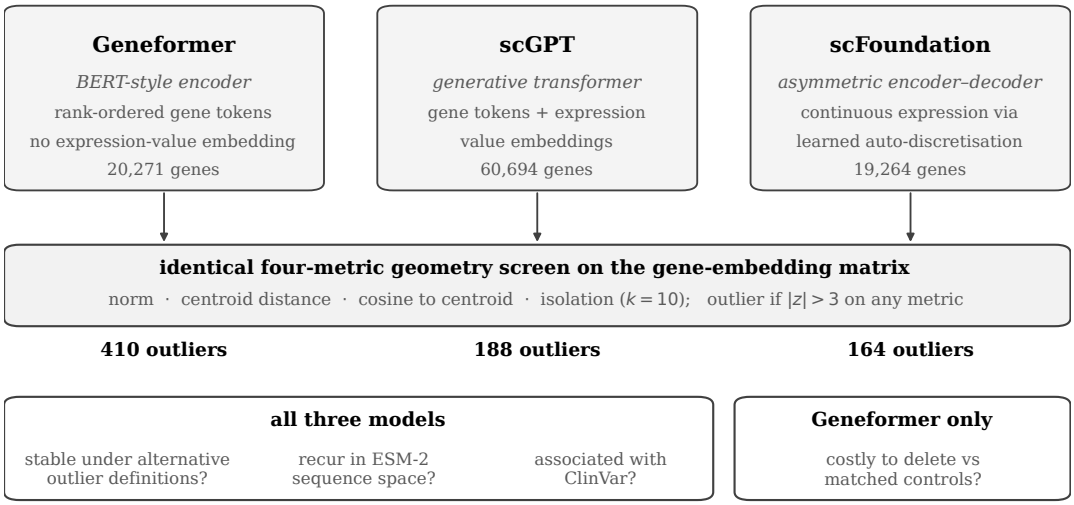


Fig. 1. The three models and the shared screen. The models differ in architecture, expression encoding and vocabulary size, and also in pretraining objective, corpus and scale, which are not shown. Robustness to alternative outlier definitions, ESM-2 recurrence and covariate-adjusted ClinVar association are evaluated for all three; matched deletion is the Geneformer follow-up.

a 34-fold excess (adjusted $p = 3 \times 10^{-99}$), yet their Jaccard index, the shared outliers as a fraction of all outliers called in either model, is only 0.172 and the genome-wide correlation between their anomaly scores is $\rho = 0.186$. Geneformer and scFoundation share 8 against 3.3 expected (raw $p = 0.018$; Bonferroni-adjusted $p = 0.055$), and scGPT and scFoundation share a single gene against 0.9 expected (adjusted $p = 1$), which is not detectably above chance. One gene, *HSP90AA1*, is an outlier in all three. The two models that represent genes as discrete tokens agree more with each other than either does with the continuous-expression model, while still agreeing only weakly in absolute terms. This comparison cannot isolate expression encoding from differences in corpus, objective, vocabulary or scale.

Table 1. Direct cross-model agreement on the 18,915 genes shared by all three vocabularies. Overlap p: one-sided hypergeometric, Bonferroni-adjusted over the three pairs.

Model pair	Overlap	Expected	Jaccard	ρ	Adjusted overlap p
Geneformer, scGPT	72	2.1	0.172	0.186	3×10^{-99}
Geneformer, scFoundation	8	3.3	0.015	0.026	0.055
scGPT, scFoundation	1	0.9	0.004	0.007	1

Over the same 18,915 shared genes, class composition also differed. Mitochondrial genes were enriched among Geneformer and scGPT outliers (OR 163 and 455), whereas scFoundation had no mitochondrial outliers at all; ribosomal ORs were 44.8, 18.9 and 4.4 across the same three models. Expression separated the models in the same way: median maximum HPA expression among outliers was 19 times the non-outlier median in Geneformer and 12 times

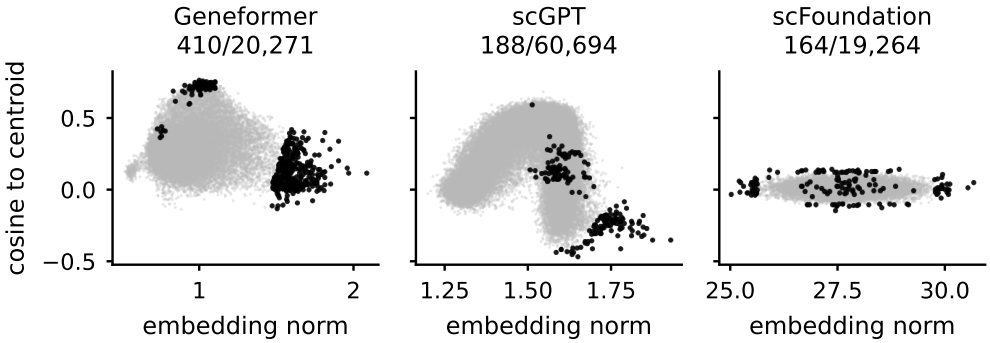


Fig. 2. Gene-embedding geometry in each model, with called outliers in black; the cosine axis is shared, the norm axis is model-specific. Geneformer and scGPT each separate into two clouds along the norm axis, the structure behind scGPT's sensitivity to the outlier definition (Section 3.2); scFoundation instead forms a narrow band, with cosine to the centroid 0.009 ± 0.035 and a 2.5% coefficient of variation in embedding norm.

in scGPT, but slightly below it in scFoundation. Geneformer outliers were also enriched for loss-of-function-constrained genes (OR 1.46 under the mutually exclusive scheme), and its 50 highest-anomaly genes had much higher median maximum HPA expression than the vocabulary overall (1,813 against 42 nTPM). Downstream effects could therefore reflect expression and class rather than geometry.

3.2. *Outlier-set stability differs among models*

The preceding section compared models under one outlier definition. Because the metric distributions are skewed or bimodal, outlier sets were also re-derived within each model with the mean and standard deviation replaced by the median and median absolute deviation (MAD) (Section 2.2). Under the MAD-based $|z| > 3$ definition, all 410 original Geneformer outliers were recovered, so containment, the fraction of the original set recovered, is 100%; the anomaly-score rankings are highly concordant ($\rho = 0.956$) and the top-50 genes are identical. scFoundation behaves similarly (containment 97%, $\rho = 0.999$, top-50 overlap 49/50).

Geneformer's Jaccard index against that MAD set is nonetheless only 0.492, because the MAD definition recovers all 410 original outliers and adds a further 423; the original set is a strict subset, so the low value reflects the larger selected set rather than loss of the original outliers, which form a conservative core.

Containment must be read against its ceiling, since a stricter rule selecting fewer genes cannot contain them all; Figure 3a marks these ceilings. This matters most for scFoundation under the stricter MAD-based $|z| > 3.5$ definition, where containment falls to 16%: that rule selects only 27 genes, all among the original 164, so the bar sits on its ceiling and the low value reflects the number of genes selected rather than instability. scGPT is the exception: under MAD-based $|z| > 3$ it retains 83% of its original outliers, but only 22 of its original top-50 genes remain in the top 50, so individual gene identity is sensitive to the definition, a consequence of its bimodal norm distribution, under which mean- and median-based standardisation weight the tails differently. Class-level associations are nevertheless stable across the MAD family for

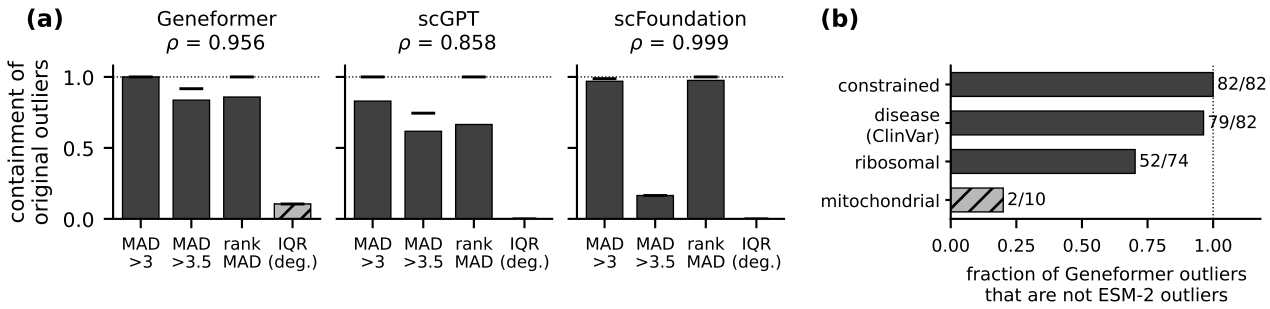


Fig. 3. (a) Outlier identity under alternative outlier definitions. Bars give containment, the fraction of original $|z| > 3$ outliers recovered; ticks mark the maximum attainable containment $\min(n_{\text{new}}, n_{\text{orig}})/n_{\text{orig}}$, so a bar meeting its tick is saturated, not unstable; the hatched IQR bar is one of three degenerate rules reported but excluded (Section 2.2); ρ is the Spearman correlation between original and MAD-based anomaly scores. (b) Fraction of Geneformer outliers that do not recur in ESM-2, by mutually exclusive class over 19,017 shared genes, labelled with the scFM-specific count; the hatched mitochondrial bar is short because only 2 of its 10 outliers are scFM-specific, the other 8 being ESM-2 outliers as well.

Geneformer and scGPT alike; no class significant under the original call loses significance.

In the simulated matrices of Section 2.1, which test the screen rather than the models, the screen flagged 0.8–0.9% of genes on average by union (0.25–0.35% per metric) when nothing was planted; recovery of planted genes crossed 50% near a mean $|z|$ of 3 on the targeted metric across dimensions, and unplanted genes were flagged at 0.4–0.8% on average.

3.3. *Outliers are largely model-specific, with a smaller sequence-related component*

If geometric outliers reflected properties of the encoded protein sequences that another representation would also expose, the same screen applied to protein-sequence embeddings should recover them (Section 2.3). Across 19,017 shared genes, Geneformer and ESM-2 outlier sets had Jaccard 0.054 and anomaly-score correlation $\rho = 0.072$ (averaged over all three scFMs, 0.023 and 0.042). Forty-nine genes overlapped against 11.7 expected ($p < 10^{-15}$), but only four appeared in both top-50 lists. Thus most scFM outliers did not recur in ESM-2, with a smaller sequence-related component.

The shared component included eight mitochondrially encoded, 22 ribosomal and seven olfactory-receptor genes. Under mutually exclusive classes (Figure 3b), all 82 constrained and 79 of 82 disease-associated outliers were scFM-specific (outliers in Geneformer but not in ESM-2), as were 52 of 74 ribosomal outliers. Mitochondrial genes were the exception, with only 2 of 10 scFM-specific because the other 8 recurred in ESM-2. These encode hydrophobic membrane proteins of distinctive composition,[36] so their recurrence is compatible with sequence distinctiveness, although the properties responsible were not tested.

3.4. *Matched deletion in Geneformer detected no excess annotation loss*

If geometric extremeness marked genes on which Geneformer depends, deleting them from its input should cost more cell-type annotation performance than deleting comparable genes.

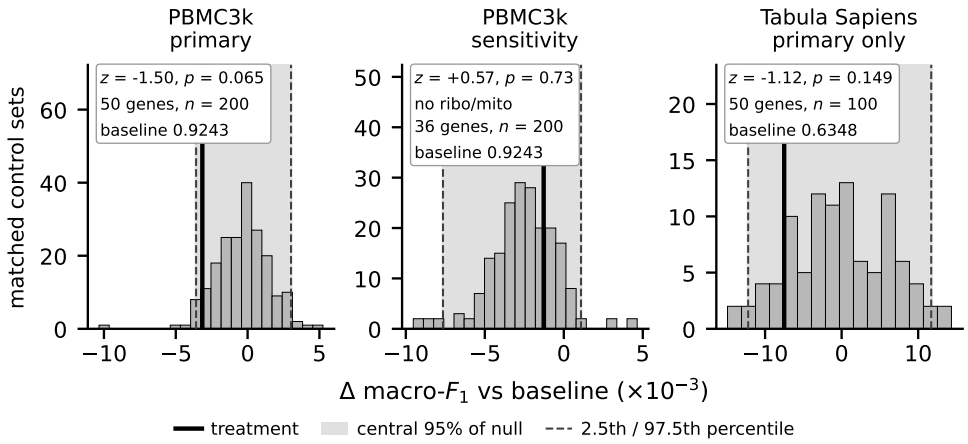


Fig. 4. Matched-control deletion results for Geneformer. Each panel gives Δ macro-F_1 for gene sets matched to the treatment on expression level, breadth, gene length and class; shading is the 2.5th–97.5th percentile null band, the vertical line is the treatment, and panel labels give z and the one-sided empirical p. Axes are not shared, so the three histograms are not a common null. Left, the 50 highest-anomaly Geneformer genes in PBMC3k. Centre, the 36 remaining after removing ribosomal and mitochondrial genes, against their own null. Right, the primary arm on Tabula Sapiens; its sensitivity arm was not replicated.

From a PBMC3k baseline of 0.924, deleting the 50 highest-anomaly Geneformer genes left macro-F_1 0.00280 below the mean of 200 matched controls ($z = -1.50$). Twelve controls were at least as damaging ($p = 0.065$), placing the treatment inside the empirical 95% null band (Figure 4).

Ribosomal and mitochondrial genes account for 28% of the treatment set, so any cost might follow from removing those classes rather than from geometry. The 36 genes remaining once they were excluded sat 0.00122 above their own control mean ($z = +0.57$; 145 of 200 controls at least as damaging, $p = 0.73$). The weak degradation seen in the full treatment was not observed in the non-ribosomal, non-mitochondrial sensitivity arm; because each arm was assessed against its own matched null, the analysis does not test the difference between arms.

Replication in a broader annotation task. The same treatment was applied to a Tabula Sapiens immune subset, extending the test from eight PBMC3k cell types to 22 and from a baseline macro-F_1 of 0.924 to 0.635. It left macro-F_1 0.0073 below the mean of 100 controls, again inside the null ($z = -1.12$; 14 controls at least as damaging, $p = 0.149$). Token exposure was comparable between treatment and controls, so the result is not explained by the treatment genes appearing less often in the analysed cells; matching was exact on class though imperfect on expression (SMD 0.283) and length (0.112). Agreement across two datasets with very different baselines supports the primary result, though it does not show the effect to be exactly zero.

Table 2. ClinVar association over 18,911 complete cases. GF is Geneformer; n is the outlier count within this universe. **(a)** covariate-adjusted (Firth logistic regression, profile-likelihood intervals); GF ∩ scGPT is exploratory, selected post hoc. **(b)** unadjusted, under two annotation schemes that estimate different quantities.

(a) Adjusted

Gene set	Unadj. OR	Adjusted OR (95% CI)	p
Geneformer ($n = 388$)	1.16	1.05 (0.82–1.34)	0.72
scGPT ($n = 102$)	2.95	1.21 (0.76–1.98)	0.42
scFoundation ($n = 161$)	0.80	0.98 (0.69–1.38)	0.91
GF ∩ scGPT (exploratory; $n = 72$)	3.59	1.44 (0.79–2.69)	0.23

(b) Unadjusted, by annotation scheme

Gene set	Class scheme	OR	p
GF ∩ scGPT ($n = 72$)	Overlapping	3.59	5.0×10^{-7}
GF ∩ scGPT ($n = 72$)	Mutually exclusive	0.385	1.6×10^{-3}
All Geneformer ($n = 388$)	Overlapping	1.16	0.15
All Geneformer ($n = 388$)	Mutually exclusive	0.519	3.8×10^{-8}

3.5. *No model's outliers show a detectable adjusted ClinVar association*

Geometric outliers are enriched for functionally central gene classes, inviting the inference that they are disease-relevant. ClinVar[30] membership was therefore regressed on outlier status, adjusting for expression, breadth, length, constraint and gene class (Section 2.6). Across 18,911 complete cases, none of the three outlier sets showed a detectable association (Table 2a). Geneformer gave an adjusted OR of 1.05 (95% confidence interval, CI, 0.82–1.34, $p = 0.72$), and scFoundation 0.98 (0.69–1.38, $p = 0.91$). For scGPT the adjustment removed a strong unadjusted signal: its OR of 2.95 ($p = 1.8 \times 10^{-7}$) fell to 1.21 (0.76–1.98, $p = 0.42$) under the same adjustment. The intervals remain compatible with effects in either direction, including a nearly twofold association for scGPT, so these are nulls of detection rather than demonstrations of exact independence.

The 72-gene Geneformer–scGPT subset was similarly enriched before adjustment (OR 3.59, $p = 5.0 \times 10^{-7}$), but not afterwards (1.44, 0.79–2.69, $p = 0.23$).

That same unadjusted enrichment is also sensitive to something other than covariates. Under direct membership the 72 shared outliers give a ClinVar odds ratio of 3.59 ($p = 5.0 \times 10^{-7}$); under the mutually exclusive scheme the same 72 genes give 0.385 ($p = 1.6 \times 10^{-3}$), an apparent depletion (Table 2b). The reversal follows from the precedence order. Constrained genes are labelled before disease genes, 22% of ClinVar genes are also constrained, and these outliers are enriched for constraint at OR 7.36, so the residual disease category is stripped of the genes that would have carried the enrichment. Both are correctly computed and answer different questions. A gene-class enrichment reported without its scheme can therefore point either way.

4. Discussion

The same screen returned sets that differed in size, class composition and stability under alternative outlier definitions, so a finding in one scFM should not be assumed to transfer to another. The discrete-token models agreed most, but architecture, encoding, vocabulary, objective, corpus and scale all varied, with one model per design. This ordering motivates controlled comparison; it does not identify its cause or establish that one design is preferable.

In Geneformer and scGPT, outliers had much higher maximum tissue expression and were enriched for functionally central classes; scFoundation showed much less of both, with outlier expression slightly below the vocabulary median. Expression breadth did not follow the same pattern, running lower among Geneformer outliers and higher among scGPT's. Pretraining token frequencies were not measured, so it cannot be shown that these are the genes each model encountered most often, and expression measured in a held-out dataset is only a proxy for exposure during training. Gene expression is heavy-tailed,[37] and the observed enrichment toward the head of that distribution is an inference rather than a demonstration. This is still useful information about a model: it describes how learned gene representations are distributed across a vocabulary, and it differs markedly between the models examined here.

Matched input-deletion tests of the highest-anomaly Geneformer genes did not detect an excess loss of cell-type annotation performance relative to matched controls in either dataset. The result should be read narrowly: deletion removes a gene's measurement from the model input and, with the probe retrained after each deletion, asks whether cell-type information remains recoverable from the remaining inputs beyond matched controls, not how much a fixed classifier relied on the deleted genes. It is not a knockout, and it says nothing about gene function, viability or perturbation response. Separately, none of the three models' outlier sets showed a detectable adjusted ClinVar association; scGPT provides the clearest example, with its odds ratio falling from 2.95 to 1.21. Genes selected on an expression-correlated criterion can therefore appear functionally central while showing no detectable additional signal for the outcomes tested here, cell-type annotation under matched deletion and covariate-adjusted ClinVar membership. That is the distinction matched or covariate-adjusted controls are there to make.

The absence of a detectable adjusted association is itself interesting. Geneformer and scGPT outliers were strongly enriched for ribosomal, mitochondrial and constrained genes, whereas scFoundation showed weaker ribosomal enrichment and no mitochondrial outliers; yet none of the three sets showed a residual link to genes already curated as disease-associated once expression, breadth, length, constraint and class were accounted for. Enrichment for familiar biological classes did not, by itself, imply additional information for the outcomes tested here. An expression-correlated score preferentially selects highly expressed genes and, through correlated covariates, genes over-represented in constraint and disease panels, so an unmatched enrichment cannot show that the score itself carries information. The annotation-scheme result makes the same point from another direction: overlapping and mutually exclusive panels estimate different quantities, so a gene-class enrichment should state its scheme and report both where they disagree.

The transferable contribution is therefore the audit design rather than a universal list

of outlier genes: a checkpoint-level score generates a hypothesis, and matched or covariate-adjusted analyses test whether it adds information for a specified outcome. The outlier sets reported here are specific to three checkpoints and should not be assumed to transfer. The screen half of that design is also cheap, which matters as models proliferate. When a released checkpoint exposes a static gene-embedding matrix and identifiers, the screen needs no evaluation data and thus avoids the contamination question that downstream benchmarking raises whenever overlap with the pretraining corpus is undocumented.

Limitations. This work concentrates on three models chosen for their different representational designs, whose geometries differ enough that conclusions about another model would need the same analysis repeated.[38] Each design is represented by a single checkpoint, so differences cannot be attributed to architecture, and the deletion test was run for Geneformer only, so its downstream annotation conclusion is a Geneformer result, whereas the ClinVar test covers all three models. The deletion experiments were restricted to two immune datasets and cell-type annotation, though nothing in the design prevents the same test on other tissues, or on perturbation prediction,[39] network inference or atlas integration.[40] Within those tests, Tabula Sapiens uses a filtered 22-class subset with 100 draws and no sensitivity arm, and matching remains imperfect (expression SMD 0.32 in PBMC3k and 0.28 in Tabula Sapiens; length 0.11 in Tabula Sapiens).[35] The screen also describes geometry without defining what a well-behaved embedding space would look like, and reads static input embeddings rather than contextual representations computed for individual cells. Finally, the pretraining corpora are too large to work with here, so why particular genes occupy extreme positions cannot be established. Work on language models has traced anomalous tokens to under-exposure during training,[17,18] and the equivalent account for scFMs would need that access.

5. Conclusion

Geneformer, scGPT and scFoundation produce substantially different gene-embedding geometries, and genes exceptional in one model are largely not exceptional in another. Most are also not outliers in ESM-2 protein-sequence space, indicating limited concordance between the geometric extremes of the scFM and protein-sequence representations. In the three checkpoints examined, and on the two definitions of importance tested, matched input-deletion tests in Geneformer did not detect an excess loss of cell-type annotation performance, and no model's outlier set showed a detectable ClinVar association after covariate adjustment. The screen itself is cheap, needing only a released checkpoint's static gene-embedding matrix and identifiers, and that design, rather than any particular list of genes, is what carries forward.

Acknowledgements. This work was supported by a Schweppe Scholar Award to J.P.W.

Data and code availability. Code, inputs, outputs, a traceability table linking each analysis to its script and output, and the environment and checksums for reproduction are at `https://github.com/jpwhalley/matched-null-audit` (release `psb-2027-camera-ready`, commit `392517b9d950`).

References

1. C. V. Theodoris, L. Xiao, A. Chopra, M. D. Chaffin, Z. R. Al Sayed, M. C. Hill, H. Mantineo, E. M. Brydon, Z. Zeng, X. S. Liu and P. T. Ellinor, Transfer learning enables predictions in network biology, *Nature* **618**, 616 (2023).
2. H. Cui, C. Wang, H. Maan, K. Pang, F. Luo, N. Duan and B. Wang, scGPT: toward building a foundation model for single-cell multi-omics using generative AI, *Nature Methods* **21**, 1470 (2024).
3. F. Yang, W. Wang, F. Wang, Y. Fang, D. Tang, J. Huang, H. Lu and J. 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 (2022).
4. M. Hao, J. Gong, X. Zeng, C. Liu, Y. Guo, X. Cheng, T. Wang, J. Ma, X. Zhang and L. Song, Large-scale foundation model on single-cell transcriptomics, *Nature Methods* **21**, 1481 (2024).
5. F. Pedrocchi, F. Barkmann, A. Joudaki and V. Boeva, Sparse autoencoders reveal interpretable features in single-cell foundation models, *bioRxiv* (2025), bioRxiv 2025.10.22.681631; latest revision 16 June 2026.
6. I. Kendiukhov, Sparse autoencoders reveal organized biological knowledge but minimal regulatory logic in single-cell foundation models: a comparative atlas of Geneformer and scGPT, *arXiv* (2026), arXiv:2603.02952 [q-bio.GN].
7. H. Chen, M. S. Venkatesh, J. Gómez Ortega, S. V. Mahesh, T. N. Nandi, R. K. Madduri, K. Pelka and C. V. Theodoris, Scaling and quantization of large-scale foundation model enables resource-efficient predictions in network biology, *Nature Computational Science* **6**, 450 (2026).
8. R. Boiarsky, N. M. Singh, A. Buendia, A. P. Amini, G. Getz and D. Sontag, Deeper evaluation of a single-cell foundation model, *Nature Machine Intelligence* **6**, 1443 (2024).
9. K. Z. Kedzierska, L. Crawford, A. P. Amini and A. X. Lu, Zero-shot evaluation reveals limitations of single-cell foundation models, *Genome Biology* **26**, p. 101 (2025).
10. C. Ahlmann-Eltze, W. Huber and S. Anders, Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines, *Nature Methods* **22**, 1657 (2025).
11. T. Liu, K. Li, Y. Wang, H. Li and H. Zhao, Evaluating the utilities of foundation models in single-cell data analysis, *Advanced Science* **13**, p. e14490 (2026).
12. I. Kendiukhov, What topological and geometric structure do biological foundation models learn? Evidence from 141 hypotheses, *PLOS ONE* **21**, p. e0344826 (2026).
13. K. 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)*, eds. K. Inui, J. Jiang, V. Ng and X. Wan (Association for Computational Linguistics, Hong Kong, China, November 2019).
14. J. Mu and P. Viswanath, All-but-the-top: Simple and effective postprocessing for word representations, in *International Conference on Learning Representations (ICLR)*, (ICLR, Vancouver, Canada, 2018).
15. M. Oyama, S. Yokoi and H. Shimodaira, Norm of word embedding encodes information gain, in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, eds. H. Bouamor, J. Pino and K. Bali (Association for Computational Linguistics, Singapore, December 2023).
16. M. Radovanović, A. Nanopoulos and M. Ivanović, Hubs in space: Popular nearest neighbors in high-dimensional data, *Journal of Machine Learning Research* **11**, 2487 (2010).
17. J. Rumbelow and M. Watkins, SolidGoldMagikarp (plus, prompt generation) (2023), Blog post, LessWrong / AI Alignment Forum, `https://www.lesswrong.com/posts/aPeJE8bSo6rAFoLqg/solidgoldmagikarp-plus-prompt-generation`, posted 5 February 2023; accessed 22 Septem-

ber 2026.
18. S. Land and M. Bartolo, Fishing for Magikarp: automatically detecting under-trained tokens in large language models, in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, eds. Y. Al-Onaizan, M. Bansal and Y.-N. Chen (Association for Computational Linguistics, Miami, Florida, USA, November 2024).
19. W. Gilpin, The cell as a token: high-dimensional geometry in language models and cell embeddings, *Bioinformatics* **41**, p. btaf595 (2025).
20. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido and A. Rives, Evolutionary-scale prediction of atomic-level protein structure with a language model, *Science* **379**, 1123 (2023).
21. J. Morales, S. Pujar, J. E. Loveland, A. Astashyn, R. Bennett, A. Berry, E. Cox, C. Davidson, O. Ermolaeva, C. M. Farrell, R. Fatima, L. Gil, T. Goldfarb, J. M. Gonzalez, D. Haddad, M. Hardy, T. Hunt, J. Jackson, V. S. Joardar, M. Kay, V. K. Kodali, K. M. McGarvey, A. McMahon, J. M. Mudge, D. N. Murphy, M. R. Murphy, B. Rajput, S. H. Rangwala, L. D. Riddick, F. Thibaud-Nissen, G. Threadgold, A. R. Vatsan, C. Wallin, D. Webb, P. Flicek, E. Birney, K. D. Pruitt, A. Frankish, F. Cunningham and T. D. Murphy, A joint NCBI and EMBL-EBI transcript set for clinical genomics and research, *Nature* **604**, 310 (2022).
22. The UniProt Consortium, UniProt: the Universal Protein Knowledgebase in 2023, *Nucleic Acids Research* **51**, D523 (2023).
23. E. A. Stuart, Matching methods for causal inference: a review and a look forward, *Statistical Science* **25**, 1 (2010).
24. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, Scikit-learn: machine learning in Python, *Journal of Machine Learning Research* **12**, 2825 (2011).
25. B. Phipson and G. K. Smyth, Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn, *Statistical Applications in Genetics and Molecular Biology* **9**, p. Article 39 (2010).
26. G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson and J. H. Bielas, Massively parallel digital transcriptional profiling of single cells, *Nature Communications* **8**, p. 14049 (2017).
27. F. A. Wolf, P. Angerer and F. J. Theis, SCANPY: large-scale single-cell gene expression data analysis, *Genome Biology* **19**, p. 15 (2018).
28. The Tabula Sapiens Consortium, The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans, *Science* **376**, p. eabl4896 (2022).
29. CZI Cell Science Program, S. Abdulla, B. Aevermann, P. Assis, S. Badajoz, S. M. Bell, E. Bezzi, B. Cakir, J. Chaffer, S. Chambers, J. M. Cherry, T. Chi, J. Chien, L. Dorman, P. Garcia-Nieto, N. Gloria, M. Hastie, D. Hegeman, J. Hilton, T. Huang, A. Infeld, A.-M. Istrate, I. Jelic, K. Katsuya, Y. J. Kim, K. Liang, M. Lin, M. Lombardo, B. Marshall, B. Martin, F. McDade, C. Megill, N. Patel, A. Predeus, B. Raymor, B. Robatmili, D. Rogers, E. Rutherford, D. Sadgat, A. Shin, C. Small, T. Smith, P. Sridharan, A. Tarashansky, N. Tavares, H. Thomas, A. Tolopko, M. Urisko, J. Yan, G. Yeretssian, J. Zamanian, A. Mani, J. Cool and A. Carr, CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data, *Nucleic Acids Research* **53**, D886 (2025).
30. M. J. Landrum, J. M. Lee, M. Benson, G. R. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart,

D. Hoffman, W. Jang, K. Karapetyan, K. Katz, C. Liu, Z. Maddipatla, A. Malheiro, K. McDaniel, M. Ovetsky, G. Riley, G. Zhou, J. B. Holmes, B. L. Kattman and D. R. Maglott, ClinVar: improving access to variant interpretations and supporting evidence, *Nucleic Acids Research* **46**, D1062 (2018).

31. K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J. X. Chong, K. E. Samocha, E. Pierce-Hoffman, Z. Zappala, A. H. O'Donnell-Luria, E. V. Minikel, B. Weisburd, M. Lek, J. S. Ware, C. Vittal, I. M. Armean, L. Bergelson, K. Cibulskis, K. M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M. E. Talkowski, C. A. Aguilar Salinas, T. Ahmad, C. M. Albert, D. Ardissino, G. Atzmon, J. Barnard, L. Beaugerie, E. J. Benjamin, M. Boehnke, L. L. Bonnycastle, E. P. Bottinger, D. W. Bowden, M. J. Bown, J. C. Chambers, J. C. Chan, D. Chasman, J. Cho, M. K. Chung, B. Cohen, A. Correa, D. Dabelea, M. J. Daly, D. Darbar, R. Duggirala, J. Dupuis, P. T. Ellinor, R. Elosua, J. Erdmann, T. Esko, M. Färkkilä, J. Florez, A. Franke, G. Getz, B. Glaser, S. J. Glatt, D. Goldstein, C. Gonzalez, L. Groop, C. Haiman, C. Hanis, M. Harms, M. Hiltunen, M. M. Holi, C. M. Hultman, M. Kallela, J. Kaprio, S. Kathiresan, B.-J. Kim, Y. J. Kim, G. Kirov, J. Kooner, S. Koskinen, H. M. Krumholz, S. Kugathasan, S. H. Kwak, M. Laakso, T. Lehtimäki, R. J. F. Loos, S. A. Lubitz, R. C. W. Ma, D. G. MacArthur, J. Marrugat, K. M. Mattila, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, J. B. Meigs, O. Melander, A. Metspalu, B. M. Neale, P. M. Nilsson, M. C. O'Donovan, D. Ongur, L. Orozco, M. J. Owen, C. N. A. Palmer, A. Palotie, K. S. Park, C. Pato, A. E. Pulver, N. Rahman, A. M. Remes, J. D. Rioux, S. Ripatti, D. M. Roden, D. Saleheen, V. Salomaa, N. J. Samani, J. Scharf, H. Schunkert, M. B. Shoemaker, P. Sklar, H. Soininen, H. Sokol, T. Spector, P. F. Sullivan, J. Suvisaari, E. S. Tai, Y. Y. Teo, T. Tiinamaija, M. Tsuang, D. Turner, T. Tusie-Luna, E. Vartiainen, M. P. Vawter, J. S. Ware, H. Watkins, R. K. Weersma, M. Wessman, J. G. Wilson, R. J. Xavier, B. M. Neale, M. J. Daly, D. G. MacArthur and G. A. D. Consortium, The mutational constraint spectrum quantified from variation in 141,456 humans, *Nature* **581**, 434 (2020).
32. K. Chao, J. Goodrich, P. Darnowsky, R. Panchal, K. Laricchia, M. Wilson and K. Samocha, gnomAD v4.1.1 (2026), gnomAD browser news, `https://gnomad.broadinstitute.org/news/2026-03-gnomad-v4-1-1/`, posted 30 March 2026; accessed 22 September 2026.
33. M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen and F. Pontén, Tissue-based map of the human proteome, *Science* **347**, p. 1260419 (2015).
34. D. Firth, Bias reduction of maximum likelihood estimates, *Biometrika* **80**, 27 (1993).
35. P. C. Austin, Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples, *Statistics in Medicine* **28**, 3083 (2009).
36. R. Eaglesfield and K. Tokatlidis, Targeting and insertion of membrane proteins in mitochondria, *Frontiers in Cell and Developmental Biology* **9**, p. 803205 (2021).
37. C. Furusawa and K. Kaneko, Zipf's law in gene expression, *Physical Review Letters* **90**, p. 088102 (2003).
38. P. Qiu, Q. Chen, H. Qin, S. Fang, Y. Zhang, Y. Zhang, T. Xia, L. Cao, Y. Zhang, X. Fang, Y. Li and L. Hu, BioLLM: A standardized framework for integrating and benchmarking single-cell

foundation models, *Patterns* **6**, p. 101326 (2025).
39. J. M. Replogle, R. A. Saunders, A. N. Pogson, J. A. Hussmann, A. Lenail, A. Guna, L. Mascibroda, E. J. Wagner, K. Adelman, G. Lithwick-Yanai, N. Iremadze, F. Oberstrass, D. Lipson, J. L. Bonnar, M. Jost, T. M. Norman and J. S. Weissman, Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq, *Cell* **185**, 2559 (2022).
40. M. D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché and F. J. Theis, Benchmarking atlas-level data integration in single-cell genomics, *Nature Methods* **19**, 41 (2022).