Systems Biology and Network Analysis: From Multi-omics Integration to Causal and Mechanistic Understanding

Gurkan Bebek
Center for Proteomics and Bioinformatics, Case Western Reserve University, Cleveland, OH 44106, USA
E-mail: gurkan@case.edu

Rachael Hageman Blair
Department of Biostatistics, University at Buffalo, Buffalo, NY, USA
E-mail: hageman@buffalo.edu

Filipa Blasco Tavares Pereira Lopes
Mass Spectrometry and Proteomics Core, University of Utah, Salt Lake City, UT, USA
E-mail: u6071085@utah.edu

Joshua Welch
Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA
E-mail: welch@umich.edu

Onur Mutlu
Information Technology and Electrical Engineering Department, ETH Zürich, Zurich, 8092, Switzerland
E-mail: aonur.mutlu@inf.ethz.ch

Serguei Pakhomov
Department of Pharmaceutical Care & Health Systems, University of Minnesota, Minneapolis, MN 55455, USA
E-mail: pakh0002@umn.edu

1. Session Introduction: Systems Biology and Network Analysis

Advances in high-throughput molecular profiling continue to generate increasingly complex datasets spanning multiple biological scales. These developments have created new opportunities to investigate biological systems through integrative computational approaches that combine molecular measurements with network, pathway, and knowledge-based representations. At the same time, extracting mechanistic insight from these data remains a fundamen-

© 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.

tal challenge. The papers presented in this session address this challenge through advances in network analysis, representation learning, causal knowledge modeling, perturbation analysis, and multi-modal integration. Collectively, they illustrate how systems-level approaches can facilitate biological discovery, generate mechanistic hypotheses, and support translational applications.

2. Session Summary

This session includes an invited talk and twelve peer-reviewed papers selected for oral and poster presentations. The contributions span multiple areas of systems biology, including single-cell modeling, multi-omics integration, network propagation, causal knowledge representation, and translational network biology. Although the methodological approaches vary, a common theme is developing computational frameworks that connect large-scale molecular data to biological mechanisms.

2.1. *Representation Learning and Foundation Models*

Several contributions focus on the opportunities and challenges associated with large-scale representation learning. Whalley examines gene embedding structures across several widely used single-cell foundation models and demonstrates that embedding outliers are not consistently associated with biologically important genes, highlighting limitations in the biological interpretation of learned representations. Kim and Hwang address computational scalability by introducing a cell-adaptive token allocation strategy that substantially reduces the computational cost of single-cell foundation model inference while preserving downstream performance. Together, these studies illustrate the growing role of foundation models in computational biology while emphasizing the need for both interpretability and efficiency.

2.2. *Multi-Omics Integration and Dynamic Modeling*

A second group of papers focuses on integrating heterogeneous molecular data. Mo and Olshen present iClusterPlus2, an extension of the iCluster framework that enables scalable integration of bulk and single-cell multi-omics datasets through shared latent representations. Zhang et al. introduce scMultiNODE, which combines neural ordinary differential equations with optimal transport methods to model cellular trajectories from unaligned single-cell modalities. Zheng et al. investigate relationships between polygenic and proteomic risk scores and show that network-based approaches can reveal complementary biological signals that are not apparent when each data source is considered independently. These studies highlight continuing efforts to bridge multiple molecular layers within unified analytical frameworks.

2.3. *Perturbation Analysis and Network Propagation*

Understanding how perturbations propagate through biological systems remains a central challenge in systems biology. Liang and Singh introduce COMPASS, a framework that decomposes perturbation responses into shared and perturbation-specific components in large-scale Perturb-seq experiments. By separating common cellular responses from gene-specific effects,

the method improves interpretation of regulatory perturbation data. Bobak et al. examine signal propagation in co-expression networks by adapting concepts from network interference and spillover analysis. Their work shows that interactions between network modules can reveal biologically relevant signals that direct measurements alone may not detect.

2.4. *Knowledge Graphs and Mechanistic Representation*

Several papers emphasize the importance of structured biological knowledge for mechanistic interpretation. Hoehndorf et al. introduce PhysioMap, a causal knowledge graph that represents physiological processes and supports analyses based on mechanistic relationships and interventions. Caufield et al. present KGBioportal, a platform that transforms biomedical ontologies into graph-based resources that can be incorporated into knowledge graph applications, network analyses, and machine learning workflows. Together, these contributions reflect increasing interest in combining data-driven approaches with formal biological knowledge representations.

2.5. *Translational Applications and Methodological Evaluation*

The remaining papers demonstrate how network-based approaches can support translational research while also highlighting the importance of methodological rigor. Karunakaran et al. reconstruct a GPX4-centered interaction network to investigate disease mechanisms and identify potential therapeutic opportunities in Sedaghatian-type spondylometaphyseal dysplasia. Nemecio and Brubaker integrate microbial, host, and pharmacological information to study microbial drug metabolism across human-associated microbiomes. Debnath and Ghosh examine benchmark datasets used for drug-synergy prediction and show that commonly used validation strategies can substantially overestimate performance because of information leakage. Collectively, these studies illustrate both the promise of network-based discovery and the importance of careful model evaluation.

3. Summary

Several themes emerge across the papers in this session. Biological networks continue to provide a common framework for integrating heterogeneous datasets, interpreting machine-learning models, and connecting molecular measurements to higher-level phenotypes. At the same time, there is increasing emphasis on approaches that support mechanistic interpretation rather than prediction alone, whether through perturbation modeling, causal knowledge representation, or multi-modal data integration. The contributions presented in this session illustrate how advances in computational methods are helping bridge the gap between large-scale molecular data and biological mechanisms.

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,

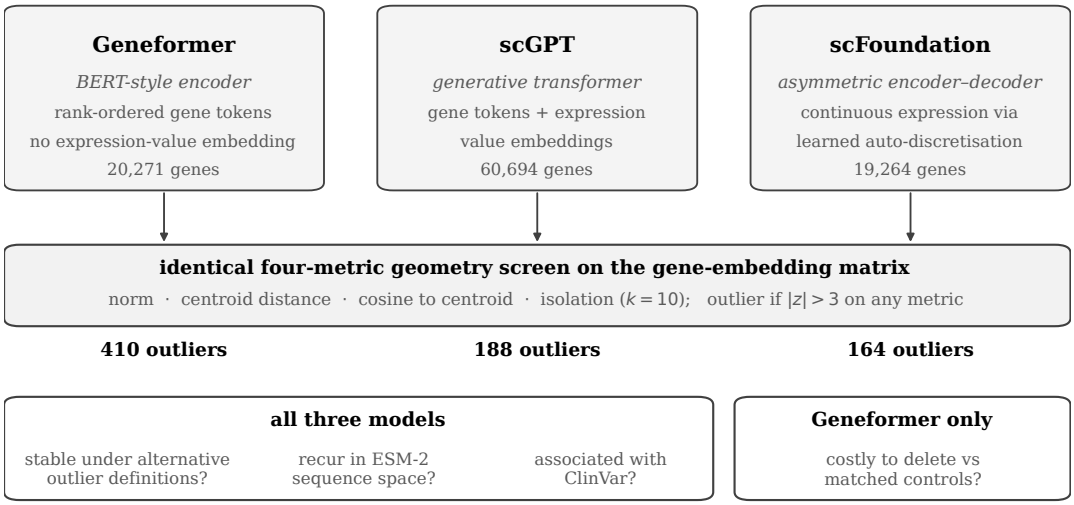


Fig. 1. The three models and the shared screen. The models differ in architecture, expression encoding and vocabulary size, and also in pretraining objective, corpus and scale, which are not shown. Robustness to alternative outlier definitions, ESM-2 recurrence and covariate-adjusted ClinVar association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

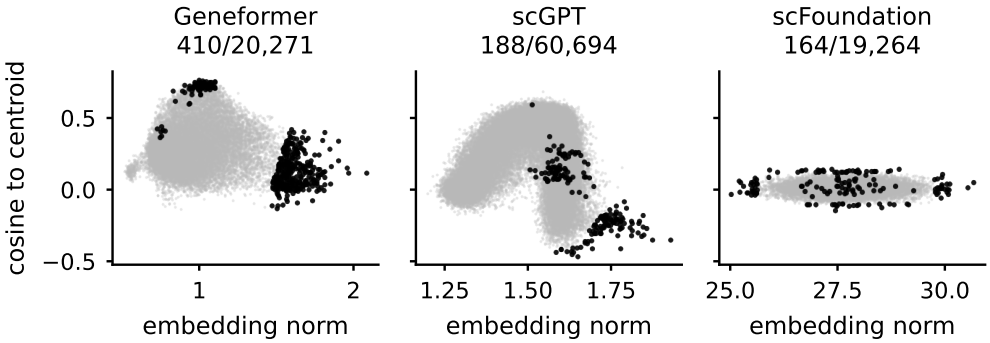


Fig. 2. Gene-embedding geometry in each model, with called outliers in black; the cosine axis is shared, the norm axis is model-specific. Geneformer and scGPT each separate into two clouds along the norm axis, the structure behind scGPT's sensitivity to the outlier definition (Section 3.2); scFoundation 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

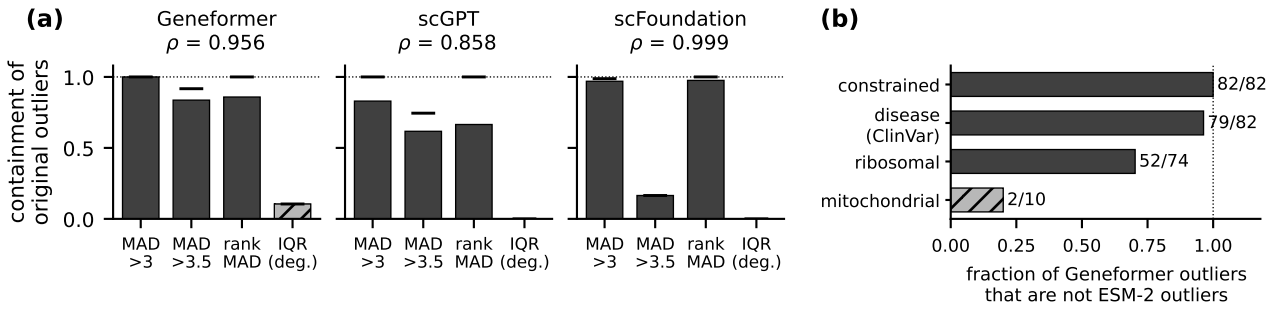


Fig. 3. (a) Outlier identity under alternative outlier definitions. Bars give containment, the fraction of original $|z| > 3$ outliers recovered; ticks mark the maximum attainable containment $\min(n_{\text{new}}, n_{\text{orig}})/n_{\text{orig}}$, so a bar meeting its tick is saturated, not unstable; 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.

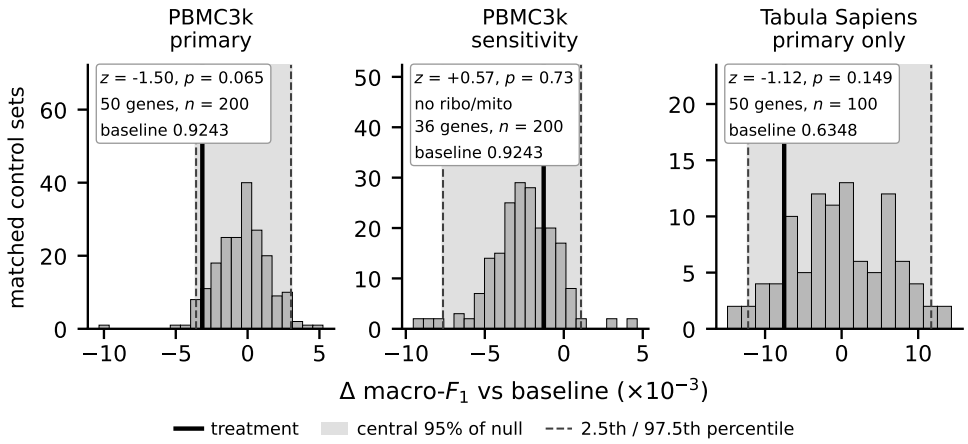


Fig. 4. Matched-control deletion results for Geneformer. Each panel gives Δ macro-F_1 for gene sets matched to the treatment on expression level, breadth, gene length and class; shading is the 2.5th–97.5th percentile null band, the vertical line is the treatment, and panel labels give z and the one-sided 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).

AdaGeneBudget: Cell-Adaptive Gene-Token Allocation for Efficient Single-Cell Foundation Models

Dohee Kim and Uiwon Hwang†

Division of Artificial Intelligence & Software, Ewha Womans University, Seoul, 03760, Republic of Korea

†*E-mail: uiwon.hwang@ewha.ac.kr*

Abstract

Single-cell foundation models (scFMs) represent each cell using sequences of gene-associated tokens, making embedding extraction increasingly costly as the number of cells and expressed genes grows. Existing input policies typically rely on fixed input budgets, with retained genes determined by random subsampling, model-native ranking, or a dataset-level highly variable gene (HVG) panel. However, they do not jointly determine, for each cell, which genes to retain and how many tokens to allocate.

We introduce **AdaGeneBudget**, a training-free gene-token selection method that combines each gene's expression with reference-derived inverse detection frequency and retains the shortest ranked prefix that captures a target fraction of the cell's expression-specificity score mass. The resulting cell-specific budget is bounded by minimum and maximum lengths, requires no cell-type labels, and leaves the pretrained backbone unchanged.

We evaluated AdaGeneBudget in a frozen-backbone inference setting using pretrained scGPT and Geneformer models on Kang and PBMC reference-mapping tasks, with an additional scPRINT comparison against its official HVG policy and an expressed-only HVG control. Across four scGPT and Geneformer backbone–dataset pairs, AdaGeneBudget substantially reduced mean gene-token counts and peak GPU memory while increasing embedding-extraction throughput by up to 4.63×. Despite this compression, it preserved native-level aggregate annotation utility and numerically outperformed token-matched random and Top-Expression baselines. AdaGeneBudget also preserved fine-grained and low-support cell identities and retained lineage-marker programs and stimulation-associated pathway genes under compression. In scPRINT, both HVG controls achieved higher annotation macro-F1, whereas AdaGeneBudget more faithfully preserved the stimulation-induced embedding direction. These results establish biologically informed, cell-adaptive gene-token allocation as a practical complement to architectural and systems-level efficiency methods for applying existing scFMs to new datasets. They also suggest that cell-adaptive input allocation may be a useful design principle for future models operating under finite token budgets.

Keywords: single-cell foundation models; efficient inference; gene selection; adaptive token allocation; single-cell RNA sequencing

© 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.

1. Introduction

The scale of single-cell RNA sequencing (scRNA-seq) datasets has increased rapidly, with public resources now aggregating tens of millions of cells across tissues and diseases.[1] This growth has enabled single-cell foundation models (scFMs) to learn transferable gene and cell representations from large single-cell transcriptomic corpora. Geneformer and scGPT were pretrained on approximately 30 million and more than 33 million profiles, respectively, while scFoundation and scPRINT scale to 50 million or more cells.[2–5] Once pretrained, these models provide reusable cell-level embeddings for downstream analyses such as clustering, cell-type annotation, and batch integration. As scFMs are applied to increasingly large atlas- and cohort-scale datasets, repeatedly extracting embeddings from long per-cell gene-token sequences can become a practical bottleneck under finite GPU memory and computation.

Despite differences in tokenization and expression encoding, many transformer-based scFMs represent each cell as a sequence of gene tokens. The number of vocabulary-matched expressed genes can vary substantially across cells and datasets and reach thousands,[2,3,5] while pairwise computation in standard dense self-attention scales quadratically with sequence length.[6] These computational demands become particularly important at atlas or cohort scale: more cells increase the total workload and longer per-cell sequences increase computation and memory per cell. Cell subsampling changes the set of cells retained for analysis, and uniform random subsampling may underrepresent rare populations or incompletely capture transcriptomic heterogeneity.[7] We instead retain every cell while reducing the number of gene tokens used to represent each cell.

Outside single-cell modeling, related approaches include instance-wise feature selection, sparsity-inducing feature selection, and dynamic Transformer token pruning.[8–10] These approaches typically rely on learned task-dependent selection mechanisms or pruning modules within the model. In contrast, AdaGeneBudget selects gene tokens before the native input pipeline of a frozen scFM, without modifying the backbone or requiring additional training.

Existing scFMs use model-specific but largely fixed input-construction rules. Geneformer rank-orders detected genes and truncates long inputs,[2] whereas the official scGPT embedding pipeline randomly subsamples nonzero gene tokens when its maximum sequence length is exceeded.[3,11] scPRINT was pretrained with randomly sampled gene contexts[5] and also supports a dataset-level HVG policy that selects a fixed panel of highly variable genes.[12] Random subsampling is simple but does not explicitly prioritize genes by cell-specific relevance. HVG selection provides a practical global policy because it ranks genes by their standardized cell-to-cell variability within the reference dataset and yields a shared ranking and panel across cells rather than adapting them to each cell.[13]

More specialized single-cell feature-selection and targeted-panel methods also construct compact gene sets: SCMER selects features to preserve cellular manifold structure, geneBasis selects a targeted gene panel to preserve transcriptomic structure, and PERSIST learns gene panels for transcriptome reconstruction or task-specific prediction.[14–16] These methods typically produce gene sets shared across cells for their respective objectives, whereas AdaGeneBudget performs cell-specific gene prioritization and budget allocation to reduce gene-token processing cost while preserving downstream utility in an existing frozen scFM. Moreover, a

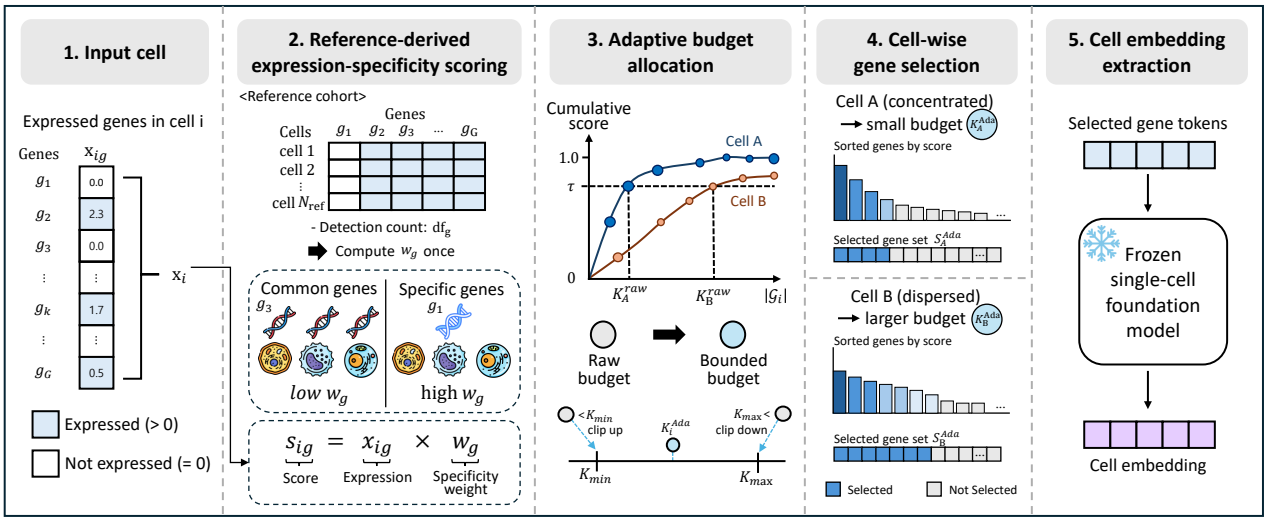


Fig. 1. **Overview of AdaGeneBudget.** AdaGeneBudget computes reference-derived gene specificity, ranks expressed genes using expression-specificity scores, and assigns a cell-specific token budget according to cumulative score mass before constructing the native model input.

common maximum length does not determine how many tokens are sufficient for each cell: some cells may be represented by a compact gene set, whereas others may require a broader one. This motivates treating input construction as two coupled, cell-specific decisions—which genes to retain and how many tokens to allocate—while bounding the resulting budgets to avoid extreme input lengths.

We introduce **AdaGeneBudget**, a training-free gene-token selection method for pretrained scFMs. As illustrated in Fig. 1, AdaGeneBudget combines reference-derived inverse detection-frequency weights with each cell's expression values and retains the shortest ranked prefix that reaches a target score mass within budget bounds. Cells with concentrated score mass therefore receive shorter inputs, whereas cells with more diffuse score mass receive longer inputs. The method requires no cell-type labels, estimates gene prevalence from the reference cohort only, and passes selected genes through each backbone's native ordering and expression-encoding pipeline. We evaluate input policies with frozen backbones under the same reference-mapping protocol, isolating the effects of input construction from task-specific parameter adaptation, and show that AdaGeneBudget reduces embedding-extraction cost while preserving annotation utility and biologically relevant signals under gene-token compression.

Our contributions are as follows. First, we formulate gene-token input construction in scFMs as two coupled cell-specific decisions: which genes to prioritize and how many tokens to allocate to each cell, without model retraining. Second, across scGPT and Geneformer, we demonstrate computational savings while preserving reference-mapping utility, using the fixed-budget Top-Expression baseline and the Fixed TF–IDF ablation to help distinguish the contribution of gene prioritization from that of adaptive budget allocation. Third, we show that its benefits extend beyond aggregate annotation, including preservation of fine-grained cell identities, lineage-marker programs, and stimulation-associated signals. Code and supplementary materials are available at `https://github.com/D0heeKim/AdaGeneBudget`.

2. Method

2.1. *Problem formulation*

For cell i, let $\mathbf{x}_i = (x_{i1}, \ldots, x_{iG}) \in \mathbb{R}^G_{\geq 0}$ denote its expression vector, where x_{ig} is the expression value used to score gene g. Let $\mathcal{V} \subseteq \{1, \ldots, G\}$ denote the set of genes recognized by the pretrained backbone. The eligible expressed-gene set and its size are

$$\mathcal{G}_i = \{g \in \mathcal{V} \mid x_{ig} > 0\}, \qquad n_i = |\mathcal{G}_i|. \tag{1}$$

Under native input construction, each backbone constructs its model-specific input sequence from $\mathcal{G}_i$ according to its gene ordering, expression encoding, and input-length conventions. Our objective is instead to select a subset $\mathcal{S}_i \subseteq \mathcal{G}_i$, with $K_i = |\mathcal{S}_i|$, to reduce gene-token processing cost while preserving downstream utility and biologically relevant information in the resulting frozen-backbone representation.

2.2. *Reference-derived expression-specificity score*

Expression magnitude captures within-cell abundance but not how frequently a gene is detected across the reference population. We therefore combine each gene's expression value with an inverse reference-detection-frequency weight, following classical inverse-frequency weighting principles from information retrieval,[17,18] yielding a TF–IDF-inspired expression-specificity score that reflects within-cell expression and inverse prevalence across the reference cohort.

Let the reference cohort contain N_{ref} cells. The reference detection count of gene g, analogous to document frequency, is

$$\mathrm{df}_g = \sum_{r=1}^{N_{\text{ref}}} \mathbb{I}[x_{rg} > 0], \tag{2}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function. We then define the smoothed inverse-frequency weight for gene g and its expression-specificity score in cell i as

$$w_g = \log\left(\frac{N_{\text{ref}} + 1}{\mathrm{df}_g + 1}\right) + 1, \tag{3}$$

$$s_{ig} = x_{ig} w_g, \qquad g \in \mathcal{G}_i. \tag{4}$$

As df_g increases, w_g decreases toward 1; genes detected in many reference cells are therefore upweighted less, although they may still rank highly when x_{ig} is large.

We refer to s_{ig} as an *expression-specificity score*, rather than standard TF–IDF, because x_{ig} represents single-cell expression rather than textual term frequency. The weight vector is estimated once from the reference cohort and reused for both reference and held-out query cells. For each query cell, s_{ig} is computed from its own expression values using the fixed reference-derived weights; query-cohort prevalence and cell-type labels from either cohort are not used for ranking or selection.

2.3. *Cell-adaptive score-mass allocation*

For each cell i, let $g_i^{(1)}, \ldots, g_i^{(n_i)}$ denote the eligible genes ordered by decreasing expression-specificity score:

$$s_{i,g_i^{(1)}} \geq s_{i,g_i^{(2)}} \geq \cdots \geq s_{i,g_i^{(n_i)}}. \tag{5}$$

Ties are resolved deterministically according to each backbone's fixed gene order. The normalized score mass captured by the top k genes is

$$M_i(k) = \frac{\sum_{j=1}^{k} s_{i,g_i^{(j)}}}{\sum_{g \in \mathcal{G}_i} s_{ig}}. \tag{6}$$

For a target score-mass threshold $\tau \in (0, 1]$, the unconstrained cell-specific budget is the length of the shortest prefix whose score mass reaches that threshold. Given $1 \leq K_{\min} \leq K_{\max}$, the unconstrained budget, the realized budget, and the selected gene set are defined as

$$K_i^{\text{raw}} = \min\{k \in \{1, \ldots, n_i\} \mid M_i(k) \geq \tau\}, \tag{7}$$

$$K_i^{\text{Ada}} = \min\{n_i,\, K_{\max},\, \max(K_{\min}, K_i^{\text{raw}})\}, \tag{8}$$

$$\mathcal{S}_i^{\text{Ada}} = \left\{g_i^{(1)}, \ldots, g_i^{(K_i^{\text{Ada}})}\right\}. \tag{9}$$

A cell with concentrated score mass reaches τ with a shorter prefix, whereas more diffuse score mass requires a longer prefix within the budget bounds. All main experiments use $\tau = 0.90$, $K_{\min} = 128$, and $K_{\max} = 600$. Before applying the bounds, setting $\tau = 0.90$ yields the shortest prefix that captures at least 90% of the cell's total expression-specificity score mass; it does not retain a fixed proportion of the expressed genes. Sensitivity analyses are reported in Supplementary Sec. S1.

2.4. *Backbone integration and computational implications*

AdaGeneBudget operates before backbone-specific sequence construction and changes only the gene subset supplied to the backbone. Let T_m denote the sequence-construction function for backbone m, which maps the selected genes and their associated expression information to the model-specific input sequence, and let f_{θ_m} denote the corresponding embedding function. The resulting cell embedding is

$$\mathbf{z}_i^{(m)} = f_{\theta_m}\left(T_m\left(\mathcal{S}_i^{\text{Ada}}, \mathbf{x}_{i,\mathcal{S}_i^{\text{Ada}}}\right)\right), \tag{10}$$

where $\mathbf{x}_{i,\mathcal{S}_i^{\text{Ada}}}$ denotes the expression values of cell i restricted to the selected genes. After selection, each backbone's native conventions for gene ordering, expression encoding, and special-token handling are retained; model-specific special tokens are not counted in K_i^{Ada}. All backbone parameters remain fixed during embedding extraction. This separation allows AdaGeneBudget to be applied to other gene-token-based scFMs whose input pipelines accept an externally specified gene subset.

Given precomputed reference-derived weights, expression-specificity scoring requires $O(n_i)$ time, and the deterministic full ranking used in our implementation requires $O(n_i \log n_i)$ time. Let L_i^{native} and L_i^{Ada} denote the realized unpadded input lengths, including model-specific special tokens, under native input construction and AdaGeneBudget, respectively. For the

same backbone, reducing the input length reduces the pairwise token-interaction term in dense self-attention to approximately $\left(L_i^{\mathrm{Ada}}/L_i^{\mathrm{native}}\right)^2$ of its native value. This approximation applies only to the pairwise attention term; end-to-end runtime and memory also depend on selection overhead, token-wise computation, and batch padding. We therefore report mean realized gene-token count, end-to-end throughput, and peak GPU memory in Section 4.1, with selection overhead included in the throughput measurements.

3. Experimental setup

Datasets and backbones. We evaluated AdaGeneBudget on Kang[19] (19,583 reference cells from seven donors and 5,090 query cells from held-out donor 1015) and the PBMC RNA dataset from the Seurat v4 multimodal atlas[20] (132,137 reference cells and 19,957 query cells from held-out donor P5), with both held-out splits fixed a priori based on dataset composition. The primary backbones were the scGPT whole-human checkpoint[3,11] and Geneformer V1-10M.[2,21] We additionally evaluated scPRINT on Kang because its dataset-level HVG option enables a same-backbone comparison between AdaGeneBudget and global HVG selection.[5,12] We also audited evaluation-study overlap with documented pretraining corpora. PBMC overlaps with the source corpus used by scGPT, whereas no documented Kang source-study overlap was identified for scGPT or scPRINT. Kang and PBMC were absent from the documented Genecorpus-30M used by Geneformer V1. To assess transfer beyond the two primary immune-cell datasets, we additionally evaluated AdaGeneBudget on a triple-negative breast cancer (TNBC) dataset.[22] Experimental details and results are provided in Supplementary Sec. S5.

Input policies and evaluation. We compared AdaGeneBudget with Native Bucketed, Matched Random, and Top-Expression across both backbones, and additionally with Native Rank for Geneformer. Native Bucketed used each backbone's native input policy with the same length-aware bucketing applied to all methods. Matched Random uniformly sampled eligible expressed genes without replacement using a fixed budget matched to AdaGeneBudget's mean realized budget on the reference cohort. Top-Expression retained, for each cell, the highest-expressed eligible genes using the same reference-fitted fixed budget. For Geneformer, Native Rank retained genes in the backbone's native rank-value order using the same fixed budget. Fixed budgets were matched to AdaGeneBudget's mean realized budget on the reference cohort and then held fixed for query cells, without using query-set information for budget selection. For scPRINT, reference-only global HVG controls were computed using the Seurat v3 procedure;[13] the complementary budget-matching schemes are described in Sec. 4.4.

Reference and query cells were embedded under each policy, and query cells were annotated by cosine 5-nearest-neighbor majority voting over the labeled reference embeddings; query labels were used only as ground truth for evaluation. Throughput included method-specific selection and bucketing but excluded shared data and model loading. Measurements were repeated five times, speedups were computed within each backbone–dataset pair, and all inference experiments used an NVIDIA RTX A6000 GPU.

Biological-signal preservation metrics. PBMC lineage programs comprised 20 donor-consistent markers per lineage derived from the reference donors. Marker coverage was the

fraction of a cell's expressed lineage-program genes retained, and enrichment was coverage divided by the uniform-retention expectation K_i/n_i. Cells lacking expressed genes from their corresponding program were excluded, and cell-level values were averaged within each lineage.

For Kang, paired-donor pseudobulk differential expression was performed in CD14+ monocytes from the seven reference donors, and genes with $q < 0.05$ and $|\log_2 \mathrm{FC}| \geq 1$ were tested for one-sided Fisher over-representation in MSigDB Hallmark gene sets.[23] After Benjamini-Hochberg correction,[24] the 12 pathways with the smallest adjusted P values were selected for analysis. For each significant DE gene belonging to a selected pathway, mapped to the Geneformer input space and expressed in at least one held-out CD14+ monocyte, conditional retention was the fraction of expressing control and stimulated cells retaining that gene, and pathway retention was the unweighted mean across genes.

For scPRINT, Hallmark IFN-α response retention was the pooled fraction of expressed gene-cell pairs retained as positive input tokens in stimulated query cells.[23] For cell type c and input policy m, perturbation-direction preservation was the cosine similarity between $\Delta_{m,c} = \bar{\mathbf{z}}_{m,c,\mathrm{stim}} - \bar{\mathbf{z}}_{m,c,\mathrm{ctrl}}$ and the corresponding all-expressed reference direction from the same backbone, averaged equally across cell types with at least 20 query cells per condition.

4. Results

4.1. *AdaGeneBudget reduces gene-token processing cost*

Table 1 summarizes query-cell mean selected-gene counts, end-to-end embedding-pipeline throughput, and peak GPU memory, with the same length-aware bucketing applied to all policies and comparisons restricted within each backbone–dataset pair. Across the four backbone–dataset pairs, AdaGeneBudget reduced mean gene-token counts by 27.6–69.1%, yielding 1.19–4.63× speedups and 42.5–70.6% lower peak memory than Native Bucketed.

Table 1. Computational efficiency. Speedup is relative to Native Bucketed within each backbone—dataset pair. Fixed-budget matching used reference cells, so query-set mean lengths may differ.

Backbone	Dataset	Input policy	Mean genes ↓	Throughput (cells/s) ↑	Speedup	Peak memory (GB) ↓
scGPT	Kang	Native Bucketed	564.1	1129.9 ± 2.2	1.00×	1.376
		Matched Random	391.9	1438.7 ± 3.4	1.27×	0.594
		Top-Expression	391.9	1447.8 ± 38.5	1.28×	0.594
		AdaGeneBudget	405.8	1349.5 ± 2.1	1.19×	0.791
	PBMC	Native Bucketed	1184.7	443.9 ± 0.8	1.00×	1.376
		Matched Random	598.3	1031.2 ± 0.7	2.32×	0.788
		Top-Expression	598.3	990.2 ± 3.1	2.23×	0.788
		AdaGeneBudget	596.3	940.7 ± 5.1	2.12×	0.791
Geneformer	Kang	Native Bucketed	558.3	810.7 ± 7.4	1.00×	13.527
		Native Rank	390.3	1243.7 ± 0.5	1.53×	2.654
		Matched Random	390.3	1199.0 ± 0.4	1.48×	2.654
		Top-Expression	390.3	1192.2 ± 8.8	1.47×	2.654
		AdaGeneBudget	404.2	1101.6 ± 0.4	1.36×	3.981
	PBMC	Native Bucketed	1927.5	146.7 ± 2.7	1.00×	13.527
		Native Rank	598.2	721.6 ± 2.1	4.92×	3.971
		Matched Random	598.2	744.1 ± 1.0	5.07×	3.971
		Top-Expression	598.2	716.1 ± 4.9	4.88×	3.971
		AdaGeneBudget	596.1	678.6 ± 0.3	4.63×	3.977

Matched Random was 6.6–9.7% faster because it avoided expression-specificity scoring, full ranking, and adaptive cutoff determination. On Geneformer–Kang, AdaGeneBudget selection accounted for 14.8% of the end-to-end embedding-pipeline time; within selection, the TF–IDF and native-order sorts dominated the profiled cost (68.3%) compared with 12.2% to scoring and 11.2% to cutoff determination. Nevertheless, AdaGeneBudget retained 91.2–93.8% of Matched Random throughput. Its higher peak memory than fixed-budget policies with similar mean lengths reflects the longest padded batches, as variable allocation allows cells to reach $K_{\max}$.

4.2. *AdaGeneBudget preserves reference-mapping annotation utility under compression*

Under the reference-mapping protocol in Section 3, AdaGeneBudget numerically outperformed Matched Random and Top-Expression on all three annotation metrics and all four backbone–dataset pairs (Table 2), with macro-F1 gains of 2.24–18.04 percentage points over Matched Random. Despite substantially compressed inputs, it also numerically exceeded Native Bucketed on all three metrics across all four backbone–dataset pairs and Geneformer Native Rank in both datasets, indicating that which genes are retained matters, not only how many.

Table 2. Reference-to-query cell-type annotation performance. Stochastic policies are reported as mean ± standard deviation across selection seeds, whereas deterministic policies are reported once. Bold indicates the best numerical result within each backbone–dataset pair.

Backbone	Dataset	Input policy	Mean genes	Accuracy ↑	Macro-F1 ↑	Balanced accuracy ↑
scGPT	Kang	Native Bucketed	564.1	0.9481 ± 0.0014	0.8406 ± 0.0023	0.8368 ± 0.0037
		Matched Random	391.9	0.9352 ± 0.0010	0.8314 ± 0.0154	0.8263 ± 0.0192
		Top-Expression	391.9	0.9434	0.8339	0.8318
		AdaGeneBudget	405.8	**0.9493**	**0.8538**	**0.8685**
	PBMC	Native Bucketed	1184.7	0.9169 ± 0.0017	0.8741 ± 0.0030	0.8588 ± 0.0033
		Matched Random	598.3	0.8843 ± 0.0018	0.8345 ± 0.0027	0.8198 ± 0.0024
		Top-Expression	598.3	0.9244	0.8861	0.8681
		AdaGeneBudget	596.3	**0.9338**	**0.8932**	**0.8771**
Geneformer	Kang	Native Bucketed	558.3	0.8804	0.6614	0.6577
		Native Rank	390.3	0.8617	0.7012	0.6753
		Matched Random	390.3	0.8370 ± 0.0025	0.5905 ± 0.0049	0.5821 ± 0.0079
		Top-Expression	390.3	0.8729	0.7091	0.6910
		AdaGeneBudget	404.2	**0.8857**	**0.7709**	**0.7563**
	PBMC	Native Bucketed	1927.5	0.8408	0.7685	0.7712
		Native Rank	598.2	0.8239	0.7557	0.7585
		Matched Random	598.2	0.7352 ± 0.0015	0.6198 ± 0.0037	0.6161 ± 0.0036
		Top-Expression	598.2	0.8299	0.7596	0.7606
		AdaGeneBudget	596.1	**0.8434**	**0.7727**	**0.7758**

A Fixed TF–IDF ablation using the same score and ranking as AdaGeneBudget but a shared fixed budget supported the contribution of adaptive allocation: on Geneformer–Kang, AdaGeneBudget increased macro-F1 from 0.7130 to 0.7709 under comparable realized budgets. The default configuration remained close to the best tested across all four backbone–dataset pairs (Supplementary Sec. S1). Supplementary Sec. S2–S4 show that larger macro-F1 gains were supported by bootstrap intervals, AdaGeneBudget better preserved local, within-

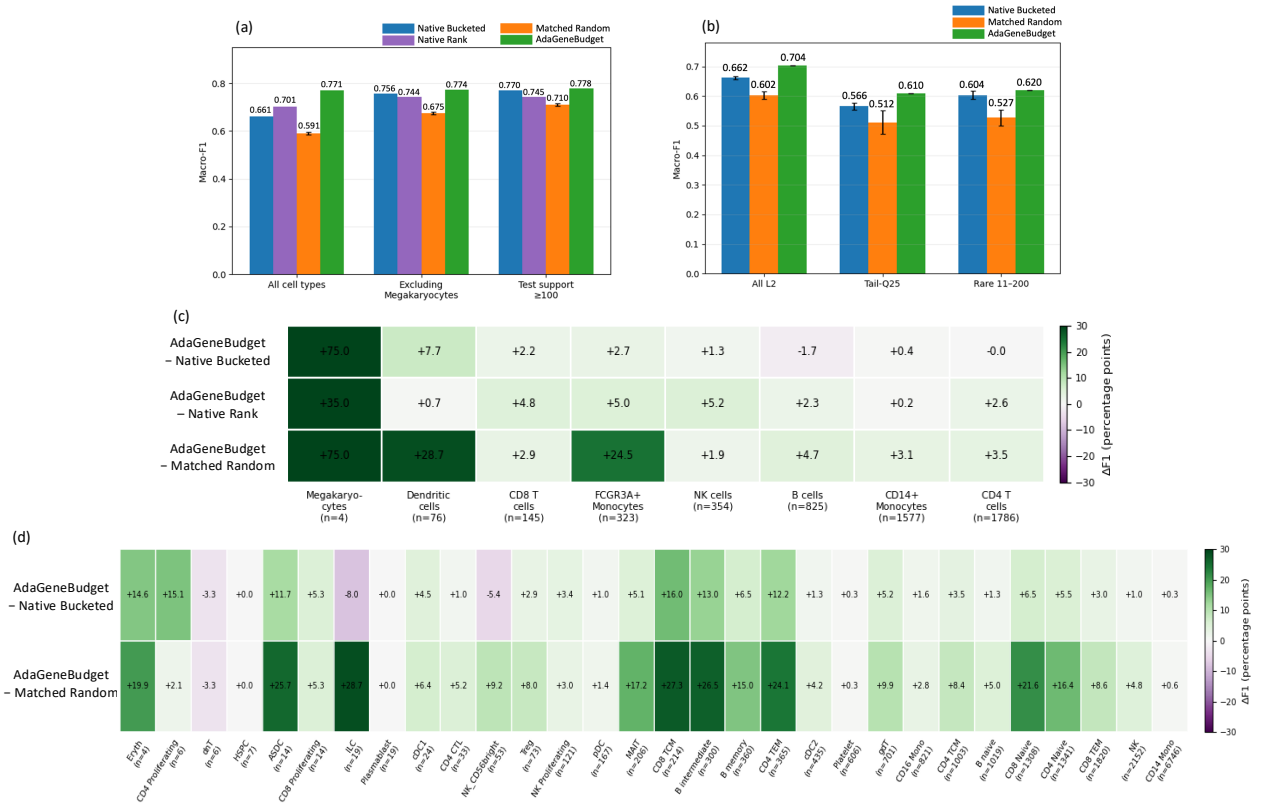


Fig. 2. **Fine-grained and low-support annotation under gene-token reduction.** (a) Geneformer–Kang L1 macro-F1 across all cell types and support-restricted subsets. (b) scGPT–PBMC L2 macro-F1 across all subtypes and low-support subsets. (c,d) Per-population F1 differences, in percentage points, between AdaGeneBudget and the indicated comparators for Geneformer–Kang (c) and scGPT–PBMC (d), ordered by support. Error bars show standard deviation across five seeds for stochastic policies; deterministic policies are single runs.

cell-type, and expression-referenced structure than Matched Random and tested reference variations had limited effects on Geneformer–Kang.

Supplementary donor-disjoint validation on a 427,823-cell TNBC cohort further extended the efficiency–utility tradeoff beyond the two primary immune-cell datasets. AdaGeneBudget achieved 1.80× and 3.46× speedups over Native Bucketed with scGPT and Geneformer, respectively, while attaining the highest numerical macro-F1 among the evaluated policies for both backbones (0.5411 and 0.3392; Supplementary Sec. S5).

4.3. *AdaGeneBudget preserves fine-grained and low-support cell identities*

Aggregate annotation metrics can mask failures in fine-grained or low-support populations. We therefore examined Geneformer–Kang performance by L1 support and scGPT–PBMC performance using finer L2 labels (Fig. 2).

On Kang, AdaGeneBudget achieved the highest macro-F1 across all cell types (0.771), after excluding Megakaryocytes (0.774) and among classes with at least 100 query cells (0.778) (Fig. 2a,c). Thus, the aggregate gain was not driven by an extremely low-support population.

For Megakaryocytes ($n = 4$), AdaGeneBudget correctly annotated three cells (F1=0.750), compared with F1=0 for Native Bucketed and Matched Random and 0.400 for Native Rank; given the very small support, we treat this result as an illustrative example rather than independent evidence of rare-cell robustness.

On PBMC, AdaGeneBudget likewise achieved the highest macro-F1 across all 30 L2 subtypes (0.704), in Tail-Q25 (0.610), and among subtypes with 11–200 cells (0.620) (Fig. 2b). Among subtypes with more than 10 query cells, it exceeded Native Bucketed for 23 and Matched Random for 25, including all 11 T-cell-related subtypes (Fig. 2d). The only subtypes above 10-cell support for which AdaGeneBudget was lower than Native Bucketed were ILC and NK CD56$^{\text{bright}}$, by 8.0 and 5.4 points, respectively, while remaining above Matched Random by 28.7 and 9.2 points. These exceptions may reflect the difficulty of distinguishing related lymphoid populations, since human ILC states are transcriptionally heterogeneous and some ILC1 and NK states show transcriptional similarity.[25,26] Moreover, the original PBMC subtype labels were derived from joint RNA and surface-protein analysis,[20] which may make exact recovery from RNA-only embeddings challenging. Among the four subtypes below the 11-cell support, only dnT was lower than both comparators (3.3 points); with only six cells, both Native Bucketed and Matched Random achieved mean F1 of 0.033, with a nonzero result in just one of five seeds for each policy. We therefore do not interpret this small difference as a stable comparator advantage. Overall, AdaGeneBudget preserved fine-grained and low-support identities across support-restricted analyses, with isolated exceptions confined to extremely small or closely related lymphoid populations.

4.4. *AdaGeneBudget preserves lineage and stimulation-associated signals under compression*

Beyond annotation utility, we evaluated whether compression retained stimulation-associated pathways, lineage-marker programs, and stimulation-induced embedding shifts (Fig. 3).

In Geneformer–Kang, AdaGeneBudget achieved mean conditional retention of approximately 0.81, compared with 0.67 for Native Rank and 0.61 for Matched Random (Fig. 3a). It exceeded Native Rank in 10 of 12 pathways and Matched Random in all 12, with high retention for IL6–JAK–STAT3, TNFα–NFκB, and KRAS signaling.

In Geneformer–PBMC, AdaGeneBudget retained approximately 0.82 mean lineage-marker coverage versus 0.92 for Native Bucketed, 0.59 for Native Rank, and 0.26 for Matched Random (Fig. 3b). This recovered about 89% of native coverage while reducing mean input from 1,927.5 to 596.1 genes and exceeded comparators across all six lineages. Bootstrap uncertainty estimates for both biological-preservation analyses are provided in Supplementary Sec. S2.

With scPRINT, the official fixed-panel HVG control was matched to AdaGeneBudget's mean total gene-token count on the reference cohort, including zero-valued panel entries. HVG achieved higher macro-F1 (0.767 versus 0.739), whereas AdaGeneBudget retained more Hallmark IFN-α-response genes (0.718 versus 0.178) and more faithfully preserved the stimulation-induced embedding direction (0.975 versus 0.408; Fig. 3c).

To separate the effect of zero-expression overhead, we also evaluated an expressed-only HVG control by expanding the reference-fitted global candidate panel to 11,929 genes to

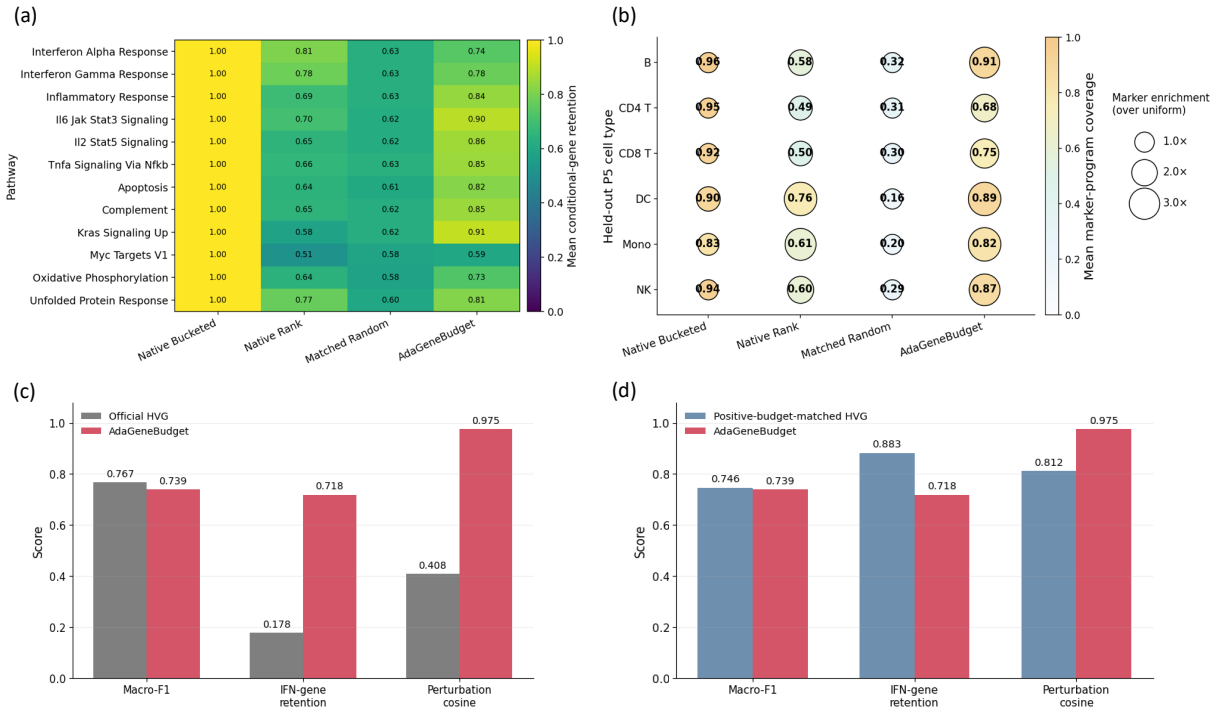


Fig. 3. **Biological-signal preservation under compression.** (a) Geneformer–Kang mean conditional retention of stimulation-associated pathway genes. (b) Geneformer–PBMC lineage-marker coverage in held-out donor P5 (color and values) and enrichment over uniform retention at each policy's realized budget (bubble area). (c) scPRINT's reference-only Seurat-v3 fixed-panel HVG policy versus AdaGeneBudget under a matched total-token budget. (d) Expressed-only HVG versus AdaGeneBudget under a matched positive-gene budget. Panels (c,d) report macro-F1, IFN-α-response retention, and cosine similarity to the all-expressed stimulation direction.

match AdaGeneBudget's mean positive-input length (Fig. 3d). HVG achieved slightly higher macro-F1 (0.746 versus 0.739) and IFN-response retention (0.883 versus 0.718), whereas AdaGeneBudget better preserved the stimulation direction (0.975 versus 0.812), with higher similarity in all seven eligible cell types. Thus, the direction-preservation advantage persisted even under this broader expressed-only HVG control.

5. Discussion

We introduced AdaGeneBudget as a training-free method that treats gene-token input construction in scFMs as a cell-specific allocation problem. Across scGPT and Geneformer, it reduced embedding-extraction cost while preserving native-level annotation utility; its advantage over token-matched random selection and Geneformer Native Rank indicates that representation quality depends on which genes are retained as well as how many. Our empirical conclusions concern frozen-backbone embedding extraction and reference mapping; we did not evaluate pretraining, fine-tuning, or task-specific model adaptation. Held-out donors were unseen only with respect to the reference-mapping split, not necessarily model pretraining: corpus audits confirmed PBMC overlap for scGPT and at the scPRINT source-corpus

level, whereas no source-study match for Kang or PBMC was identified in the documented Geneformer V1 corpus. We therefore do not characterize all evaluated settings as pretraining-unseen.

The biological analyses further demonstrate why evaluation should extend beyond aggregate annotation metrics. Support-stratified analyses showed that gains over random compression persisted after excluding the extremely low-support Megakaryocyte population and among more highly supported cell types. AdaGeneBudget also retained lineage-marker programs and stimulation-associated pathway genes at rates far above random selection. In scPRINT, both HVG controls achieved higher annotation macro-F1, whereas AdaGeneBudget more faithfully preserved the stimulation-induced embedding direction under both fixed-panel and expressed-only comparisons. These findings suggest that biological-state fidelity should be considered alongside cell-type annotation when evaluating compressed scFM inputs.

AdaGeneBudget complements rather than replaces architectural and systems-level efficiency improvements. Because it operates before the native input pipeline, requires no backbone updates or cell-type labels, and computes reference-derived statistics once, it can be applied to frozen models without retraining. The same principle may also inform future scFM design: when candidate genes exceed an explicit or practical context budget, identical maximum lengths may not be the most effective use of computation.

Several limitations remain. First, the main evaluation focused on two immune-cell datasets and three pretrained backbones. Supplementary donor-disjoint TNBC validation broadened the tissue and disease settings with a 427,823-cell cohort, but broader validation across tissues, sequencing settings, and pretrained checkpoints is still needed. Evaluating whether similar benefits extend to pretraining or task-specific fine-tuning also remains future work. Second, the expression-specificity weights depend on the reference cohort. Although performance was stable under the tested changes in reference size, donor composition, and under cross-cohort weight estimation, robustness to broader biological or technical reference–query shifts remains to be established. Because inverse-frequency weighting favors low-frequency genes, rare technical artifacts may also receive increased priority. In scGPT–PBMC, low-frequency genes were enriched among selected inputs but remained a small absolute fraction; detailed results are provided in Supplementary Sec. S6. Third, the current scoring rule evaluates genes individually and does not explicitly account for redundancy among selected genes. Under a limited token budget, highly correlated or co-expressed genes may therefore occupy multiple input positions while providing overlapping information, potentially displacing genes with complementary signals. Finally, adaptive scoring and full ranking introduce additional preprocessing overhead relative to simpler fixed-budget policies. Sorting accounted for most of this cost, suggesting that partial-ranking or approximate top-k implementations could reduce it. Because runtime and memory were evaluated on a single GPU platform, the observed computational gains should not be assumed to transfer unchanged across hardware environments.

Overall, AdaGeneBudget shows that cell-adaptive, biologically informed token allocation can provide a favorable balance between computational efficiency and representation fidelity. This establishes input gene selection as a practical and complementary direction for scaling gene-token-based scFMs.

6. Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00561169), Global – Learning & Academic research institution for Master's · PhD students, and Postdocs (G-LAMP) Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. RS-2025-25442252), Institute of Information & communications Technology Planning & Evaluation (IITP) under the Leading Generative AI Human Resources Development (IITP-2027-RS-2026-25546026) grant funded by the Korea government (MSIT), and the NAVER Digital Bio Innovation Research Fund, funded by NAVER Corporation (Grant No. 3720260100).

References

1. 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).
2. 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).
3. 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).
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. J. Kalfon, J. Samaran, G. Peyré and L. Cantini, scPRINT: Pre-training on 50 million cells allows robust gene network predictions, *Nature Communications* **16**, p. 3607 (2025).
6. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems*, (Curran Associates, Inc., 2017).
7. B. Hie, H. Cho, B. DeMeo, B. Bryson and B. Berger, Geometric sketching compactly summarizes the single-cell transcriptomic landscape, *Cell Systems* **8**, 483 (2019).
8. J. Yoon, J. Jordon and M. van der Schaar, INVASE: Instance-wise variable selection using neural networks, in *International Conference on Learning Representations*, 2019.
9. Y. Yamada, O. Lindenbaum, S. Negahban and Y. Kluger, Feature selection using stochastic gates, in *Proceedings of the 37th International Conference on Machine Learning*, eds. H. Daumé, III and A. Singh, Proceedings of Machine Learning Research, Vol. 119 (PMLR, 13–18 Jul 2020).
10. Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou and C.-J. Hsieh, DynamicViT: Efficient vision transformers with dynamic token sparsification, in *Advances in Neural Information Processing Systems*, 2021.
11. H. Cui, C. Wang and K. Pang, scGPT: Towards building a foundation model for single-cell multi-omics using generative AI Zenodo, (2024), Version v1; Software.
12. J. Kalfon, scPRINT: Pre-training on 50 million cells allows robust gene network predictions Zenodo, (2025), Version 1.6.4; Software.

13. T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, III, Y. Hao, M. Stoeckius, P. Smibert and R. Satija, Comprehensive integration of single-cell data, *Cell* **177**, 1888 (2019).
14. S. Liang, V. Mohanty, J. Dou, Q. Miao, Y. Huang, M. Müftüoğlu, L. Ding, W. Peng and K. Chen, Single-cell manifold-preserving feature selection for detecting rare cell populations, *Nature Computational Science* **1**, 374 (2021).
15. A. Missarova, J. Jain, A. Butler, S. Ghazanfar, T. Stuart, M. Brusko, C. Wasserfall, H. Nick, T. Brusko, M. Atkinson, R. Satija and J. C. Marioni, geneBasis: An iterative approach for unsupervised selection of targeted gene panels from scRNA-seq, *Genome Biology* **22**, p. 333 (2021).
16. I. Covert, R. Gala, T. Wang, K. Svoboda, U. Sümbül and S.-I. Lee, Predictive and robust gene selection for spatial transcriptomics, *Nature Communications* **14**, p. 2091 (2023).
17. K. Spärck Jones, A statistical interpretation of term specificity and its application in retrieval, *Journal of Documentation* **28**, 11 (1972).
18. G. Salton and C. Buckley, Term-weighting approaches in automatic text retrieval, *Information Processing & Management* **24**, 513 (1988).
19. H. M. Kang, M. Subramaniam, S. Targ, M. Nguyen, L. Maliskova, E. McCarthy, E. Wan, S. Wong, L. Byrnes, C. M. Lanata, R. E. Gate, S. Mostafavi, A. Marson, N. Zaitlen, L. A. Criswell and C. J. Ye, Multiplexed droplet single-cell RNA-sequencing using natural genetic variation, *Nature Biotechnology* **36**, 89 (2018).
20. Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck, III, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, P. Hoffman, M. Stoeckius, E. Papalexi, E. P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L. M. Fleming, B. Yeung, A. J. Rogers, J. M. McElrath, C. A. Blish, R. Gottardo, P. Smibert and R. Satija, Integrated analysis of multimodal single-cell data, *Cell* **184**, 3573 (2021).
21. C. V. Theodoris, Geneformer: Geneformer-V1-10M pretrained checkpoint `https://huggingface.co/ctheodoris/Geneformer/tree/main/Geneformer-V1-10M`, (2022), Hugging Face model repository, accessed August 4, 2026.
22. Y. Yan, Y. Lin, T. Kumar, S. Bai, A. Thennavan, J. Li, E. Sei, T. Tran, M. Hu, M. Rao, C. Tang, S. He, A. Casasent, E. Ravenberg, G. M. Rauch, A. R. Clayborn, D. Tripathy, A. Thompson, B. Lim, L. Huo, S. Moulder, C. Yam and N. Navin, Ecotypes of triple-negative breast cancer in response to chemotherapy, *Nature* **654**, 1088 (2026).
23. A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov and P. Tamayo, The molecular signatures database (MSigDB) hallmark gene set collection, *Cell Systems* **1**, 417 (2015).
24. Y. Benjamini and Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289 (1995).
25. L. Mazzurana, P. Czarnewski, V. Jonsson, L. Wigge, M. Ringnér, T. C. Williams, A. Ravindran, Å. K. Björklund, J. Säfholm, G. Nilsson, S.-E. Dahlén, A.-C. Orre, M. Al-Ameri, C. Höög, C. Hedin, S. Szczegielniak, S. Almer and J. Mjösberg, Tissue-specific transcriptional imprinting and heterogeneity in human innate lymphoid cells revealed by full-length single-cell rna-sequencing, *Cell Research* **31**, 554 (2021).
26. L. Rebuffet, J. E. Melsen, B. Escalière, D. Basurto-Lozada, A. Bhandoola, N. K. Björkström, Y. T. Bryceson, R. Castriconi, F. Cichocki, M. Colonna, D. M. Davis, A. Diefenbach, Y. Ding, M. Haniffa, A. Horowitz, L. L. Lanier, K.-J. Malmberg, J. S. Miller, L. Moretta, E. Narni-Mancinelli, L. A. J. O'Neill, C. Romagnani, D. G. Ryan, S. Sivori, D. Sun, C. Vagne and E. Vivier, High-dimensional single-cell analysis of human natural killer cell heterogeneity, *Nature Immunology* **25**, 1474 (2024).

Efficient Integration of Bulk and Single-Cell Multi-omics Data Using Latent Factor Models

Qianxing Mo†
Department of Biostatistics and Bioinformatics, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida 33612, USA
Email: qianxing.mo@moffitt.org

Adam B. Olshen
Department of Epidemiology and Biostatistics, Helen Diller Family Comprehensive Cancer Center, University of California at San Francisco, San Francisco, CA 94143, USA

Integrative clustering (iCluster) methods are designed for joint clustering analysis of bulk multi-omics data measured on the same samples and have been successfully applied to identify integrative molecular subtypes for a variety of cancers. The iCluster methods are based on latent factor models that require intensive computation for feature and model selection. The computational cost has limited their application to single cell sequencing studies, which typically generate multi-omics data for tens of thousands of cells in a single experiment. Here, we present a new workflow and a new parameter-tuning algorithm for the iCluster methods that substantially reduces the computational burden, thereby enabling the application of iCluster methods to the integrative analysis of single-cell multi-omics data. We use one bulk multi-omics dataset and two single cell multi-omics datasets to demonstrate the updated iCluster methods. Our results show that the updated iCluster methods compare favorably to the Weighted Nearest Neighbor (WNN) method and the Multi-Omics Factor Analysis (MOFA2) method for integrative analysis of single cell multi-omics data. The new algorithm has been implemented in R functions iClusterPlus2 and tune.iCluster2b, which are available at https://bioconductor.org/packages/release/bioc/html/iClusterPlus.html.

Keywords: Integrative clustering; iCluster; Latent factor; Multi-omics; iClusterPlus; iClusterPlus2.

1. Introduction

Over the past 30 years, high-throughput technologies, such as microarrays and next-generation sequencing, have enabled the generation of massive amounts of molecular data, including copy number, gene expression, DNA methylation, somatic mutations, and other molecular features of cancer. The promise of these technologies has led to large-scale collaborative initiatives, such as The Cancer Genome Atlas (TCGA)[1, 2] and the International Cancer Genome Consortium (ICGC)[3]. In these projects, as well as in similar but smaller-scale studies conducted by individual laboratories, the same biological samples were often subjected to multiple molecular assays.

† To whom correspondence should be addressed.

© 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.

In the 2010s, single-cell sequencing emerged as a prominent technology. While initially focused primarily on gene expression, single-cell sequencing has been expanded to enable the measurement of many other molecular features, including DNA mutations, DNA methylation, and DNA–protein interactions. Common examples of multimodal single-cell experiments include cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq)[4], RNA sequencing combined with assay for transposase-accessible chromatin using sequencing (ATAC-seq)[5], and single-cell nucleosome, transcriptome and methylation sequencing (scNMT-seq)[6].

Although each type of multi-omics data could be analyzed separately, integrative analysis is biologically and scientifically appealing because different molecular measurements can provide complementary information about the underlying biology of cancer and other diseases. However, integrative analysis is challenging because multi-omics data are typically high-dimensional and exhibit different noise characteristics, scales and distributions. The iCluster method[7, 8] was one of the first approaches designed for integrative analysis of continuous multi-omics data. It was subsequently extended to iClusterPlus[9], which can jointly model multiple types of omics data, including continuous, binary, multi-categories and count data. In addition, a Bayesian extension of iClusterPlus, iClusterBayes[10], was developed to incorporate Bayesian variable selection[11]. To transform high-dimensional omics data into a low-dimensional space while retaining the most important patterns, the iCluster methods use low dimensional latent factors to model the high-dimensional multi-omics data. In addition, they employ lasso[12] or Bayesian variable selection approaches to identify informative features. Other integrative methods, including Cobolt[13] and MOFA[14] also use latent factor models for dimensionality reduction. A major distinction among these methods is the approach to estimate the latent factors. For example, the iCluster methods use expectation maximization (EM) algorithm and classical Bayesian methods, while Cobolt uses multimodal variational autoencoder (MVAE) and MOFA2 uses a variational Bayesian approach.

Although the iCluster methods have been successfully applied to bulk multi-omics data to identify integrative cancer subtypes across a variety of cancers, they have not been applied to analysis of single cell sequencing experiments, which typically generate multi-omics data for tens of thousands of cells. We realize that a major limitation of the iCluster methods is their intensive computational requirements for feature and model selection, which has limited their applications to large-scale multi-omics datasets. Therefore, in this manuscript, we introduce a new workflow and a novel parameter-tuning algorithm designed to substantially reduce the computational burden of the iCluster methods, thereby enabling their application to large single-cell multi-omics datasets.

The manuscript is organized as follows. In the Methods section, we briefly introduce the iCluster methods, the new workflow and the lasso parameter-tuning algorithm. In the Application section, we apply the new methods to one bulk multi-omics dataset and two single cell multi-omics datasets, and compare their performance with that of the popular Seurat WNN[15, 16] method and MOFA2. Finally, in the Discussion, we summarize and discuss the proposed methods and their potential applications.

2. Methods

2.1. *A Latent factor model for joint dimension reduction of continuous multi-omics data*

Suppose there are n samples assayed by T high-throughput techniques generating T omics datasets. These multi-omics data can be represented as data matrices for integrative analyses. Let X_t be the tth data matrix with dimensions $p_t \times n$, where p_t is the number of features (rows) and n is the number of samples (columns). Shen et al. [7, 8] proposed a Gaussian latent factor model for joint decomposition of X_t ($t = 1, 2, \ldots, \mathrm{T}$), which can be expressed as

$$X_t = W_t Z + E_t, \quad t = 1, 2, \ldots, \mathrm{T}. \tag{1}$$

In model (1), W_t denotes the weight (loading) matrix with dimensions $p_t \times k$, corresponding to p_t features and k latent factors; Z denotes the latent factor matrix with dimensions $k \times n$, corresponding to k latent factors and n samples; and E_t denotes the random error matrix with dimensions $p_t \times n$, where the column vector $e_{ti} (i = 1, \ldots, n)$ represent independent random errors that follow a multivariate normal distribution with mean 0 and a diagonal covariance matrix $\Psi_t = (\sigma_1^2, \ldots, \sigma_{p_t}^2)$. The model (1) can be viewed as a high-rank matrix X_t decomposed into two low-rank matrices W_t and Z ($k \ll p_t, n$) plus a random error matrix E_t. Sample i ($i = 1, 2, \ldots, \mathrm{n}$) is associated with the ith column vector $z_i = (z_{i1}, z_{i2}, \ldots, z_{ik})'$ of matrix Z that captures the correlation structure of multi-omics data. As a result, the latent factor vectors can be used for sample clustering. The jth row vector $w_{tj} = (w_{tj1}, \ldots, w_{tjk})$ of W_t indicate how strongly feature j is associated with each factor. For feature j in the tth data matrix, the larger the absolute value of $|w_{tjk}|$ the stronger the association with factor k. Therefore, the weight matrix can be used to identify the important multi-omics features that contribute to the joint sample clustering. Shen, Wang and Mo[8] used a penalized EM algorithm to estimate the optimal matrices W_t and Z in which a lasso-type penalty term is imposed on W_t to introduce sparsity. When an appropriate lasso parameter is chosen, some rows of W_t become all zeros, indicating their corresponding features are non-informative, while the informative features are not all zeros. Therefore, the latent factor model (1) can be used to perform both sample clustering and feature selection, which are both important for integrative clustering analysis of multi-omics data.

2.2. *Latent factor models for joint dimension reduction of heterogeneous multi-omics data*

The original iCluster method is designed to integrate continuous multi-omics data and assumes that each data type follows a normal distribution. Although this assumption is generally appropriate for some normalized molecular measurements, it is not suitable for many commonly encountered data types. In practice, multi-omics datasets are heterogeneous, which include continuous, count, binary, and categorical variables. For example, gene expression in RNA-seq is usually measured as read counts, DNA copy number alterations are often classified as gain, loss, or normal, and somatic mutation is typically represented by a binary variable denoting the presence or absence of a mutation. To accommodate this heterogeneity, Mo et al. developed the iClusterPlus[9] method that extends the original latent factor framework by modeling each data type using an appropriate member of the generalized linear model (GLM) family. This capability represents a substantial advancement over the original iCluster method by broadening its

applicability to diverse multi-omics datasets. In the following, we briefly describe the models implemented in iClusterPlus.

In the iClusterPlus model, x_{ijt} denote a variable for the jth ($j = 1, \cdots, p_t$) omics feature of the ith ($i = 1, \cdots, n$) sample in the tth ($t = 1, \cdots, T$) dataset. Assume sample i is associated with a latent factor $z_i = \left(z_{i1,}\ z_{i2,} \ldots,\ z_{ik}\right)'$ with k elements. When x_{ijt} is a continuous variable, it is assumed to follow a Gaussian distribution, with its mean modeled as a linear function of z_i:

$$x_{ijt} = \alpha_{jt} + \beta_{jt} z_i + \varepsilon_{ijt}, \varepsilon_{ijt} \sim N(0,\ \sigma_{jt}^2), \quad (2)$$

where α_{jt} is the intercept, $\beta_{jt} = (\beta_{1jt}, \ldots, \beta_{kjt})$ are the slope coefficients, and ε_{ijt} is the random error assumed to be independent, identically and normally distributed with mean 0 and variance σ_{jt}^2. When $x_{ijt}$is a count variable, it is assumed to follow a Poisson distribution, with its mean modeled as a function of z_i through a Poisson regression with a log link function:

$$\log(\lambda(x_{ijt}|\ z_i)) = \alpha_{jt} + \beta_{jt} z_i, \quad (3)$$

where $\lambda(x_{ijt}|\ z_i)$ is the conditional mean of the count given z_i. When $x_{ijt}$is a binary variable with realized value 0 or 1, it is assumed to follow a Bernoulli distribution, with the probability of being 1 modeled as a function of z_i through a logistic regression model:

$$\log\left(\frac{P(x_{ijt}=1|z_i)}{1-P(x_{ijt}=1|z_i)}\right) = \alpha_{jt} + \beta_{jt} z_i, \quad (4)$$

where $P(x_{ijt} = 1|z_i)$ is the probability of $x_{ijt} = 1$ given z_i. Finally, when x_{ijt} is a multi-categorical variable, it is assumed to follow a multinomial distribution, with category probabilities modeled as functions of z_i through a multinomial logistic regression model:

$$P(x_{ijt}=\mathrm{c}|z_i) = \frac{\exp(\alpha_{jct}+\beta_{jct} z_i)}{\sum_{l=1}^{C}\exp(\alpha_{jlt}+\beta_{jlt} z_i)},\ \mathrm{c} = 1, 2, \ldots, \mathrm{C}, \quad (5)$$

where $P(x_{ijt}=\mathrm{c}|z_i)$ denotes the probability of x_{ijt} being the *c* category.

In the iClusterPlus model, all the omics data are linked through a common latent factor representing the underlying biological structure shared across multi-omics platforms. Parameter estimation is performed like in iCluster using a penalized EM algorithm with sparsity-inducing penalties to identify informative features while simultaneously estimating the latent factors and model parameters.

2.3. *A dynamic algorithm for searching optimal lasso parameters in T-dimensional space*

In the joint modeling of multi-omics data, iCluster imposes a lasso penalty on parameter estimation to facilitate identifying the omics features that contribute to sample clustering. When an appropriate lasso parameter is chosen, the coefficients β_{jt} corresponding to non-informative

features are shrunk to zero, whereas those associated with informative features retain non-zero values. Sample clusters are determined based on the estimated latent factors z_i. It is a computationally intensive process to search for optimal lasso parameters in a T-dimensional space, especially when $T \geq 2$. In order to efficiently search for the optimal combination of penalty parameters, iClusterPlus uses the uniform design to generate good lattice points from the search domain. For example, the recommended lattice points in a 4-dimensional space are 307, 562, 701, 1019, 2129, 3001, 4001, 5003 and 6007. Usually, the more lattice points that are searched, the more likely an optimal model is identified. Originally, iCluster was developed to identify cancer subtypes. In practice, that is typically a small number of subtypes. Therefore, it is only necessary to test a small number of latent factors (e.g., $k = 1 - 6$). Suppose there were four omics datasets and we chose $k = 1 - 6$, we would need to evaluate at least 307×6 models. The optimal *k* is selected at the elbow point of the plot of the deviance ratio or the Bayesian information criterion (BIC) versus *k*, that is where the deviance ratio or BIC levels off. Using this empirical rule, iCluster divides n samples into $k + 1$ clusters.

Although iCluster models work well for joint clustering analysis of bulk cancer multi-omics datasets, they face challenges in the analysis of larger single cell multi-omics data, which can consist of dozens of cell types among thousands of cells. To reduce computational burden, we propose the following workflow and parameter-tuning algorithm to optimize the iCluster methods.

(1) Determine an optimal number of factors by performing principal component analysis (PCA) of concatenated multi-omics data. An optimal number of factors can be identified from elbow of the PCA variance plot.
(2) For each of the multi-omics dataset, set the search domain for the lasso parameter and calculate the change of lasso parameter for each the iteration of the EM algorithm. Let λ_{t0} and λ_{t1} denote the minimum and maximum lasso parameters of dataset t, respectively. The change of lasso parameter ($\lambda_{t\Delta}$) for each iteration of the EM algorithm is calculated as
$$\lambda_{t\Delta} = (\lambda_{t1} - \lambda_{t0})/N,$$
where N is the maximum number of EM iterations.
(3) For dataset t, increase or decrease lasso parameter λ_t based on the BIC change. The following is the pseudo code for updating λ_t at each EM iteration.

```
Initialize BIC_min_t = BIC of the initial model for dataset t
Initialize BIC_rate_cutoff_t = 0.05
for i = 1 to N do
  for t = 1 to T do
     BIC_current_t = BIC of current model for dataset t
     BIC_rate_t = (BIC_current_t - BIC_min_t)/BIC_min
     if BIC_rate_t > BIC_rate_cutoff_t then
       λ_t = λ_t + λ_tΔ
     else if BIC_rate_t < −BIC_rate_t then
       BIC_min_t = BIC_current_t
       λ_t = λ_t − λ_tΔ
  end for
end for
```

(4) After the iterations, perform uniform manifold approximation and projection (UMAP) on the latent factor matrix to obtain a two-dimensional embedding and visualize the sample/cell clusters in the resulting UMAP space.
(5) Perform clustering analysis for the samples or cells using partition-based methods (e.g. k-means) or graph-based methods (e.g., Louvain[18] method), depending on the clustering structure observed in the UMAP embedding.

In the above algorithm, we set a cutoff for the BIC rate of change between 2 consecutive iterations. If the rate of change is less than the cutoff (e.g., 0.05), λ_t will remain the same. Suppose the number of iterations is N. The possible lattice points in a T-dimensional space will be N^T, which is quickly prohibitive with the old algorithm. We use three real examples to illustrate the new workflow.

2.4. *Comparisons with other methods*

We ran WNN using default parameters. For iCluster and MOFA2, we chose the same number of latent factors. We ran MOFA2 using the same overall workflow as iCluster except in the step used to estimate the latent factors.

3. Application

3.1. Integrative clustering analysis of uveal melanoma (UM) multi-omics data.

We first applied the new algorithm to the bulk uveal melanoma (UM) multi-omics data generated by TCGA[2]. The UM dataset consists of somatic mutation, DNA copy number variation, DNA methylation and RNA-seq gene expression data for 80 UM primary samples. The level-3 UM multi-omics data available at http://firebrowse.org were processed to form 4 data matrices in which the columns represent the samples, and the rows represent the omics features. The data procession was detailed in Mo et al.[17]. Briefly, somatic mutation was coded as 1 and wild-type was coded as 0. Genes with a somatic mutation rate $\geq 2\%$ in the samples were used for iCluster analysis. We logit transformed the methylation beta values and then selected the top 25% most variable genes. We log2 transformed the mRNA-seq normalized counts plus 1 and selected the top 25% most variable genes. For copy number data, we condensed the chromosome segments to 2632 chromosome regions using the CNregions function in the iClusterPlus package. Finally, we standardized omics feature values across samples for the methylation, mRNA and copy number data.

To determine the number of factors used for the latent factor model, we concatenated the standardized methylation, mRNA and copy number data matrix and then performed PCA on the concatenated matrix. Figure 1 shows the proportion of the total variance explained by the first 30 principal components (PCs). The explained variance decreases rapidly for the first few PCs, after which the rate of decline slows markedly beginning around the fifth PC. The optimal number of latent factors for iCluster modeling can be selected near this inflection point. We chose $k = 7$ as the number of latent factors for iCluster integration. The resulting sample clustering was highly consistent for values of $k \geq 7$, indicating a stable solution.

It took about a minute to run the new algorithm on a Mac Studio with a M4 Max chip, which was implemented in a new function named iClusterPlus2. In contrast, the first version of iClusterPlus took about 12 hours to evaluate 701 combinations of lasso parameters for each of the latent factors ($k = 1 - 5$). Both versions of iClusterPlus generated the same clustering results.

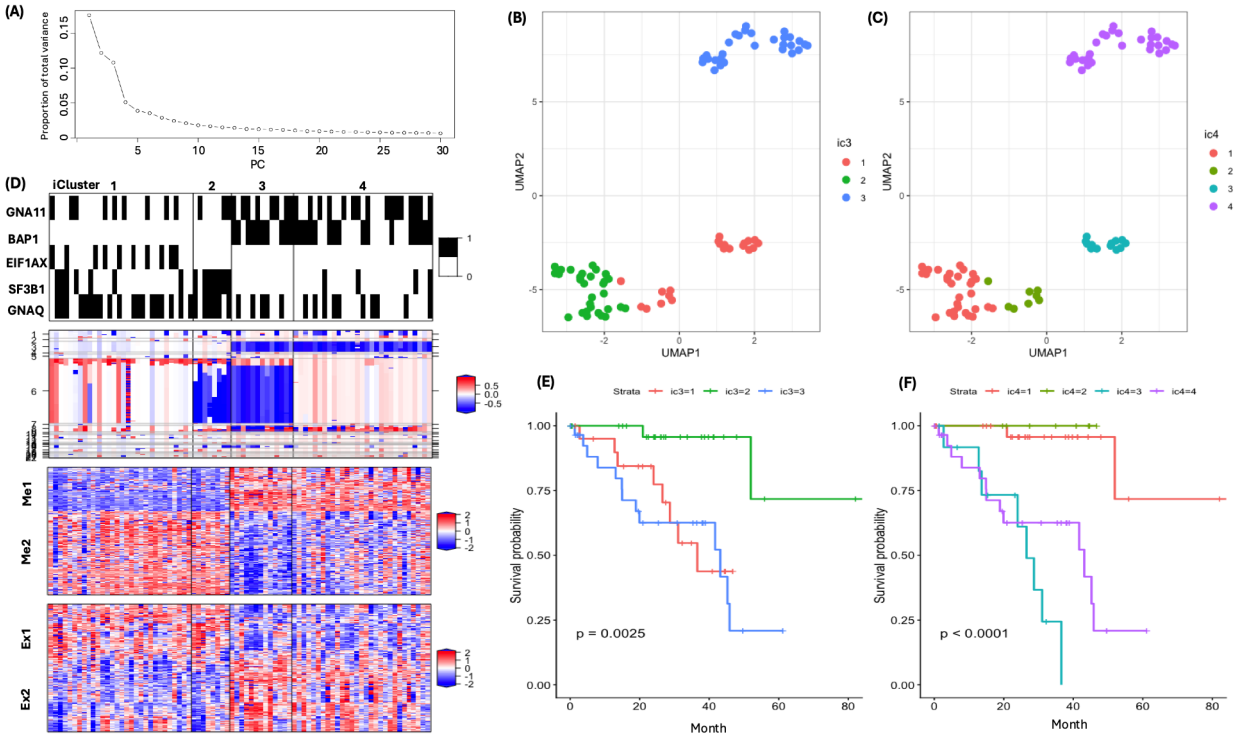


Figure 1. iCluster analysis of UM multi-omics data. (A) PCA variance plot: proportion of total variance vs. PC. (B) UM samples distributed on UMAP 2-dimensional space; UM samples are grouped into 3 clusters by K-means clustering. (C) UM samples distributed on UMAP 2-dimensional space; UM samples are grouped into 4 clusters by K-means clustering. (D) Heatmap of the top driver features of the multi-omics data. Somatic mutation (the top panel) is indicated by black bar; DNA copy number (the 2nd panel from the top): copy number loss, normal and gain are indicated by blue, white and red, respectively; Methylation (the 3rd panel from the top): hypomethylation and hypermethylation are indicated by blue and red, respectively. mRNA (the bottom panel): low and high expression are indicated by blue and red, respectively. (E) Overall survival of the 3 iClusters. (F) Overall survival of the 4 iClusters.

To visualize the sample clusters in the latent factor space, we performed UMAP on the latent factor matrix. Figure 1B and 1C show the 3 and 4 major clusters on the UMAP 2-dimensional space, respectively. It can be seen that cluster 1 of Figure 1B is further divided into 2 sub-clusters (clusters 2 and 3) of Figure 1C. The iCluster analysis identified the somatic mutations *GNAQ*, *SF3B1*, *EIF1AX*, *BAP1* and *GNA11* were drivers of UM subtyping (Figure 1D, top panel). The driver copy number alterations included chr3 loss, chr6 gain and loss, and chr8 gain (Figure 1D, the 2nd panel from the top). The driver methylated genes formed 2 major gene clusters labeled as Me1 and Me2, which had opposite methylation patterns from each other (Figure 1D, the 3rd panel from the top). In terms of gene expression, the driver genes also formed 2 major gene clusters labeled as Ex1 and Ex2 that had opposite expression patterns (Figure 1D, the bottom panel). Overall, it appears that the multi-omics patterns drive the samples clustering into two major clusters 1+2 and 3+4, which coincides with the chr3 status (chr3 normal in clusters 1+2, chr3 loss

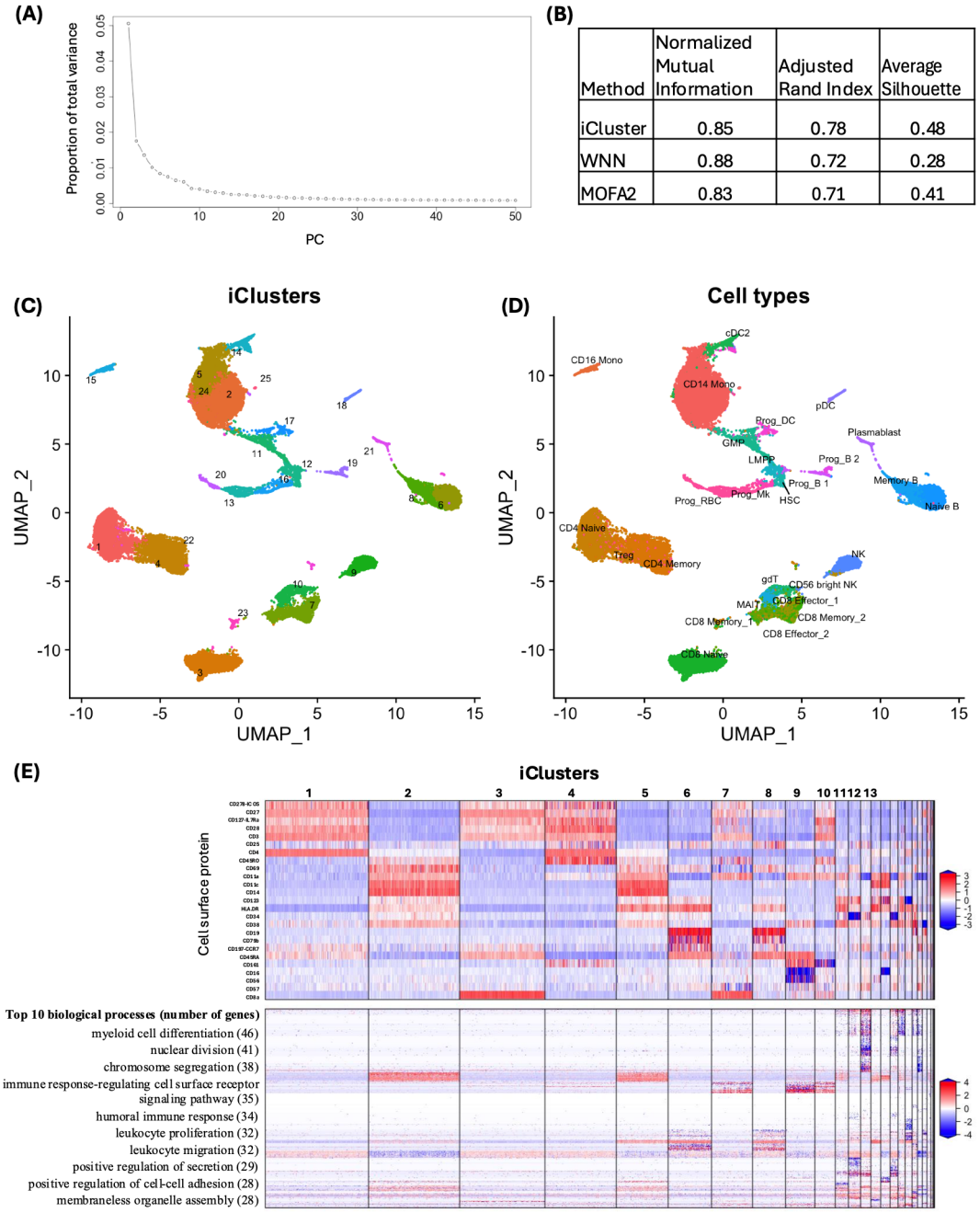
in clusters 3+4). The loss in chr6q arm is the major visible omics pattern that divides clusters 1+2 and 3+4 further into two smaller clusters, where clusters 2 and 3 are associated with chr6q loss, while clusters 1 and 4 are not. It is well known that the most common chromosomal aberration in this disease is the loss of one copy of chr3 (monosomy 3). Patients with monosomy 3 are associated with a worse overall survival than those with normal chr3, while there is no significant difference in overall survival between clusters 1 and 2, nor between clusters 3 and 4 (Figure 1F). In the 3-cluster solution, both clusters 2 and 3 include patients with chr3 loss, which may explain why there is not significant survival difference between the two clusters (Figure 1E). Statistically, it is challenging to determine which clustering result (the 3-cluster solution or the 4-cluster solution) is better. Clinically, it is clear that the 4-cluster solution is better because it is more consistent with the clinical observation.

3.2. Integrative clustering analysis of CITE-seq data.

Our second application is to single cell multi-omics data collected using the CITE-seq method[9]. The dataset consists of single cell RNA-seq (scRNA-seq) and antibody-derived tag (ADT) data from the same 30,672 cells from human bone marrow cells first introduced by Stoecklus et al.[9] with the experiment performed by Stuart et al.[16]. The data can be obtained by installing the SeuratData R package (https://github.com/satijalab/seurat-data). The cells include 17009 RNA features and 25 ADT features. We processed the CITE-Seq data following the Seurat Weighted Neighbor Analysis (WNN) vignette (https://satijalab.org/seurat/articles/weighted_nearest_neighbor_analysis), which included normalization, scaling and feature selection. As a result, the 2000 most highly variable RNA-seq genes and all the 25 ADT features were used for iCluster and WNN analyses.

To estimate the ideal number of factors to use for iCluster modeling, we performed PCA of the concatenated matrix of ADT and RNA-seq. We found that the explained variances level off around the 10th PC (Figure 2A). For iCluster modeling of single cell multi-omics data, we suggest selecting the PC after which the explained variance levels off. As an example, we selected $k = 20$ as the number of latent factors for iCluster modeling. Actually, we found that the clustering results were generally consistent for a range of factors (e.g., $k = 20 - 30$). Because features had already been selected, we utilized a small value for the lasso parameters (λ=0.1) for both the RNA and ADT data. The model ran in only 17 seconds on the M4 Mac Studio mentioned earlier.

We utilized the Louvain[18] method (a graph-based clustering method implemented in the RphenoGraph function in R) to identify cell clusters in the latent factor space. As a result, we identified 25 clusters ranging in size from 4736 to 29 cells. These cell clusters are visualized on the 2-dimensional UMAP of the latent factors (Figure 2C). The annotated cell clusters are shown on Figure 3D. The most abundant cells are CD4 naïve, CD14 mono, CD8 naïve, CD4 memory, CD14 mono and naïve B cells (Figure 3D). The ADT and RNA expression patterns are shown on Figure 2E. We can see that the iClusters 1 (CD4 naïve), 3 (CD8 naïve) and 4 (CD4 memory) are characterized by high expression of CD4, CD8a, CD4 and CD45RO, respectively (Figure 2E, top panel). In addition, the three iClusters are also characterized by high expression of CD278-ICOS, CD27, CD127-IL7Ra, CD28 and CD3 (Figure 2E, top panel).



Method	Normalized Mutual Information	Adjusted Rand Index	Average Silhouette
iCluster	0.85	0.78	0.48
WNN	0.88	0.72	0.28
MOFA2	0.83	0.71	0.41

Figure 2. iCluster analysis of CITE-seq data. (A) The PCA variance plot of the CITE-seq data. (B) Comparison of iCluster, WNN and MOFA2 in terms of normalized mutual information, adjusted rand index and average silhouette. (B) Cell clusters identified by iCluster. (D) Annotated cell clusters. (E) Heatmaps of cell surface markers of the ADT data (top panel) and the top 500 genes of the scRNA-seq data (bottom panel).

To identify the most important genes that drove the cell clustering, we ranked the RNA-seq genes by the sum of absolute values of the model parameters (βs) from the largest to the smallest.
Figure 2E shows the top 500 genes of the RNA-seq data. Immune related biological processes (gene ontology terms) are enriched in the top 500 RNA-seq genes, which include myeloid cell differentiation (46 genes), immune response-regulating cell surface receptor signaling pathway (35

genes), humoral immune response (34 genes), leukocyte proliferation (32 genes) and migration (32 genes) (Figure 2E, bottom panel).

To evaluate the performance of iCluster, WNN and MOFA2, we calculated three performance metrics for which higher values indicate better performance: normalized mutual information (NMI), adjusted rand index (ARI) and average silhouette (AS). These three metrics were calculated by comparing the cell clusters identified by each method with the annotated cell types. iCluster achieved the highest ARI and AS values, and the second-highest NMI value (Figure 2B). Overall, iCluster demonstrated the best performance.

3.3. Integrative clustering analysis of scRNA-seq and scATAC-seq data.

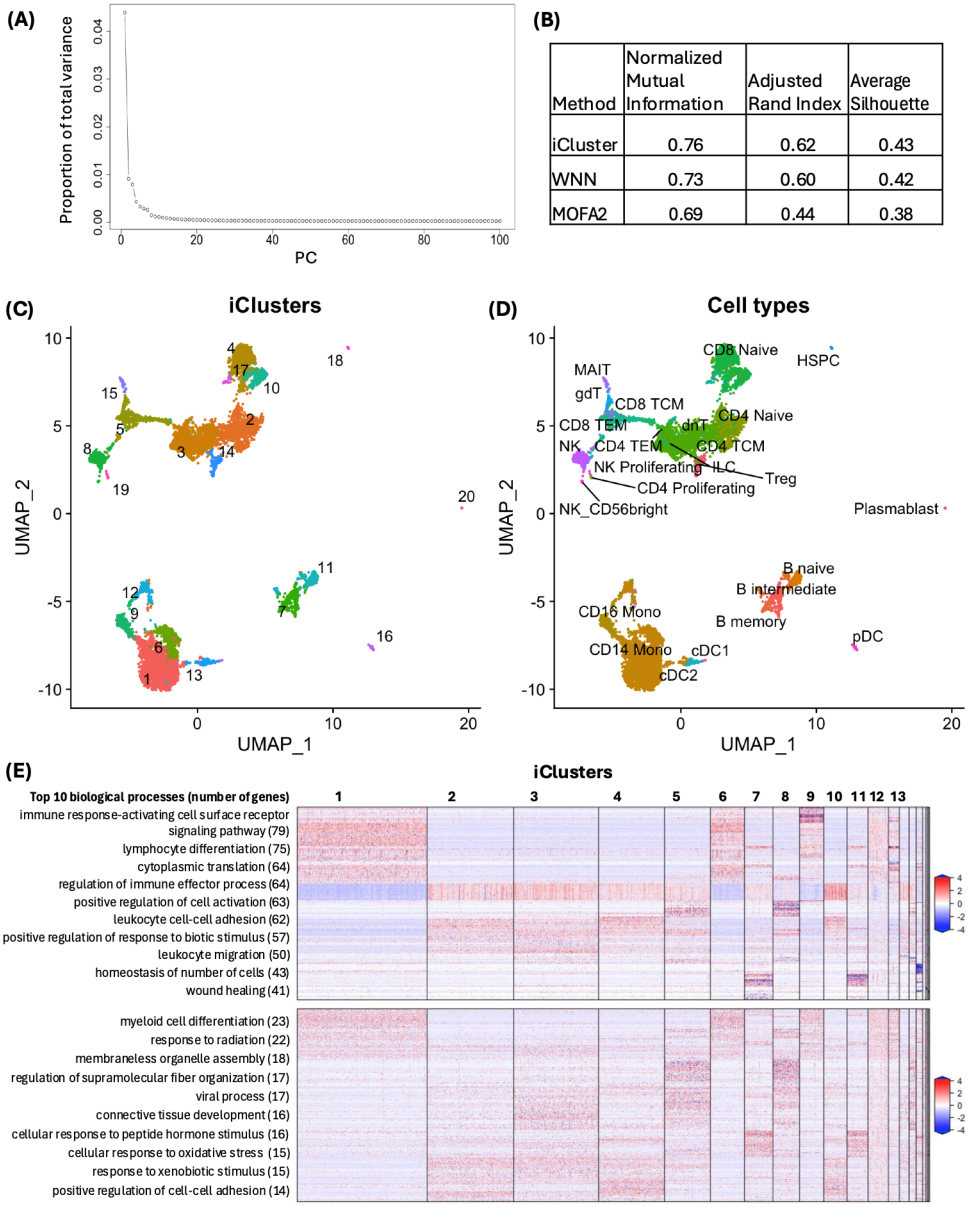
Our third example is another common type of single cell experiment where scRNA-seq and scATAC-seq[5] are applied to the same set of cells, in this case 11,909 PBMCs, and can be found at the 10x Genomics website. ATAC-seq is a technique that is used to measure chromosome accessibility. ATAC-seq data are correlated within regions, with high regional counts implying open chromatin regions. Integrating scRNA-seq and scATAC-seq should allow better identification of cell types.

We used the Seurat package to process the scRNA-seq and ATAC-seq data, which included filtering out cells with low quality data, normalization and scaling of the sequence counts and feature selection. Specifically, we filtered out cells with percent of RNA mapped to mitochondrial genes $\geq 20\%$, total ATAC-seq counts $\leq 5{,}000$ or $\geq 70{,}000$, and total RNA-seq counts $\leq 1{,}000$ or $\geq 25{,}000$. As a result, 10,412 cells remained for iCluster analysis. We selected the 3000 most variable genes for the RNA-seq data and top 20% most variable regions for the ATAC-seq data. In addition, we merged neighboring ATAC regions into the same regions, which led to 15,441 regions for iCluster analysis.

To determine the number of factors used for iCluster modeling, we performed PCA of the concatenated matrix of RNA-seq and ATAC-seq. We found that the variances level off around the 10th PC (Figure 3A). However, since the first PC was dominated by sequencing depth, Seurat only used the 2nd to the 50th PCs for the WNN integrative analysis (see Seurat WNN tutorial). In order to have a fair comparison, we chose $k = 50$ latent factors for iCluster modeling. Then, we identified the cell clusters in the latent factor space using the Louvain method and visualized the cell clusters on a 2-dimensional UMAP of the latent factors. We annotated the cells using Azimuth software (https://azimuth.hubmapconsortium.org/) with PBMC as the reference.

We identified 20 cell clusters with sizes ranging from 2138 to 18 cells (Figure 3C). The most abundant cell types are CD14 Mono, CD4 Naïve, CD4 TCM, CD8 Naïve and CD8 TEM (Figure 3D). HSPC, Plasmablast and pDC are among the least abundant cells and they are well separated from the other mature immune cells on the UMAP plot, which are expected (Figure 3D). To identify the most important features that drove the cell clustering, we ranked the RNA-seq genes and ATAC-seq regions by the sum of absolute values of the model parameters (βs) from the largest to the smallest. Figure 3E shows the top 750 genes of the RNA-seq data and 1544 regions of the ATAC-seq data. The top biological processes (gene ontology terms) involved in the top 750 RNA-seq genes are immune related, which include the immune response-activating cell

surface receptor signaling pathway (79 genes), lymphocyte differentiation (75 genes), regulation of immune effector process (64 genes), positive regulation of cell activation (63 genes), leukocyte cell-cell adhesion (62 genes) and positive regulation of response to biotic stimulus (57 genes). A lot of genes falling in the top ATAC-seq regions are also involved in immune-related processes, which include myeloid cell differentiation (23 genes), response to radiation (22 genes) and viral process (17 genes); the others are involved in cellular signal pathways such as cellular response to peptide hormone stimulus, cellular response to oxidative stress (15 genes) and positive regulation of cell-cell adhesion (14 genes).



Method	Normalized Mutual Information	Adjusted Rand Index	Average Silhouette
iCluster	0.76	0.62	0.43
WNN	0.73	0.60	0.42
MOFA2	0.69	0.44	0.38

Figure 3. iCluster of RNA-seq and ATAC-seq data. (A) The PCA variance plot of the combined RNA-seq and ATAC-seq matrices. (B) Comparison of iCluster, WNN and MOFA2 in terms of normalized mutual information, adjusted rand index and average silhouette. (C) Cell clusters identified by iCluster. (D) Annotated cell clusters. (E) Heatmap of the top features in the RNA-seq data (top panel) and the ATAC-seq data (bottom panel).

Using the cell types predicted by Azimuth as the reference, we calculated the NMI, ARI and AS for the cell clusters identified by iCluster, WNN and MOFA2. Compared to WNN and MOFA2, iCluster achieved the highest values for all the three metrics, indicating the best overall performance (Figure 3B).

Discussion

In this manuscript, we introduce a new workflow and parameter-tuning algorithm for the iCluster methods to integrate multi-omics generated from the same samples or cells. In the new approach, we first determine the optimal number of latent factors for iCluster modeling using principal component analysis of the concatenated data matrix. Then, we use a dynamic algorithm to search for optimal lasso penalty parameters during iCluster modeling. Next, we visualize sample or cell clusters on a two-dimensional UMAP embedding of the latent factor matrix. Finally, we determine cluster memberships utilizing either partition-based or graph-based clustering of the latent factors depending on the clustering structure observed in the UMAP embedding.

This new approach overcomes a major limitation of the previous iCluster workflow, which requires an extensive grid search over different lasso penalty parameters and numbers of latent factors. As a result, the new approach substantially reduces the computational burden of the iCluster methods while maintaining their ability to integrate multi-omics data. Using TCGA uveal melanoma and two single-cell multi-omics datasets, we demonstrate that the new approach can be used not only to identify cancer molecular subtypes but also to identify cell clusters in single-cell multi-omics datasets comprising tens of thousands of cells. These results demonstrate the broad applicability and computational efficiency of the proposed approach for both bulk and single-cell multi-omics data analyses.

We compared the updated iCluster methods to the popular Seurat WNN method and the MOFA2 method. Overall, the updated iCluster methods demonstrated overall superior performance compared with the WNN and MOFA2 methods across three evaluation metrics: NMI, ARI, and AS score. In addition, because the iCluster methods are model-based, the model parameters can be used to rank multi-omics features to identify the most important features that drive the cell clustering. This feature-selection capability is an important advantage of model-based methods. In contrast, the WNN method does not provide a direct approach for identifying and ranking multi-omics features that contribute most to cell clustering in single-cell multi-omics data analysis.

One limitation of our study is that we compared the iCluster methods only with the WNN and MOFA2 methods. Several other methods have been developed for integrative analysis of multi-omics data and could also be considered for comparison. In future work, we plan to conduct more extensive comparisons with additional methods.

In summary, we developed a new computationally efficient approach for the iCluster methods that substantially improves their applicability to both bulk and single-cell multi-omics data. We believe that this approach will be a useful computational tool for multi-omics data analysis in biological research.

References

1.Cancer Genome Atlas Research N. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489(7417):519–25.
2.Robertson AG, Shih J, Yau C, Gibb EA, Oba J, Mungall KL, et al. Integrative Analysis Identifies Four Molecular and Clinical Subsets in Uveal Melanoma. Cancer Cell. 2018;33(1):151.
3.International Cancer Genome C, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, et al. International network of cancer genome projects. Nature. 2010;464(7291):993–8.
4.Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, et al. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods. 2017;14(9):865–8.
5.Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science. 2018;361(6409):1380–5.
6.Clark SJ, Argelaguet R, Kapourani CA, Stubbs TM, Lee HJ, Alda-Catalinas C, et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nat Commun. 2018;9(1):781.
7.Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;25(22):2906–12.
8.Shen R, Wang S, Mo Q. Sparse Integrative Clustering of Multiple Omics Data Sets. Ann Appl Stat. 2013;7(1):269–94.
9.Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc Natl Acad Sci U S A. 2013;110(11):4245–50.
10.Mo Q, Shen R, Guo C, Vannucci M, Chan KS, Hilsenbeck SG. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostatistics. 2018;19(1):71–86.
11.Savitsky T, Vannucci M, Sha N. Variable Selection for Nonparametric Gaussian Process Priors: Models and Computational Strategies. Stat Sci. 2011;26(1):130–49.
12.Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology 1996;58(1):267–88.
13.Gong B, Zhou Y, Purdom E. Cobolt: integrative analysis of multimodal single-cell sequencing data. Genome Biol. 2021;22(1):351.
14.Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, Stegle O. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 2020; 21(1):111.
15.Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177(7):1888–902 e21.
16.Hao Y, Hao S, Andersen-Nissen E, Mauck WM, 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573–87 e29.
17.Mo Q, Wan L, Schell MJ, Jim H, Tworoger SS, Peng G. Integrative Analysis Identifies Multi-Omics Signatures That Drive Molecular Classification of Uveal Melanoma. Cancers (Basel). 2021;13(24).
18.Levine JH, Simonds EF, Bendall SC, Davis KL, Amir el AD, Tadmor MD, et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell. 2015;162(1):184–97.

scMultiNODE: Recovering Developmental Dynamics in Weakly Resolved Modalities through Multi-Modal Integration

Jiaqi Zhang

Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21202, USA

Manav Chakravarthy

Department of Computer Science, Brown University, Providence, RI 02906, USA

Ritambhara Singh†

Siebel School of Computing and Data Science, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA
Center for Computational Molecular Biology, Brown University, Providence, RI 02906, USA
Department of Computer Science, Brown University, Providence, RI 02906, USA
† *Corresponding author: ritsingh@illinois.edu*

Measuring multiple genomic profiles (i.e., modalities) over time enhances our understanding of cell development. However, these modalities are not equally analyzable. Sparse and poorly separated measurements, such as chromatin accessibility, are difficult to interpret on their own, and the limited availability of temporal single-cell co-assay data leaves such modalities without the cross-modal cell correspondence needed to borrow structure from richer ones. As a result, critical developmental analyses and cross-modal comparisons remain underperformed for these weak modalities. Additionally, existing methods cannot integrate separately sequenced measurements while preserving cellular dynamics. We present scMultiNODE, an integration model that combines gene expression and chromatin accessibility measurements from multiple discrete timepoints for single cells, without relying on any prior cell-to-cell correspondence across modalities or time. scMultiNODE employs a scalable Quantized Gromov-Wasserstein optimal transport method to align cells across measurements, and neural ordinary differential equations with a dynamic regularization term to model cell development in a shared latent space, capturing cellular dynamics while preserving cell-type variations. Experiments across four developmental single-cell datasets demonstrate that scMultiNODE integrates separately sequenced measurements over time more effectively than existing methods that overlook cellular dynamics. Crucially, by anchoring chromatin accessibility to the shared dynamical latent space, scMultiNODE recovers developmental trajectories and cell annotations from scATAC-seq that this modality cannot support in isolation. The resulting joint latent space further supports important multi-modal downstream analyses, including complex trajectory investigation and cross-modal label transfer for cell annotation. The data, code, and supplementary notes are publicly available at `https://github.com/rsinghlab/scMultiNODE`.

Keywords: single-cell development, multi-modal data integration, optimal transport

© 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

1. Introduction

Biological systems are inherently dynamic, constantly changing and adapting over time, and function at different scales, like organisms, cells, and molecules. Analyzing dynamic processes at the cell level reveals how cells grow, divide, and differentiate into various cell types.[1,2] Understanding these cellular dynamics enhances our understanding of cell development and disease progression. Advances in single-cell measurements taken over time support this study by capturing high-resolution genomic snapshots of different cell states. Recent research increasingly involves temporally profiling multiple developmental single-cell measurements (or modalities), uncovering heterogeneity within the same tissue type and cellular dynamics across developmental stages.[3,4] For instance, Fleck et al.[5] collected gene expression and chromatin accessibility profiles across a developmental time course of brain organoids, capturing key processes such as neurogenesis. These modalities, however, are not equally informative. They differ not only in the biological signal they capture but in how readily that signal can be analyzed, and this asymmetry shapes what can be learned from each alone. In this paper, we present a deep learning model that integrates gene expression and chromatin accessibility modalities across discrete single-cell timepoints, enabling the characterization of cellular dynamics.

While existing studies[6–9] have provided valuable insights on modeling cellular development using either gene expression or chromatin accessibility, each modality offers only a partial and unevenly reliable view. Gene expression provides detailed transcriptional readouts but is affected by dropouts and mRNA degradation,[10] whereas chromatin accessibility reports regulatory potential yet is sparse, high-dimensional, and of low specificity, leaving it difficult to interpret in isolation and typically lacking cell-type annotations.[11] This reflects not just a difference in content, but in analyzability. Core developmental analyses, such as reconstructing trajectories, ordering cells in pseudotime, and assigning cell identities, are unreliable or altogether infeasible when performed on a sparse, weakly separated modality such as chromatin accessibility on its own, because its cells do not cleanly resolve into the groups and paths these analyses depend on. Integration is usually motivated by completeness, such that combining complementary modalities yields a fuller description of a biological system.[12,13] We emphasize a second, less appreciated consequence. Because the modalities share underlying biology, integration offers the opportunity to transfer the analytical strength of a richer modality to a weaker one, anchoring the poorly resolved modality within a shared representation so that analyses it cannot support alone become possible. Realizing this opportunity, rather than merely fusing modalities, is a central aim of our work.

In practice, however, the structure of the data works against this. Single-cell sequencing is inherently destructive, so each modality is typically measured in different cell populations, yielding datasets that are "unaligned", where no cell-to-cell correspondence exists between modalities. Co-assay technologies that profile multiple modalities within the same cells (e.g., gene expression and chromatin accessibility)[14,15] do provide such correspondence, but they remain resource-intensive and are often limited by low sequencing depth,[16,17] which restricts

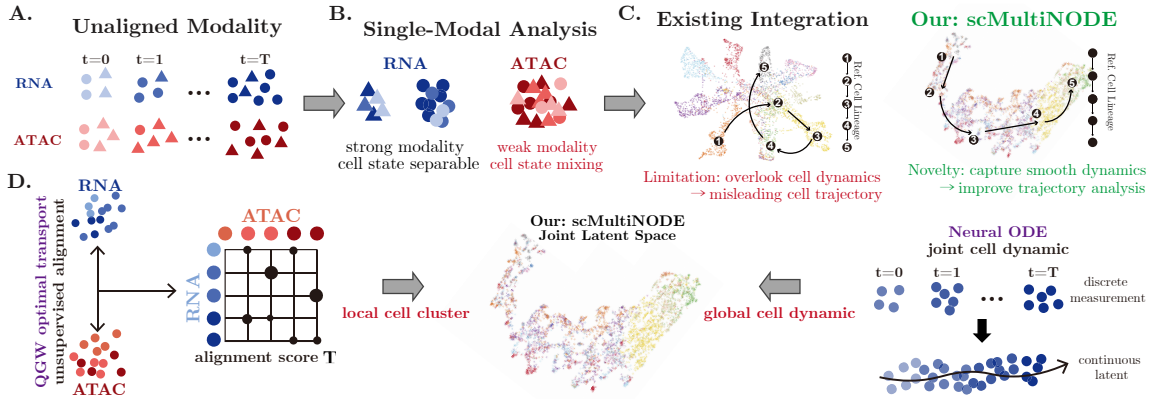


Fig. 1. **Model overview.** (**A**) We integrate multi-timepoint scRNA-seq (gene expression) and scATAC-seq (chromatin accessibility) data without prior cell-to-cell correspondences across modalities or timepoints. (**B**) Single-modal analysis on the weaker modality limits accurate cell-state identification. (**C**) Existing integration methods (e.g., SCOTv2[18]) improve the weaker modality by leveraging the stronger one, but often fail to capture continuous cellular dynamics, yielding trajectories (solid arrows) that break continuity, contradict reference lineages, or connect unrelated cell groups. In contrast, our `scMultiNODE` explicitly models smooth and continuous dynamics, improving temporal analysis. (**D**) `scMultiNODE` aligns multi-modal datasets using QGW optimal transport to infer cross-modal cell correspondences while preserving cell-type structure, and models continuous cell dynamics with neural ODEs.

their scalability across developmental timepoints and complex tissues. The unaligned setting sharpens the asymmetry described above. Without measured correspondence, there is no direct bridge along which a weak modality can borrow structure from a strong one, and the weak modality frequently lacks annotations altogether or carries only sparse, biased ones that are difficult to use. Consequently, most temporally resolved multi-modal datasets remain *unaligned across modalities and time*, so the analyses a weak modality would benefit from most, like trajectory inference and cross-modal comparison, are the hardest to perform on it.

Currently, there are no computational methods that can jointly integrate unaligned multi-modal temporal datasets and explicitly model cellular dynamics over time. Existing approaches, primarily applied to static snapshot datasets, either align different modalities by identifying correlated patterns or common biological signals across modalities[19,20] or adopt alignment techniques that better capture intricate cross-modality structures.[18,21–23] They achieve flexible integration, particularly by aligning cell types across modalities, but treat integration as a symmetric fusion aimed at completeness and do not model how cells evolve over time. Although the recent scTIE model[24] combines multiple timepoints from temporal data into a shared space for gene regulation analysis, it is not applicable to unaligned datasets and does not explicitly model cellular development. As a result, the integrated latent spaces of existing methods do not naturally encode cellular dynamics,[25] limiting their ability to resolve temporally dependent transitions and branching developmental trajectories. Fig. 1B and Sec. 4.2 show state-of-the-art integration methods (e.g., SCOTv2[18]) learning an alignment of developmental multi-modal single-cell data but yielding trajectories that break continuity, contradict reference lineages, or connect unrelated cell groups, which has a direct and overlooked cost for weak modalities. A method that cannot represent developmental structure at

all cannot transfer that structure to a poorly resolved modality, so even where static labels can be shared, dynamic analyses on the weak modality remain out of reach. Closing this gap requires a method that integrates unaligned temporal multi-modal data, explicitly models cellular dynamics, and thereby carries developmental structure to the modality that cannot recover it alone.

We propose single-cell Multi-modal Neural Ordinary Differential Equation (`scMultiNODE`), which integrates gene expression (scRNA-seq) and chromatin accessibility (scATAC-seq) profiles while explicitly modeling cellular dynamics over multiple timepoints (Fig. 1). `scMultiNODE` is a novel model that does not require prior cell-to-cell correspondence information across modalities or timepoints, as opposed to other multi-modal dynamic modeling approaches. Specifically, `scMultiNODE` aligns cells across modalities and timepoints using Quantized Gromov-Wasserstein (QGW) optimal transport,[26] which identifies biologically similar cells and scales to large-scale datasets. Guided by this alignment, `scMultiNODE` constructs a joint latent space and explicitly learns smooth developmental trajectories using a single shared neural ordinary differential equation (ODE). Critically, because both modalities are governed by the same learned dynamics in this joint space, a sparse and weakly separated modality such as chromatin accessibility is placed onto the developmental flow inferred from the cleaner modality, rather than being modeled in isolation. As a result, `scMultiNODE` is the novel method to preserve both local cell relationships (e.g., cell type distinctions) and global cellular dynamics (e.g., complex developmental trajectories) across multiple modalities during integration (Fig. 1D), a capability not achieved by previous multi-modal integration methods.

We evaluate `scMultiNODE` on four developmental single-cell datasets, comprising two co-assay datasets and two unaligned datasets, using scRNA-seq and scATAC-seq assays measured at multiple timepoints across different species and tissues. Our analyses demonstrate that `scMultiNODE` integrates gene expression and chromatin accessibility measurements well in both unaligned and co-assay datasets. Moreover, `scMultiNODE` significantly outperforms baseline integrative models in capturing cellular dynamics while retaining cell-type variation. Beyond achieving high-quality integration, we demonstrate that `scMultiNODE` performs corrective integration. By leveraging complementary transcriptomic information, `scMultiNODE` recovers coherent developmental trajectories from scATAC-seq, a modality whose sparsity and weak cell-type separability often obscure trajectory structure when analyzed alone. `scMultiNODE` also identifies development-related, ATAC-specific genes and transfers annotations from RNA to unlabeled ATAC cells. Together, these results show that `scMultiNODE` produces an interpretable joint latent space that captures complex cell trajectories, supports the reconstruction of cell transition paths, and facilitates the discovery of development-related genes, which are key elements for studying cell differentiation. We envision that `scMultiNODE` will be critical to integrative analyses of multi-modal temporal single-cell datasets, especially those with unaligned measurements.

2. Methods

`scMultiNODE` aligns scRNA-seq (gene expression) and scATAC-seq (chromatin accessibility) datasets measured across time. We denote $\mathbf{X}^{(t)} \in \mathbb{R}^{m_t \times p}$ as the gene expression of m_t cells and p genes at timepoint t and $\mathbf{Y}^{(t)} \in \mathbb{R}^{n_t \times q}$ as the chromatin accessibility of n_t cells and q

chromatin regions at timepoint t. Vectors are denoted in bold lowercase letters, and matrices are in bold capital letters. Given gene expression and chromatin accessibility matrices at observed timepoints $\mathcal{T}_{\text{RNA}}, \mathcal{T}_{\text{ATAC}} \subseteq \{0, 1, 2, \cdots\}$, we aim to learn a unified low-dimensional representation that brings together multiple modalities, reflects cell type variations, and preserves dynamic cellular progression for downstream developmental analysis. We assume consistent feature sets across timepoints: the same genes for scRNA-seq and the same chromatin regions for scATAC-seq. While no prior cell-to-cell correspondence is needed, we assume the datasets share common biological signals, such as overlapping cell types and similar developmental trajectories, making joint integration both meaningful and feasible. `scMultiNODE` infers the necessary correspondence during learning and learns a single d-dimensional latent representation ($d \ll \min\{p, q\}$) that we require to: (1) place biologically similar cells from the two modalities close together, so that local cell-type structure is shared rather than confined to one modality, and (2) arrange cells so that their representations reflect continuous global progression over time.

2.1. *Scalable integration while preserving local cell relationships*

In this multi-modal integration problem, the important step is to align two single-cell modalities measured separately and representing distinct feature spaces. To achieve this, we use Gromov-Wasserstein (GW) optimal transport[27] to align cells across modalities in an unsupervised manner.

GW optimal transport aims to move data points from one feature space to another while preserving the original local geometry captured by intra-modality distances. The central concept of GW is to find the best data correspondence matrix $\mathbf{T}$ that denotes the alignment between each data point across two data modalities, through minimizing the transport cost (i.e., the GW distance) as

$$\text{GW}(\mathbf{X}_{\text{ALL}}, \mathbf{Y}_{\text{ALL}}) = \min_{\mathbf{T} \in \Pi(\nu_X, \nu_Y)} \sum_{i,i',j,j'} \| \mathbf{D}^{\text{RNA}}_{ii'} - \mathbf{D}^{\text{ATAC}}_{jj'} \|^2 \, \mathbf{T}_{ij} \mathbf{T}_{i'j'}. \tag{1}$$

Here, $\mathbf{X}_{\text{ALL}} = \text{CONCAT}(\mathbf{X}^{(t)} | t \in \mathcal{T}_{\text{RNA}})$ and $\mathbf{Y}_{\text{ALL}} = \text{CONCAT}(\mathbf{Y}^{(t)} | t \in \mathcal{T}_{\text{ATAC}})$ are concatenations of gene expression and chromatin accessibility of all cells across available timepoints. ν_X and ν_Y are empirical distributions of $\mathbf{X}_{\text{ALL}}$ and $\mathbf{Y}_{\text{ALL}}$, respectively. $\mathbf{D}^{\text{RNA}}$ and $\mathbf{D}^{\text{ATAC}}$ are intra-modality distance matrices. Due to high dimensionality and sparsity of single-cell data, `scMultiNODE` employs an autoencoder (AE)[28] to map each modality to its latent variable $\mathbf{Z}$ and construct intra-modality distance matrices in the latent space with k-nearest neighbor (kNN) distance:

$$\begin{aligned} \mathbf{D}^{\text{RNA}} = \text{kNN}(\mathbf{Z}_{\text{RNA}}) \quad \text{and} \quad \mathbf{D}^{\text{ATAC}} = \text{kNN}(\mathbf{Z}_{\text{ATAC}}) \quad \text{where} \\ \mathbf{Z}_{\text{RNA}} = \text{Enc}_{\mathbf{X}}(\mathbf{X}_{\text{ALL}}, \ \phi_{\mathbf{X}}), \quad \widehat{\mathbf{X}}_{\text{ALL}} = \text{Dec}_{\mathbf{X}}(\mathbf{Z}_{\text{RNA}}, \ \theta_{\mathbf{X}}); \\ \mathbf{Z}_{\text{ATAC}} = \text{Enc}_{\mathbf{Y}}(\mathbf{Y}_{\text{ALL}}, \ \phi_{\mathbf{Y}}), \quad \widehat{\mathbf{Y}}_{\text{ALL}} = \text{Dec}_{\mathbf{Y}}(\mathbf{Z}_{\text{ATAC}}, \ \theta_{\mathbf{Y}}), \end{aligned} \tag{2}$$

with encoder $\text{Enc}(\cdot, \phi)$ and decoder $\text{Dec}(\cdot, \theta)$ for each modality.

In the cell correspondence matrix $\mathbf{T}$, the element $\mathbf{T}_{ij}$ indicates the alignment probability of RNA cell i and ATAC cell j. The GW loss function measures how intra-modality cell-pair

distances in one dataset compare with those in the other dataset. So, the GW distance does not directly compare data points across different feature spaces, but instead finds a cross-space sample that best preserves intra-modality structure. This ensures that the transport plan $\mathbf{T}$ can preserve local geometry, like cell clusters and cell-cell neighboring relationships, in one modality and correspond to similar local structures in the other when we align two single-cell modalities. We binarize the predicted $\mathbf{T}$ such that for a RNA cell i, $\mathbf{T}_{ij} = 1$ if ATAC cell j has the highest alignment probability with i; otherwise $\mathbf{T}_{ij} = 0$. Notice that `scMultiNODE` does not require cell correspondence information as a prior and predicts it in an unsupervised manner[a].

While previous unsupervised integration methods have successfully aligned single cells using GW optimal transport,[27] the exact computation of the GW distance (Eq. 1) is NP-hard and incurs high computational costs for large-scale single-cell datasets. Therefore, `scMultiNODE` employs Quantized Gromov-Wasserstein (QGW),[26] a divide-and-conquer strategy (Supp. Sec. S1) that recursively partitions and aligns feature-space blocks, enabling efficient computation and scalability to large datasets (evaluated in Supp. Sec. S10).

With the help of the inferred cell correspondence $\mathbf{T}$, `scMultiNODE` can now integrate gene expression and chromatin accessibility measurements into a joint latent space. Specifically, `scMultiNODE` uses a neural network $\text{Fus}(\cdot,\ \omega) : \mathbb{R}^d \mapsto \mathbb{R}^d$ parameterized by ω, named the fusion layer, to map modality-specific cell latent variables of each modality to the d-dimensional joint latent space through

$$\tilde{\mathbf{Z}} = \text{Fus}(\mathbf{Z},\ \omega) \quad \text{with} \quad \mathbf{Z} \in \{\mathbf{Z}_{\text{RNA}}, \mathbf{Z}_{\text{ATAC}}\}. \tag{3}$$

We hypothesize that if cells i and j from different modalities are biologically similar, they should be aligned in the joint latent space and their latent representations $\tilde{\mathbf{z}}_i$ and $\tilde{\mathbf{z}}_j$ should be close to each other. To this end, `scMultiNODE` aligns modalities in the joint latent space by minimizing the loss function

$$\mathcal{L}_{\text{fusion}} = \alpha \sum_{i,j} \mathbf{1}\left(\mathbf{T}_{ij} \neq 0\right) \parallel \tilde{\mathbf{z}}_i - \tilde{\mathbf{z}}_j \parallel_2^2 + \text{MSE}\left(\mathbf{X}_{\text{ALL}},\ \widehat{\mathbf{X}}_{\text{ALL}}\right) + \text{MSE}\left(\mathbf{Y}_{\text{ALL}},\ \widehat{\mathbf{Y}}_{\text{ALL}}\right) \tag{4}$$

with the indication function $\mathbf{1}(\cdot)$. Latent variables $\tilde{\mathbf{z}}_i = \text{Fus}(\mathbf{z}_i, \omega)$ and $\tilde{\mathbf{z}}_j = \text{Fus}(\mathbf{z}_j, \omega)$ for RNA cell i and ATAC cell j (from Eq. 3), and

$$\begin{aligned} \widehat{\mathbf{X}}_{\text{ALL}} &= \text{Dec}_{\mathbf{X}}(\tilde{\mathbf{Z}}_{\text{RNA}},\ \theta_{\mathbf{X}}) \text{ with } \tilde{\mathbf{Z}}_{\text{RNA}} = \{\tilde{\mathbf{z}}_i \mid \forall \text{ RNA cells } i\}, \\ \widehat{\mathbf{Y}}_{\text{ALL}} &= \text{Dec}_{\mathbf{Y}}(\tilde{\mathbf{Z}}_{\text{ATAC}},\ \theta_{\mathbf{Y}}) \text{ with } \tilde{\mathbf{Z}}_{\text{ATAC}} = \{\tilde{\mathbf{z}}_j \mid \forall \text{ ATAC cells } j\}. \end{aligned} \tag{5}$$

Therefore, during fusion layer training, we minimize the distance between the latent representations of cell pairs matched by the GW optimal transport alignment (i.e., $\mathbf{1}\left(\mathbf{T}_{ij} \neq 0\right)$), weighted by the hyperparameter α. Since the decoders operate on the joint latent space, we also include the MSE reconstruction loss in the fusion objective (Eq. 4) to update the model.

2.2. *Incorporating global cellular dynamics with neural ODE*

`scMultiNODE` uses QGW to capture local cell relationships in the joint latent space. However, the joint space for now does not account for global cellular dynamics, such as developmental

[a]If cell correspondence is known (even partially) ahead, we can encode such information into $\mathbf{T}$ and skip the alignment procedure. Our experiments assume no cell correspondence is given.

trajectories or lineage bifurcations. All the previous integration methods have this limitation, hence, their latent representations poorly define natural cell paths across timepoints and hinder downstream analysis.

To address the problem, `scMultiNODE` uses neural ODEs[29] to explicitly model cell developmental dynamics in the joint latent space. Neural ODEs parameterize the temporal evolution of latent cell states by learning a neural network–based derivative function. While existing flow-based frameworks[9,30] have used neural ODEs (with respect to time) to construct continuous-time trajectories and model single-cell development from gene expression data, `scMultiNODE` extends them to model the aligned representations of multi-modal datasets. It quantifies changes of cell latent representation $\tilde{\mathbf{Z}}^{(t)}$ in the joint latent space at time t through a neural ODE: $\mathrm{d}\tilde{\mathbf{Z}}^{(t)} = \mathrm{Drift}\left(\tilde{\mathbf{Z}}^{(t)},\ \mu\right)\mathrm{d}t$. Here, $\mathrm{Drift}(\tilde{\mathbf{Z}}^{(t)}, \mu)$ is a single non-linear neural network parameterized by μ and shared across both modalities, modeling the instantaneous rate of change of any cell's latent state in the joint space. `scMultiNODE` calculates the initial condition $\tilde{\mathbf{Z}}^{(0)}$ of cells at the first time $t = 0$, defined as the earliest observed timepoint for both modalities. `scMultiNODE` then predicts the subsequent cell states step-wise at any timepoint t through (assume RNA modality has the first timepoint as $0 \in \mathcal{T}_{\mathrm{RNA}}$)

$$\begin{aligned} \tilde{\mathbf{Z}}^{(t+\Delta t)} &= \tilde{\mathbf{Z}}^{(t)} + \mathrm{Drift}\left(\tilde{\mathbf{Z}}^{(t)},\ \mu\right)\Delta t \quad \text{with} \\ \tilde{\mathbf{Z}}^{(0)} = \mathrm{Fus}\left(\mathbf{Z}^{(0)}_{\mathrm{RNA}},\ \omega\right) \quad &\text{and} \quad \mathbf{Z}^{(0)}_{\mathrm{RNA}} = \mathrm{Enc}_{\mathbf{X}}\left(\mathbf{X}^{(0)},\ \phi_{\mathbf{X}}\right), \end{aligned} \tag{6}$$

Here, hyperparameter Δt denotes step size. We use the first-order Euler method (in Eq. 6) for convenience of explanation and one can specify any ODE solver in our implementations.

To fit the continuous trajectory (controlled by Drift neural network) to the observations, `scMultiNODE` minimizes the difference between the input and the reconstructed data. Specifically, at each measured timepoint $t \in \mathcal{T}_{\mathrm{RNA}}$ or $t \in \mathcal{T}_{\mathrm{ATAC}}$, `scMultiNODE` uses the decoder $\mathrm{Dec}_{\mathbf{X}}/\mathrm{Dec}_{\mathbf{Y}}$ to convert latent variables $\tilde{\mathbf{Z}}^{(t)}$ generated from Eq. 6 back to the high-dimensional feature space. Because we have no correspondence between true cells and cells generated from the ODE model, `scMultiNODE` utilizes the 2-Wasserstein metric[31] to measure the distance between distributions defined by ground truth $\mathbf{X}$ and predictions $\widehat{\mathbf{X}}$ as

$$W_2(\mathbf{X}, \widehat{\mathbf{X}}) = \Big(\min_{\Gamma \sim \Pi(\mathbf{X}, \widehat{\mathbf{X}})} \sum_{i,j} D_{ij}^2 \Gamma_{ij}\Big)^{1/2} \quad \text{with} \quad D_{ij} = \|\mathbf{X}_i - \widehat{\mathbf{X}}_j\|_2. \tag{7}$$

Wasserstein distance measures the distribution distance in the same metric space, different from the GW version that is applied for aligning different metric spaces. Here, $\Pi(\mathbf{X}, \widehat{\mathbf{X}})$ denotes the set of all transport plans between each cell of $\mathbf{X}$ and $\widehat{\mathbf{X}}$ and D_{ij} represents the ℓ_2 distance, such that the Wasserstein metric adopts the minimal-cost transport plan Γ to measure the data dissimilarity. `scMultiNODE` utilizes Wasserstein distance as the reconstruction loss when training the neural ODE.

Furthermore, to incorporate the cellular dynamics of neural ODE, `scMultiNODE` uses a dynamic regularization term to update the joint latent space with global developmental dynamics. The dynamic regularization is proposed in our previous work[9] for modeling temporal scRNA-seq data, which incorporates cellular dynamics into the latent space to make it more robust to distribution shifts in the measurements across time. Here, we extend dynamic regularization in the multi-modal integration setting. Specifically, the dynamic regularization

minimizes the difference between the joint latent representations generated by the fusion layer (i.e., $\mathrm{Fus}\left(\mathrm{Enc}_{\mathbf{X}}(\mathbf{X}, \phi_{\mathbf{X}})\right)$ or $\mathrm{Fus}\left(\mathrm{Enc}_{\mathbf{Y}}(\mathbf{Y}, \phi_{\mathbf{Y}})\right)$) and the dynamics learned by the ODE (i.e., $\tilde{\mathbf{Z}}$). Because we have no correspondence between them, `scMultiNODE` again uses Wasserstein distance to evaluate their difference at each timepoint and defines the dynamic regularization:

$$\begin{aligned} \mathcal{R} &= \sum_{t \in \mathcal{T}_{\mathrm{RNA}}} W_2\left(\tilde{\mathbf{Z}}_{\mathrm{fus}}^{(t)},\ \tilde{\mathbf{Z}}^{(t)}\right) + \sum_{t \in \mathcal{T}_{\mathrm{ATAC}}} W_2\left(\tilde{\mathbf{Z}}_{\mathrm{fus}}^{(t)},\ \tilde{\mathbf{Z}}^{(t)}\right) \\ \text{with } &\begin{cases} \tilde{\mathbf{Z}}_{\mathrm{fus}}^{(t)} = \mathrm{Fus}\left(\mathrm{Enc}_{\mathbf{X}}(\mathbf{X}^{(t)}, \phi_{\mathbf{X}}), \omega\right) & \text{for } t \in \mathcal{T}_{\mathrm{RNA}} \\ \tilde{\mathbf{Z}}_{\mathrm{fus}}^{(t)} = \mathrm{Fus}\left(\mathrm{Enc}_{\mathbf{Y}}(\mathbf{Y}^{(t)}, \phi_{\mathbf{Y}}), \omega\right) & \text{for } t \in \mathcal{T}_{\mathrm{ATAC}} \end{cases}, \\ &\text{and} \quad \tilde{\mathbf{Z}}^{(t)} \text{ comes from neural ODE} \quad \text{(Eq. 6)} \end{aligned} \tag{8}$$

This term folds the learned dynamics back into the shared representation. Because the constraint applies identically to both modalities, a cell whose own features weakly determine its latent position is nonetheless drawn toward the shared, global developmental flow.

With these, `scMultiNODE` jointly optimizes all components by minimizing the regularized loss function

$$\mathcal{L}_{\mathrm{dyn}} = \sum_{t \in \mathcal{T}_{\mathrm{RNA}}} W_2(\mathbf{X}^{(t)},\ \widehat{\mathbf{X}}^{(t)}) + \sum_{t \in \mathcal{T}_{\mathrm{ATAC}}} W_2(\mathbf{Y}^{(t)},\ \widehat{\mathbf{Y}}^{(t)}) + \beta \mathcal{R}. \tag{9}$$

By embedding cellular dynamics, `scMultiNODE` improves upon previous integration models. This enables more effective data fitting and yields a joint latent space that is more robust and interpretable by capturing both global cellular dynamics and local cell relationships. Pseudocode for `scMultiNODE` is provided in Supp. Algorithm S1.

3. Experiment Setup

Datasets. We use four developmental single-cell datasets with scRNA-seq and scATAC-seq measured at multiple timepoints and thousands to tens of thousands of cells (Supp. Table S1): two co-assay datasets — human cortex (HC)[12] and human organoid (HO);[5] two unaligned datasets — drosophila embryogenesis (DR),[13] and mouse neocortex (MN).[32] For each modality, we keep the top 2,000 most variable features and use batch-corrected data (Supp. Sec. S3).

Baselines. We compare `scMultiNODE` with six state-of-the-art unsupervised single-cell integration methods that are capable of aligning different single-cell measurements (details in Supp. Sec. S4): `Seurat`,[19] `UnionCom`,[33] `uniPort`,[23] `Pamona`, `SCOTv1`,[22] and `SCOTv2`.[18]

Evaluation metrics. We evaluate each model's integrated latent representations from multiple perspectives (details in Supp. Sec. S5): *modality integration*, *cell type preservation*, *cell label transfer*, and *cellular dynamics*. Modality integration is assessed using `batch entropy`,[23] fraction of samples closer than the true match (`FOSCTTM`),[22] `neighborhood overlap`,[33] and Spearman correlation coefficient (`SCC`). Cell type preservation is quantified by normalized mutual information (`NMI`)[34] between Louvain clustering[35] labels in the integrated latent space and ground truth cell type. Cell label transfer is evaluated by kNN-based label transfer accuracy using cell type labels (`LTA-type`).[18,23,33] To assess cellular dynamics, we compute label transfer accuracy using timepoint labels (`LTA-time`) and the distance correlation between latent representations and timepoint labels (`time correlation`).

Implementation. scMultiNODE is trained sequentially. Modality-specific AEs are pretrained, followed by QGW cell correspondence and fusion-layer optimization (Eq. 4), and then neural ODE training (Eq. 9) to learn cellular dynamics. The full procedure is detailed in Supp. Sec. S2. For fair comparison, all methods are tuned with Optuna[36] using broad search ranges (Supp. Sec. S6) and identical evaluation criteria, without using labels during training. All methods use a joint latent dimension of 50 and are trained to convergence. Ablation studies and hyperparameter guidance for scMultiNODE are provided in Supp. Sec. S9.

4. Experiment Results

4.1. *scMultiNODE integrates modalities without sacrificing temporal or rare-population structure*

Static integration discards temporal structure while scMultiNODE preserves it. The quality of the integrated latent space is first assessed using 2D Uniform Manifold Approximation and Projection (UMAP)[37] and Principal Component Analysis (PCA) projections (Supp. Fig. S1-S4). The visualizations show that scMultiNODE effectively aligns the two modalities across all timepoints while preserving cellular dynamics. In contrast, baseline methods either fail to achieve effective cross-modal alignment over time (e.g., Pamona and UnionCom) or neglect temporal structure (e.g., SCOTv2), resulting in latent spaces dominated by cell type signals. Across all datasets, scMultiNODE consistently preserves both cell type variation and temporal cellular dynamics during integration.

Regarding the batch entropy score, scMultiNODE consistently outperforms the baselines across all datasets (Fig. 2A and E). Furthermore, scMultiNODE achieves the highest time correlation across all datasets (Fig. 2C and G), demonstrating that it aligns the two modalities well while preserving the underlying dynamics.

Then, we validate whether the integration preserves distinct cell type separations (Fig. 2B and F). scMultiNODE achieves cell-type clustering performance comparable to existing baselines across most datasets. For instance, scMultiNODE attains NMI=0.41 on HO dataset (baseline median NMI=0.405) and NMI=0.32 on MN dataset (baseline median NMI=0.30). Plotting LTA-type against LTA-time (Fig. 2D and H) demonstrates that scMultiNODE uniquely occupies the upper region of both axes, achieving the highest LTA-time at a comparable LTA-type level. In contrast, baseline methods fail to simultaneously preserve temporal and cell-type heterogeneity. These results suggest that scMultiNODE sacrifices minimal static cell-type separation to recover developmental trajectories that are overlooked by static integration methods. While incorporating cell-type supervision may improve clustering metrics, we find that it compromises modality alignment and temporal continuity (Supp. Sec. S8), likely due to rigid class constraints or noisy annotations that distort the latent space. Therefore, we adopt an unsupervised integration strategy throughout our experiments to better capture continuous cellular transitions. A complete list of metric values is provided in the Supplementary Note.

Single-modal dynamics are asymmetric across modalities while scMultiNODE overcome it. To assess the benefit of jointly modeling RNA and ATAC with scMultiNODE , we compared it against scNODE,[9] a single-modal dynamics model trained separately on RNA (scNODE (RNA)) and ATAC (scNODE (ATAC)) across all datasets (Fig. 3A-B). No-

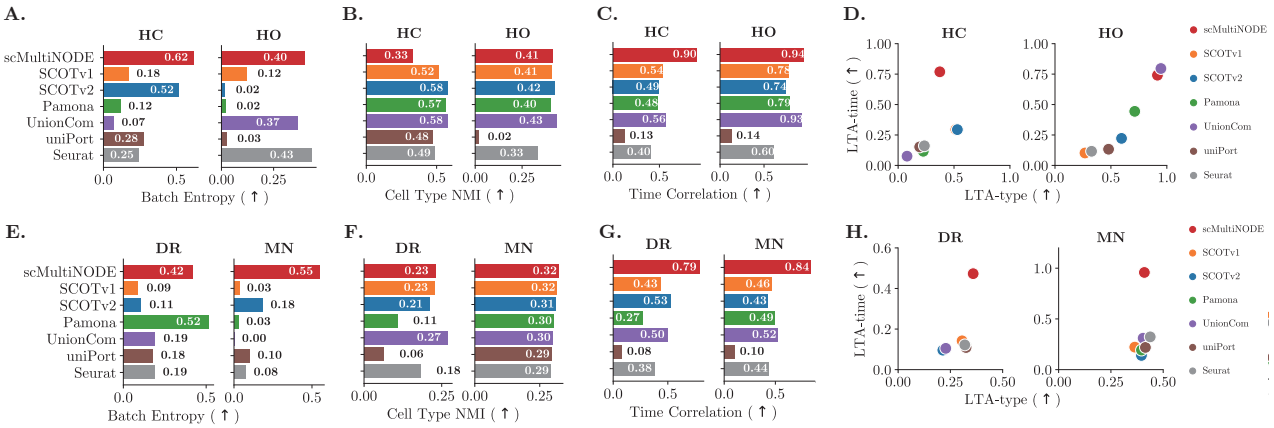


Fig. 2. Integration evaluation on all datasets. (**A** and **E**) `Batch entropy` as an example of modality integration performance. (**B** and **F**) Cell type `NMI` score. (**C** and **G**) `Time correlation` evaluates integrations on capturing cell dynamics. (**D** and **H**) Label transfer accuracy (LTA) across modalities. We plot the `LTA-type`/`LTA-time` scores on the X/Y axes for all the methods, respectively. Mean performance across five independent random seeds.

tably, `scNODE` performance is highly asymmetric, with `scNODE (ATAC)` substantially underperforming `scNODE (RNA)`, reflecting the sparser, noisier signal in ATAC data. In contrast, `scMultiNODE` largely closes this gap, significantly outperforming `scNODE (ATAC)` in `time correlation` while achieving the comparable or even better `time correlation` than `scNODE (RNA)`. A similar pattern holds for `NMI`, where `scMultiNODE` outperforms `scNODE (ATAC)` on three of four datasets, while remaining comparable to `scNODE (RNA)` on HO, the same trade-off noted above. These results indicate that `scMultiNODE` 's key advantage lies in leveraging shared cross-modal structure to substantially improve dynamics and cell-type resolution for the weaker modality, without compromising performance on the stronger RNA modality.

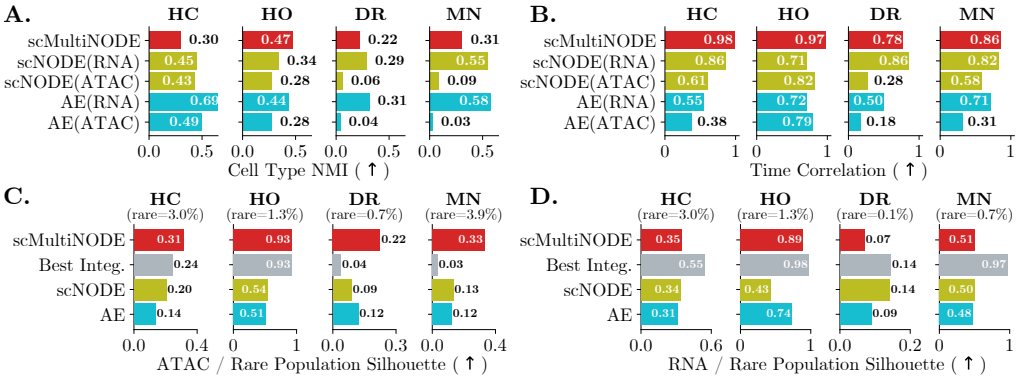


Fig. 3. (**A**, **B**) We compare `scMultiNODE` with scNODE and AE trained on each modality separately in terms of cell-type `NMI` and `time correlation`. (**C**, **D**) Silhouette score of rare cell populations in the ATAC (**C**) and RNA (**D**) modalities, where `scMultiNODE` is compared against the best integration baseline, scNODE, and AE. The fraction of rare cells is annotated above each panel ("Best Integ": best integration baseline model).

`scMultiNODE` rescues rare populations in the weak modality. Sparsity and low feature specificity make rare populations hard to resolve in scATAC-seq, limiting downstream analysis.

To evaluate rare-population separation, we first perform integration and then compute a one-vs-rest silhouette score for the rare cell type (i.e., cell types with the fewest cells) in the learned joint latent space for ATAC and RNA (Fig. 3C-D).

On the ATAC modality, methods that incorporate either multi-modal integration or cellular dynamics outperform the single-modal AE, indicating that each component independently improves rare-population separation. The best integration baseline improves over the single-modal AE on some datasets (e.g., 0.24 of best integration vs. 0.14 of AE on HC), but collapses on others and fall below the AE (e.g., DR and MN), where the integration is dominated by the stronger RNA modality at the expense of ATAC. Modeling cellular dynamics alone (i.e., `scNODE`) yields only modest gains (e.g., 0.54 vs. 0.51 on HO), suggesting that neither integration nor dynamics alone is sufficient to overcome the challenges of weak modalities. By jointly modeling cross-modal correspondence and cellular dynamics, `scMultiNODE` achieves the highest silhouette score on all datasets. For example, `scMultiNODE` attains a silhouette score of 0.31 on HC, compared with 0.24 for the best integration baseline and 0.20 for scNODE (Fig. 3C). The improvement is even more significant on unaligned MN data, where `scMultiNODE` achieves 0.33 while all other methods remain below 0.15. Notably, `scMultiNODE` is the only method that avoids ATAC collapse on these datasets, highlighting the importance of jointly modeling integration and dynamics for reliable rare-population separation in weak modalities.

Although `scMultiNODE` yields the largest improvements on the weaker ATAC modality, it remains competitive on RNA (Fig. 3D). Relative to the matched single-modal RNA representation, `scMultiNODE` retains or improves rare-population separability. It is only surpassed by the dedicated integration baselines on RNA, but those same methods collapse on ATAC. This reflects a expected trade-off that because RNA already provides strong cell-identity signals, these baselines can achieve high RNA performance by effectively discarding ATAC, whereas `scMultiNODE` learns a balanced representation that resolves rare populations across both modalities. These results demonstrate that combining multi-modal integration with dynamic modeling enables `scMultiNODE` to more reliably resolve rare cell populations, particularly in weaker modalities such as scATAC-seq.

Together, these results highlight `scMultiNODE` 's advantages. It preserves cell-type structure, better captures temporal dynamics than static integration methods, narrows the RNA–ATAC performance gap, and better resolves rare populations in the weaker modality. Ablation studies show robustness to latent-space size. `scMultiNODE` also accurately reconstructs leave-one-timepoint-out cells, demonstrating predictive capability beyond cross-modal integration (Supp. Sec. S16), while QGW captures coarse lineage structure and `scMultiNODE` improves cross-modal alignment (Supp. Sec. S14). Despite these capabilities, `scMultiNODE` remains computationally efficient and scalable (Supp. Sec. S9).

4.2. *Joint dynamics recover multifurcating trajectories that weak modality cannot support alone*

We demonstrate that `scMultiNODE` learns an interpretable joint latent space that preserves complex cellular developmental trajectories in multi-modal data. Previous methods typically infer pseudotime from the integrated latent space to reconstruct cell lineages.[19] However, because they rely on static latent representations without explicitly modeling cellular dynamics,

they often struggle to capture complex cell trajectories.

We conduct the evaluation on the human cortex (HC) dataset, which contains complex developmental trajectories. Pseudotime is inferred from each latent space using Monocle3[38] and evaluated by the Spearman correlation (ρ) with the ground-truth trajectory progression (Fig. 4A). We compare `scMultiNODE` with multi-modal integration baselines lacking dynamic modeling, as well as single-modality AE and scNODE models trained separately on RNA/ATAC. ATAC provides a particularly challenging test due to its sparsity and weak cell-type separation, allowing us to assess whether integration improves trajectory recovery in a weak modality.

To test whether `scMultiNODE` 's performance can be achieved with a simpler two-step pipeline, we also evaluate `scNODE (transfer)`. Here, `scNODE` is trained on one modality to infer pseudotime, which is then transferred to the other modality via regression after alignment with each integration baseline (Supp. Sec. S11). We evaluate both RNA-to-ATAC and ATAC-to-RNA transfer, reporting the best result across all integration baselines.

Methods that model dynamics explicitly (`scMultiNODE` and `scNODE`(RNA)) preserve global trajectory structure, whereas static integration baselines produce collapsed or misaligned branches. For example, `scMultiNODE` reaches $\rho = 0.84$ on the EN lineage against 0.47 for the best integration baseline. On the Oligodendrocyte lineage `scMultiNODE` has $\rho = 0.55$, comparable to SCOTv1 ($\rho = 0.60$) and above all others (Fig. 4B).

The uplift on ATAC is the sharpest result. From ATAC alone, no baseline recovers the branching structure. The static AE reaches $\rho = 0.14$ (Oligodendrocyte) and 0.52 (EN), and `scNODE`(ATAC) fails on the Oligodendrocyte lineage. By contrast, `scMultiNODE` attains $\rho = 0.55$ (Oligodendrocyte) and $\rho = 0.84$ (EN), recovering coherent trajectories from a modality that does not support accurate inference on its own. The two-step pipeline is not sufficient. `scNODE (transfer)` improves on static integration ($\rho = 0.61$ vs the best integration baseline's 0.47 on EN) by carrying learned dynamics across the alignment, but underperforms `scMultiNODE` on every branch and falls below single-modal `scNODE` in several cases, because errors in the intermediate alignment propagate into the transferred pseudotime. Dynamics and integration must be learned jointly rather than composed. We additionally perform this analysis using PAGA,[39] a topology-based pseudotime inference method, and obtain similar results, indicating that `scMultiNODE` integration preserves not only trajectory direction but also complex topological structures (Supp. Fig. S7).

These results demonstrate that explicitly modeling global cellular dynamics and jointly integrating modalities are essential for accurate trajectory inference. Rather than simply adding RNA information, `scMultiNODE` leverages its joint architecture to transfer developmental dynamics through a shared neural ODE, enabling weak modalities such as ATAC to recover complex trajectories that cannot be resolved independently.

Least Action Path (LAP) analysis[30] of the `scMultiNODE` integration recovered coherent trajectories toward oligodendrocytes and glutamatergic neurons (Supp. Fig. S11), with trajectory genes showing stronger temporal variation than random controls and including known developmental regulators. Notably, the weak modality contributed signals not captured by the strong modality, improving driver gene detection and highlighting complementary biological information from multi-modal integration (Supp. Sec. S12).

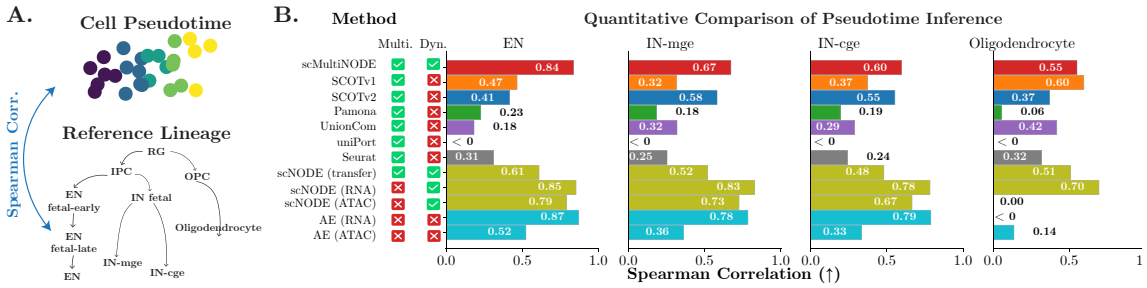


Fig. 4. (**A**) Consistency between cell pseudotime and reference lineage on the HC dataset, with RG branching into four lineages. (RG: Radial Glia, IPC: Intermediate Progenitor Cells, OPC: Oligodendrocyte Precursor Cells, EN: Excitatory Neurons, IN: Inhibitory Neurons) (**B**) Spearman correlation (ρ) between ground-truth cell-state order and Monocle3 pseudotime. "Multi.": multi-modal integration; "Dyn.": models incorporating cell dynamics.

4.3. *Label transfer makes unannotated chromatin data interpretable*

Germ layers are foundational embryonic structures that give rise to diverse organs and cell types, making their identification essential for understanding development.[40] While scRNA-seq provides strong transcriptional signals for annotation, scATAC-seq remains challenging due to limited labels.[11] This unaligned setting complicates existing approaches for modeling biological processes across modalities. To address this challenge, we leverage the integration learned by `scMultiNODE` to transfer germ layer annotations from RNA to ATAC, enabling interpretation of the weakly annotated ATAC modality (Supp. Sec. S13).

We evaluated label transfer on the unaligned DR dataset, where germ layer labels were available only for RNA. A random forest classifier trained on RNA embeddings accurately predicted ATAC labels in `scMultiNODE`'s joint latent space, producing coherent germ layer clusters consistent with developmental trajectories (Supp. Fig. S12). Marker gene analysis of ATAC-derived gene activity confirmed clear germ layer separation (Supp. Fig. S13), whereas baseline methods showed poor resolution. Gene Ontology (GO) enrichment further validated biological relevance, revealing germ layer–specific functions (Supp. Fig. S14), such as nervous system development in the ectoderm group.[41] Overall, `scMultiNODE` preserves dynamics across modalities and enables accurate cross-modal label transfer, improving the interpretability of chromatin accessibility data.

5. Conclusion

`scMultiNODE` integrates temporally resolved, multi-modal single-cell datasets across unaligned modalities and timepoints, reconstructing population-level trajectories from destructive measurements rather than tracking individual cells longitudinally. Its deterministic neural ODE may not capture stochastic or heterogeneous dynamics, population birth/death, or modality-specific cell states, motivating future stochastic and population-aware extensions.

`scMultiNODE` complements RNA-velocity methods by modeling temporally sampled, potentially unpaired modalities without requiring paired measurements or spliced/unspliced RNA counts, enabling applications where kinetic measurements are unavailable.

Acknowledgments

This project was funded by the National Institutes of Health (NIH) award 1R35HG011939-01.

References

1. D. G. Spiller, C. D. Wood, D. A. Rand and M. R. White, Measurement of single-cell dynamics, *Nature* **465**, 736 (2010).
2. W. Zakrzewski, M. Dobrzyński, M. Szymonowicz and Z. Rybak, Stem cells: past, present, and future, *Stem cell research & therapy* **10**, 1 (2019).
3. J. Ding, N. Sharon and Z. Bar-Joseph, Temporal modelling using single-cell transcriptomics, *Nature Reviews Genetics* **23**, 355 (2022).
4. R. S. Ziffra, C. N. Kim, J. M. Ross, A. Wilfert, T. N. Turner, M. Haeussler, A. M. Casella, P. F. Przytycki, K. C. Keough, D. Shin *et al.*, Single-cell epigenomics reveals mechanisms of human cortical development, *Nature* **598**, 205 (2021).
5. J. S. Fleck, S. M. J. Jansen, D. Wollny, F. Zenk, M. Seimiya, A. Jain, R. Okamoto, M. Santel, Z. He, J. G. Camp *et al.*, Inferring and perturbing cell fate regulomes in human brain organoids, *Nature* **621**, 365 (2023).
6. H. A. Pliner, J. S. Packer, J. L. McFaline-Figueroa, D. A. Cusanovich, R. M. Daza, D. Aghamirzaie, S. Srivatsan, X. Qiu, D. Jackson, A. Minkina *et al.*, Cicero predicts cis-regulatory dna interactions from single-cell chromatin accessibility data, *Molecular cell* **71**, 858 (2018).
7. G. H. T. Yeo, S. D. Saksena and D. K. Gifford, Generative modeling of single-cell time series with PRESCIENT enables prediction of cell trajectories with interventions, *Nature communications* **12**, p. 3222 (2021).
8. G. Huguet, D. S. Magruder, A. Tong, O. Fasina, M. Kuchroo, G. Wolf and S. Krishnaswamy, Manifold interpolating optimal-transport flows for trajectory inference, *Advances in Neural Information Processing Systems* **35**, 29705 (2022).
9. J. Zhang, E. Larschan, J. Bigness and R. Singh, scnode: generative model for temporal single cell transcriptomic data prediction, *Bioinformatics* **40**, ii146 (2024).
10. D. Lähnemann, J. Köster, E. Szczurek, D. J. McCarthy, S. C. Hicks, M. D. Robinson, C. A. Vallejos, K. R. Campbell, N. Beerenwinkel, A. Mahfouz *et al.*, Eleven grand challenges in single-cell data science, *Genome biology* **21**, 1 (2020).
11. H. Chen, C. Lareau, T. Andreani, M. E. Vinyard, S. P. Garcia, K. Clement, M. A. Andrade-Navarro, J. D. Buenrostro and L. Pinello, Assessment of computational methods for the analysis of single-cell atac-seq data, *Genome biology* **20**, 1 (2019).
12. K. Zhu, J. Bendl, S. Rahman, J. M. Vicari, C. Coleman, T. Clarence, O. Latouche, N. M. Tsankova, A. Li, K. J. Brennand *et al.*, Multi-omic profiling of the developing human cerebral cortex at the single-cell level, *Science Advances* **9**, p. eadg3754 (2023).
13. D. Calderon, R. Blecher-Gonen, X. Huang, S. Secchia, J. Kentro, R. M. Daza, B. Martin, A. Dulja, C. Schaub, C. Trapnell *et al.*, The continuum of drosophila embryonic development at single-cell resolution, *Science* **377**, p. eabn5800 (2022).
14. S. Ma, B. Zhang, L. M. LaFave, A. S. Earl, Z. Chiang, Y. Hu, J. Ding, A. Brack, V. K. Kartha, T. Tay *et al.*, Chromatin potential identified by shared single-cell profiling of rna and chromatin, *Cell* **183**, 1103 (2020).
15. Y. Hu, K. Huang, Q. An, G. Du, G. Hu, J. Xue, X. Zhu, C.-Y. Wang, Z. Xue and G. Fan, Simultaneous profiling of transcriptome and dna methylome from a single cell, *Genome biology* **17**, 1 (2016).
16. C. Zhu, S. Preissl and B. Ren, Single-cell multimodal omics: the power of many, *Nature methods* **17**, 11 (2020).
17. X. Wu, X. Yang, Y. Dai, Z. Zhao, J. Zhu, H. Guo and R. Yang, Single-cell sequencing to multi-omics: technologies and applications, *Biomarker Research* **12**, p. 110 (2024).
18. P. Demetci, R. Santorella, M. Chakravarthy, B. Sandstede and R. Singh, Scotv2: Single-cell multiomic alignment with disproportionate cell-type representation, *Journal of Computational*

Biology **29**, 1213 (2022).
19. Y. Hao, T. Stuart, M. H. Kowalski, S. Choudhary, P. Hoffman, A. Hartman, A. Srivastava, G. Molla, S. Madad, C. Fernandez-Granda *et al.*, Dictionary learning for integrative, multimodal and scalable single-cell analysis, *Nature biotechnology* **42**, 293 (2024).
20. J. Liu, C. Gao, J. Sodicoff, V. Kozareva, E. Z. Macosko and J. D. Welch, Jointly defining cell types from multiple single-cell datasets using liger, *Nature protocols* **15**, 3632 (2020).
21. K. Cao, Y. Hong and L. Wan, Manifold alignment for heterogeneous single-cell multi-omics data integration using pamona, *Bioinformatics* **38**, 211 (2022).
22. P. Demetci, R. Santorella, B. Sandstede, W. S. Noble and R. Singh, Scot: single-cell multi-omics alignment with optimal transport, *Journal of computational biology* **29**, 3 (2022).
23. K. Cao, Q. Gong, Y. Hong and L. Wan, A unified computational framework for single-cell data integration with optimal transport, *Nature Communications* **13**, p. 7419 (2022).
24. Y. Lin, T.-Y. Wu, X. Chen, S. Wan, B. Chao, J. Xin, J. Y. Yang, W. H. Wong and Y. R. Wang, Data integration and inference of gene regulation using single-cell temporal multimodal data with sctie, *Genome research* **34**, 119 (2024).
25. M. Connor, G. Canal and C. Rozell, Variational autoencoder with learned latent structure, in *International conference on artificial intelligence and statistics*, 2021.
26. S. Chowdhury, D. Miller and T. Needham, Quantized gromov-wasserstein, in *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21*, 2021.
27. G. Peyré, M. Cuturi *et al.*, Computational optimal transport: With applications to data science, *Foundations and Trends® in Machine Learning* **11**, 355 (2019).
28. S. Sun, J. Zhu, Y. Ma and X. Zhou, Accuracy, robustness and scalability of dimensionality reduction methods for single-cell rna-seq analysis, *Genome biology* **20**, 1 (2019).
29. R. T. Chen, Y. Rubanova, J. Bettencourt and D. K. Duvenaud, Neural ordinary differential equations, *Advances in neural information processing systems* **31** (2018).
30. X. Qiu, Y. Zhang, J. D. Martin-Rufino, C. Weng, S. Hosseinzadeh, D. Yang, A. N. Pogson, M. Y. Hein, K. H. J. Min, L. Wang *et al.*, Mapping transcriptomic vector fields of single cells, *Cell* **185**, 690 (2022).
31. M. Cuturi, Sinkhorn distances: Lightspeed computation of optimal transport, *Advances in neural information processing systems* **26** (2013).
32. W. Yuan, S. Ma, J. R. Brown, K. Kim, V. Murek, L. Trastulla, A. Meissner, S. Lodato, A. S. Shetty, J. Z. Levin *et al.*, Temporally divergent regulatory mechanisms govern neuronal diversification and maturation in the mouse and marmoset neocortex, *Nature Neuroscience* **25**, 1049 (2022).
33. K. Cao, X. Bai, Y. Hong and L. Wan, Unsupervised topological alignment for single-cell multi-omics integration, *Bioinformatics* **36**, i48 (2020).
34. M. D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Müller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché *et al.*, Benchmarking atlas-level data integration in single-cell genomics, *Nature methods* **19**, 41 (2022).
35. V. D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, Fast unfolding of communities in large networks, *Journal of statistical mechanics: theory and experiment* **2008**, p. P10008 (2008).
36. T. Akiba, S. Sano, T. Yanase, T. Ohta and M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019.
37. L. McInnes, J. Healy and J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, *arXiv preprint arXiv:1802.03426* (2018).
38. J. Cao, M. Spielmann, X. Qiu, X. Huang, D. M. Ibrahim, A. J. Hill, F. Zhang, S. Mundlos, L. Christiansen, F. J. Steemers *et al.*, The single-cell transcriptional landscape of mammalian

organogenesis, *Nature* **566**, 496 (2019).
39. F. A. Wolf, F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. Göttgens, N. Rajewsky, L. Simon and F. J. Theis, Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells, *Genome biology* **20**, p. 59 (2019).
40. Y. Zheng, R. Z. Yan, S. Sun, M. Kobayashi, L. Xiang, R. Yang, A. Goedel, Y. Kang, X. Xue, S. N. Esfahani *et al.*, Single-cell analysis of embryoids reveals lineage diversification roadmaps of early human development, *Cell stem cell* **29**, 1402 (2022).
41. S. T. Crews, Drosophila embryonic cns development: neurogenesis, gliogenesis, cell fate, and differentiation, *Genetics* **213**, 1111 (2019).

Absorption and Co-expression Modules Show Where Polygenic and Proteomic Risk Scores Diverge in Neurodegenerative Diseases

Charles Zheng*

Department of Bioengineering,
University of Pennsylvania, Philadelphia, PA, USA

Manu Shivakumar*, Li Shen and Dokyoon Kim†

Department of Biostatistics, Epidemiology and Informatics,
University of Pennsylvania, Philadelphia, PA, USA
E-mail: dokyoon.kim@pennmedicine.upenn.edu†

Polygenic and proteomic risk scores are both proposed for pre-symptomatic stratification, yet the extent to which they provide overlapping or complementary information has not been measured across neurodegenerative disease. To quantify this overlap, we define absorption as the fraction of a polygenic score's predictive contribution accounted for by an out-of-fold proteomic score and estimate it among 9,434 to 9,820 UK Biobank participants. Absorption did not track h^2, as Alzheimer's with *APOE*, Alzheimer's without *APOE* and Parkinson's carried matched SNP h^2 of 0.068, 0.061 and 0.069 yet absorbed 0.73, 0.39 and 0.19, with amyotrophic lateral sclerosis at 0.22. Residual genetic signal remained in all four, indicating that proteomic risk scores did not fully capture the predictive information contained in polygenic risk scores. The proteins associated with a polygenic score and the proteins a proteomic score selects overlap no more often than chance, converging only where *APOE* dominates. Across 33 plasma co-expression modules built in 41,358 disease-free participants, germline signal concentrates in modules rather than spreading, with the summary component of seven modules associated with the score for Alzheimer's with *APOE* and none for Parkinson's. The lysosomal module M32 has 89% of its members associated with the score for Alzheimer's with *APOE*, an association absent without *APOE*, yet none was used by the proteomic score. Rebuilding the co-expression modules in *All of Us* gave an adjusted Rand index of 0.582 against the 0.686 attainable within that cohort, and 29 of 33 modules stayed together above a permutation null. We show that transfer is stable and that the disjointness of the proteomic and genomic protein sets is not an artifact of how the sets were defined.

Keywords: Polygenic risk score; proteomic risk score; multi-omics integration; protein co-expression network; module preservation; Alzheimer's disease; Parkinson's disease; APOE.

*These authors contributed equally to this work.
†This author is the corresponding author.

© 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.

1. Introduction

Two stratification methods have been proposed and implemented for estimating patient risk before disease onset. The first uses the germline, static genome, summarized as a polygenic risk score (PRS). The second uses proteomic readings in plasma and sometimes cerebrospinal fluid (CSF), which represents the patient's state of health and physiology more dynamically and is now assayable at scale by affinity platforms.[1,2] With multi-modality data becoming increasingly available, clarifying how much one modality substitutes for the other becomes essential for determining whether PRS and ProRS provide redundant or complementary predictive information, and whether both are needed for risk stratification.

Recent work by Woerner *et al.* showed that proteomic risk scores (ProRS) trained on prevalent samples often surpassed PRS in predictive performance on incident cases while retaining complementary information, but also observed that when used in conjunction, the fraction of PRS's contribution after conditioning on ProRS scaled with heritability of the predicted trait.[3] However, Woerner *et al.* did not specifically analyze, on a protein or pathway resolution, how the two scores overlap or substitute for each other, nor did they individually categorize neurodegenerative diseases such as Alzheimer's Disease (AD), Parkinson's disease (PD), amyotrophic lateral sclerosis (ALS), and frontotemporal dementia (FTD).

Previous studies have shown that *APOE* $\varepsilon 4$ carriage leaves a reproducible imprint on the circulating proteome.[4–6] This presents evidence that plasma protein levels, from which the ProRS is built, are associated with genomic variants, so individual proteins may also associate with a PRS, which aggregates the effects of many loci. Therefore, we ask the question for each of AD, PD, ALS and FTD: to what extent can ProRS can account for proportion of genetic risk explained by the PRS signal, and does this correspond to the same individual proteins being prioritized by each score?

Proteomic and polygenic stratification may converge on the same biological processes without involving the same individual proteins. For example, genetically associated and disease-predictive proteins may occupy different roles within one pathway, and many individually weak associations may become detectable only when treated as a coordinated system. Co-expression modules provide such a higher-order representation by capturing biologically coherent processes, and protein modules recover reproducible biology across cohorts and tissues.[7–10] Prior serum work further showed that modules associated with incident disease and mortality are under cis- and trans-genetic regulation.[11] We therefore ask whether the predictive redundancy observed between PRS and ProRS reflects sparse associations across the plasma proteome or concentration within a limited number of co-expression modules, and whether the two scores converge at the module-level despite limited overlap among individual proteins.

We address these questions in UK Biobank, using 2,920 Olink analytes, polygenic scores built exclusively from non-UK-Biobank discovery GWAS, and up to 41,358 participants without a recorded diagnosis for the proteome-wide work. Our module-level analysis replicated modestly in 9,211 disease-free participants of *All of Us*, assayed on a later Olink panel. Our contributions are fourfold. We quantify the extent to which proteomic risk captures the predictive contribution of polygenic risk as an effect size with a bootstrap interval, and find it substantial for AD, partial in every disease examined. We then show that the proteins ProRS

selects and the ones associated with polygenic risk are largely different populations, overlapping no more often than chance except for AD with APOE. We then analyzed 33 data-derived co-expression modules and observed the two sets' modular separation as well as patterns of their respective enrichment. Finally, we ask whether those module definitions are stable across cohort transfer by rebuilding them in All of Us and measuring their agreement with UK Biobank, which identifies which modules are reproducible across cohorts, for use in later studies relating germline or proteomic risk to the plasma proteome.

2. Methods

2.1. *Cohorts, phenotyping and quality control*

Discovery analyses used UK Biobank,[12,13] restricted to participants of European genetic ancestry with Olink Explore 3072 proteomics at the baseline instance.[1] Cases were defined from linked hospital and death records by ICD-10 code (AD G30, PD G20, ALS G12.2, FTD G31.0) and classified as prevalent or incident relative to venepuncture. Three distinct analysis populations are used. The nested logistic models use the complete-case intersection of proteome, genotype and covariates with a downsampled control pool, giving 9,434–9,820 participants per disease. The association analyses use every participant with genotype data but without a recorded diagnosis at venepuncture, including the incident diagnoses, giving 38,602–39,807 per disease. Modules are built once in the 41,358 participants with proteomic data and no recorded AD diagnosis. These participants were not screened for PD, ALS or FTD, so about 700 of them may carry one of those diagnoses. Covariates throughout were age, age squared, sex and the first ten genotype principal components. Olink NPX values were used as released by UK Biobank, with no additional plate or processing-batch adjustment, because plate and batch identifiers were not available in our data extract.

Modules and absorption were assessed in an external cohort. *All of Us* contributed 9,969 participants with Olink Explore HT plasma proteomics.[14,15] Of the 2,920 UK Biobank module proteins, 2,819 were present on Explore HT by gene symbol and survived QC. Proteins missing in more than 20% of participants and participants missing more than 20% of proteins were dropped, and the remainder were standardized and mean-imputed, whereas the UK Biobank matrix was k-nearest-neighbour imputed.

2.2. *Polygenic risk scores*

To preclude sample-overlap inflation, discovery summary statistics were taken exclusively from non-UK-Biobank GWAS (Table 1).[16–19] Posterior effect sizes were obtained with PRS-CS under the continuous-shrinkage prior against the 1000 Genomes European LD reference,[20] and scores computed with PLINK.[21] Two AD scores were constructed, one genome-wide and one excluding chr19:44.4–46.5 Mb (GRCh37),[22] with the exclusion applied to the summary statistics before shrinkage so that the resulting posterior is *APOE*-free by construction. SNP heritability on the liability scale was estimated by LD score regression.[23]

2.3. *Proteomic risk scores*

ProRS[3] were fitted by LASSO logistic regression with `glmnet`,[24] with proteins penalized and covariates forced in unpenalized, and λ selected by inner cross-validation. A repeated stratified five-fold design was used in which all prevalent cases enter every training fold while incident cases and a downsized control sample are scored strictly out of fold; the entire procedure was repeated three times with distinct seeds, giving fifteen selection opportunities per protein. In each repetition, controls were randomly downsampled to 10,000 of the roughly 41,000 to 43,000 available, and these controls together with all incident cases were divided into five case-control-stratified folds, with prevalent cases added to every training set. This gave roughly 13 to 38 controls per case in training depending on disease, and the ratio was not varied. Because the ProRS is never exposed to the genetic score, one cached ProRS per disease is compared against every genetic score. A protein counts as ProRS-selected when it is selected in at least 7 of the 15 folds. As a sensitivity analysis on this threshold itself, the cut was swept from 1 to 15 folds; the observed-to-expected overlap between the PRS-associated and ProRS-selected sets for AD with *APOE* rose across the entire range, from parity at $t = 1$ to 5.7-fold at $t = 15$, crossing nominal significance ($p < 0.05$) at $t = 3$ and remaining significant through $t = 15$, while none of the other three scores reached nominal significance at any threshold (Supplementary Fig. S1). The reported separation is therefore not because of the selected 7-of-15 cut. As a sensitivity analysis, every ProRS was refitted by elastic net at $\alpha = 0.5$ with the penalty being the only difference.

2.4. *Proteomic absorption of polygenic predictive signal*

For each disease we fitted four nested logistic models on an identical complete-case set, these being covariates alone, covariates plus PRS, covariates plus ProRS, and both.[3] Using ΔR^2 increment,[25] absorption is

$$A \;=\; 1 - \frac{\Delta R^2(\mathrm{PRS} \mid \mathrm{cov} + \mathrm{ProRS})}{\Delta R^2(\mathrm{PRS} \mid \mathrm{cov})},$$

the proportion of the polygenic score's own predictive contribution accounted by the proteome. Confidence intervals came from a case-control-stratified bootstrap with 2000 resamples, and between-disease comparisons use a paired bootstrap that resamples the shared control pool once and applies the same resample to both diseases.[26] The DeLong test on the area under the curve is also reported.[27]

2.5. *PRS-associated proteins*

A protein was called PRS-associated if its abundance tracked that score among participants without a recorded diagnosis at proteomic assessment, including controls and incident cases. Prevalent cases were excluded so that established disease, its treatment, or its physiological consequences could not alter protein abundance and be misattributed to a PRS association. Each of the 2,920 assayed proteins was regressed individually on the standardized PRS with age, age^2, sex and PC1–10 as covariates, and the resulting p-values were Benjamini–Hochberg corrected at $q < 0.05$ within each score, and every subsequent q value carries the same correction. Two genomic regions are reported separately from the rest. The major histocompatibility

complex and the 17q21.31 inversion[1,28,29] were associated with a very wide range of traits and contain strong *cis*-pQTLs, so every polygenic score carried weight there and every score therefore appeared to associate with the proteins those regions encode. Proteins encoded in the two regions are held out of the proteome-share denominator for that reason and counted separately (Fig. 1a).

2.6. *Plasma proteome co-expression modules*

In the 41,358 disease-free participants, each protein was residualized on the covariate design by ordinary least squares, retaining residuals only. Residuals were standardized and Pearson-correlated pairwise, and the signed distance $1 - r$ clustered by Ward's minimum-variance criterion following established practice.[7] The resulting dendrogram was cut at twelve candidate resolutions spanning $K = 8$ to 50 clusters. Each resolution was characterized on reproducibility, measured by partitioning participants at random into halves and scoring each half's partition against the full-data partition by the adjusted Rand index (ARI) over five independent splits, as well as the smallest module the cut produced (Fig. 2c). Reproducibility rises monotonically across the sweep and so cannot select a resolution on its own; thus, we selected for $K = 35$ as the finest resolution before the smallest module size degrades from a plateau held at 17 proteins from $K = 20$ through $K = 35$ to 8 at $K = 40$ and 6 at $K = 50$, relative to a floor of 15 proteins per module (dashed line in Fig. 2c). The two neighboring resolutions, $K = 30$ and $K = 40$, both have split-half reproducibility within 0.03 of the selected cutoff (0.742 and 0.771 against 0.766 at $K = 35$) and agree with the retained 33-module partition at ARI 0.962 and 0.986 respectively, so $K = 35$ is stable relative to its neighbors (Supplementary Table S1). Each module was summarized by its eigenprotein, the score on the first principal component of its members' residualized abundances, which assigned one value to every participant. Any pair of eigenproteins correlating above 0.85 was merged. The retained cut requested 35 clusters, and two such merges returned the 33 modules analyzed throughout (Fig. 2c).

Modules were annotated by hypergeometric over-representation of Gene Ontology terms with the assayed panel as background. Module-level association was assessed by regressing each module's standardized eigenprotein on the standardized polygenic score with the same covariates, in disease-free participants only so that disease-consequence cannot confound any discovery. Each module was additionally tested for over-representation of the PRS-associated set and of the ProRS-selected set by hypergeometric test against the assayed panel, giving three per-module axes in total.

2.7. *Cross-cohort comparison*

Module transfer was assessed in *All of Us* using proteins shared with UK Biobank. The two cohorts were residualized and partitioned independently through the same clustering procedure, and agreement was quantified by the adjusted Rand index (ARI). As an internal reference ceiling, *All of Us* was split in half ten times, each pair of half-sample partitions was compared, and the median ARI was retained. Module-specific preservation was assessed by the fraction of each UK Biobank module's protein pairs assigned to the same *All of Us* module. This co-location fraction was compared with 2,000 size-matched random protein

sets, converted to a z-score, and tested using one-sided p-values with Benjamini–Hochberg correction across 33 modules. In a complementary analysis, UK Biobank module membership was held fixed without reclustering *All of Us*. Module coherence, defined as the variance explained by the first principal component, was recomputed in *All of Us* and compared with the corresponding UK Biobank values.[30] UK Biobank modules were derived in European-similar disease-free participants, whereas the *All of Us* analysis included all ancestries because restricting to the 2,681 European-similar participants would reduce the available sample from 9,211 by approximately fourfold.

Absorption was additionally evaluated in *All of Us* by applying the frozen UK Biobank ProRS weights without refitting. Refitting was not attempted because only 20 to 56 disease cases were available (Supplementary Table S2), so the transferred score tests the UK Biobank weights rather than the *All of Us* proteome's best achievable fit.

3. Results

3.1. *Proteomic substitution for polygenic risk is partial in every disease and its magnitude tracks one locus*

Adding ProRS to the covariates model improved discrimination in every disease and by a wider margin than what PRS achieved, with AD rising from an area under the curve of 0.819 for covariates alone to 0.855 with PRS and 0.903 with ProRS (Table 2), echoing *Woerner et al.*'s findings. Absorption was 0.727 (95% CI 0.645–0.806) for the *APOE*-containing AD score, 0.394 (0.119–0.667) for AD with *APOE* excluded, 0.221 (0.009–0.509) for ALS and 0.187 (0.001–0.380) for PD. Because the 95% confidence interval for PD absorption reaches down to zero, the data cannot rule out a complete absence of proteomic absorption of genetic risk in PD. The residual polygenic increment's interval excluded zero in all four contrasts, with $\Delta R^2(\text{PRS} \mid \text{ProRS})$ at 0.0226 (0.0134–0.0341), 0.0057 (0.0014–0.0128), 0.0081 (0.0016–0.0184) and 0.0047 (0.0013–0.0101) for AD with APOE, without APOE, ALS and PD respectively.

Table 1. *Study design. Phenotype counts are for European-ancestry UK Biobank participants with Olink plasma proteomics and complete covariates, before restriction to participants with a polygenic score. SNP-heritability is on the liability scale from LD score regression.*

Trait	ICD-10	Discovery GWAS	Prevalent	Incident	Controls	Liability h^2 (SE)
AD, with *APOE*	G30	Kunkle 2019	73	432	41,358	0.068 (0.011)
AD, no *APOE*	G30	Kunkle 2019	73	432	41,358	0.061 (0.009)
PD	G20	FinnGen R13	105	652	42,372	0.069 (0.012)
ALS	G12.2	van Rheenen 2021	36	220	42,873	0.028 (0.003)
FTD*	G31.0	Ferrari 2014	20	69	43,040	0.026 (0.022)

*FTD is unstable (the standard error of $h^2 \approx$ the estimate) and is excluded from subsequent claims.

AD with APOE had the greatest absorption of any disease model at 0.73 whereas PD was the lowest at 0.19, and SNP heritability alone cannot explain the differences between them, since liability-scale h^2 is 0.068, 0.061 and 0.069 for the AD, AD-without-*APOE* and

PD scores, all within one standard error of each other (Table 1). Direct cross-disease comparisons should be interpreted cautiously because they involve different disease-specific case populations and PRS predictive increments, and because the PD estimate is imprecise, with its lower bound touching zero. The cleaner contrast is therefore within AD, where the outcome, analysis population, covariates, and ProRS were held constant. Excluding the APOE region reduced absorption from 0.73 to 0.39, isolating the APOE region as a major contributor to the unusually high absorption observed for AD.

An attempt to repeat the analysis in *All of Us* reproduces the point estimates but cannot address the contrast. Absorption for the *APOE*-containing AD score was 0.470 (0.248–0.862) on 39 prevalent cases, an interval overlapping the UK Biobank estimate, and PD gave 0.249 in the European-similar subset, likewise overlapping (Supplementary Table S3). Because these analyses rest on very small samples, this cohort provides an underpowered assessment of absorption effect sizes rather than an independent validation of the disease-specific dissociations observed in UK Biobank. AD without *APOE* was not estimable there, its denominator $\Delta R^2(\text{PRS} \mid \text{cov})$ falling to about 5×10^{-5}. The AD-versus-PD contrast rests on 20 and 35 cases in that subset and is under-powered. The frozen UK Biobank proteomic scores transferred to *All of Us* without refitting, giving an AUC of 0.865 for AD and 0.791 for PD in all ancestries (Supplementary Table S2).

Table 2. *The four-model AUC and absorption results. ΔR^2 values are Nagelkerke increments. DeLong p reflects AUC significance of PRS given ProRS and covariates. Absorption intervals are percentile bootstrap ($B = 2000$, 3 cross-validation repeats). FTD is characterised in Table 1 but omitted here, its absorption denominator being below the stability threshold.*

	AUC				ΔR^2			
Trait	cov	+PRS	+ProRS	full	PRS|cov	PRS|ProRS	DeLong p	A [95% CI]
AD with *APOE*	0.819	0.855	0.903	0.908	0.0828	0.0226	0.056	0.73 [0.65, 0.81]
AD no *APOE*	0.819	0.825	0.903	0.905	0.0095	0.0057	0.125	0.39 [0.12, 0.67]
PD	0.745	0.750	0.805	0.809	0.0058	0.0047	0.010	0.19 [0.00, 0.38]
ALS	0.633	0.656	0.723	0.733	0.0104	0.0081	0.162	0.22 [0.01, 0.51]

3.2. *The proteins polygenic risk associates with and the proteins the proteomic score uses are generally distinct*

Regressing every assayed protein on each PRS in disease-free participants demonstrates the plasma's association with genomic signal on an individual-protein resolution. The *APOE*-containing AD score was associated with 7.9% of the proteome (227 of 2,873 proteins), against 1.2% for AD's polygenic background, 1.7% for ALS and 0.5% for PD (Fig. 1a; Supplementary Table S4). Proteins encoded in the MHC and in the 17q21.31 inversion are held out of that denominator and drawn separately, so percentages are computed over the 2,873 analytes outside those regions while the counts described below are over all 2,920. AD's polygenic back-

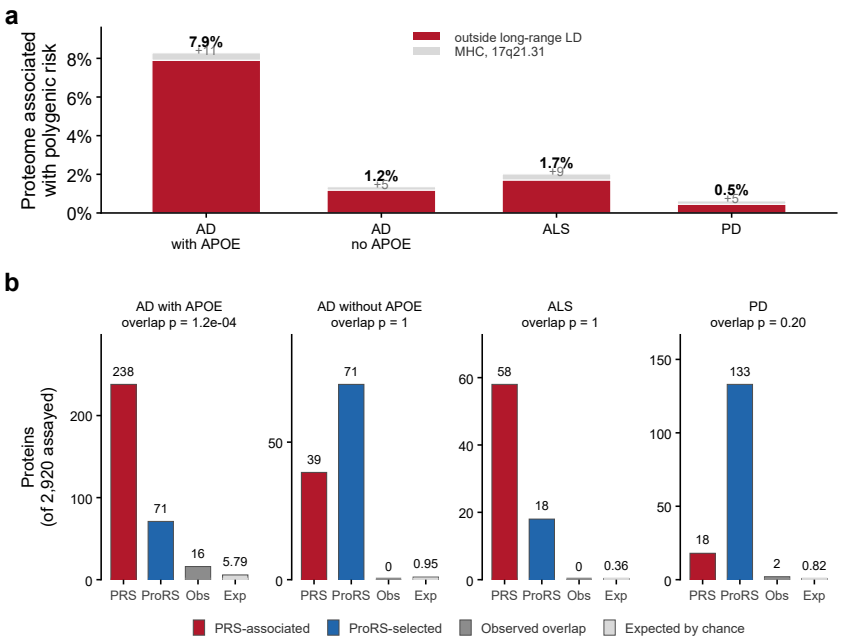


Fig. 1. *Two protein populations, in 38,602 to 39,807 participants without a recorded diagnosis at venepuncture, depending on trait. (a) Share of the assayed proteome each polygenic score is associated with at $q < 0.05$, over the 2,873 analytes outside the MHC and the 17q21.31 inversion, with the count contributed by those two regions drawn as a separate segment. (b) Absolute count of PRS-associated proteins, ProRS-selected proteins at 7 of 15 folds, their observed overlap and the overlap expected by chance, with the one-sided hypergeometric p for excess overlap above each panel.*

ground associates with the products of its own risk genes[16] (*TREM2* $q = 7.9 \times 10^{-10}$, *CD33* $q = 1.4 \times 10^{-5}$, *CR1* $q = 3.4 \times 10^{-3}$), whereas the PD score associates with neither of its main risk genes *SNCA* ($q = 0.86$) nor *GBA* ($q = 0.81$)[31] (Supplementary Table S5); *LRRK2* is not represented on the Olink panel and was not tested.

AD with APOE is the outlier here as it was for absorption. Its PRS associates with the most plasma proteins at 238, and it is the only model whose overlap with the ProRS-selected set exceeds chance, at 16 shared proteins against 5.79 expected (Supplementary Tables S6 and S7; Fig. 1b). The three other models are largely disjoint. AD without *APOE* and ALS have separate ProRS, PRS sets, and PD shares two against 0.82 expected, which is not statistically distinguishable from chance. PD has the smallest PRS-associated footprint at 18 proteins, and its low absorption further suggests that the PD score is associated with few plasma proteins, which together with its low absorption suggests a limited plasma footprint, although the PD estimates are imprecise.

3.3. *Germline signal and the proteomic score's vocabulary separate at module resolution*

Unsupervised clustering of the residualized proteome in UK Biobank yielded 33 co-expression modules after two pairs being merged from 35 modules, ranging from 17 to 717 proteins (Supplementary Table S8). Modules recovered recognizable biology, including immune cell-

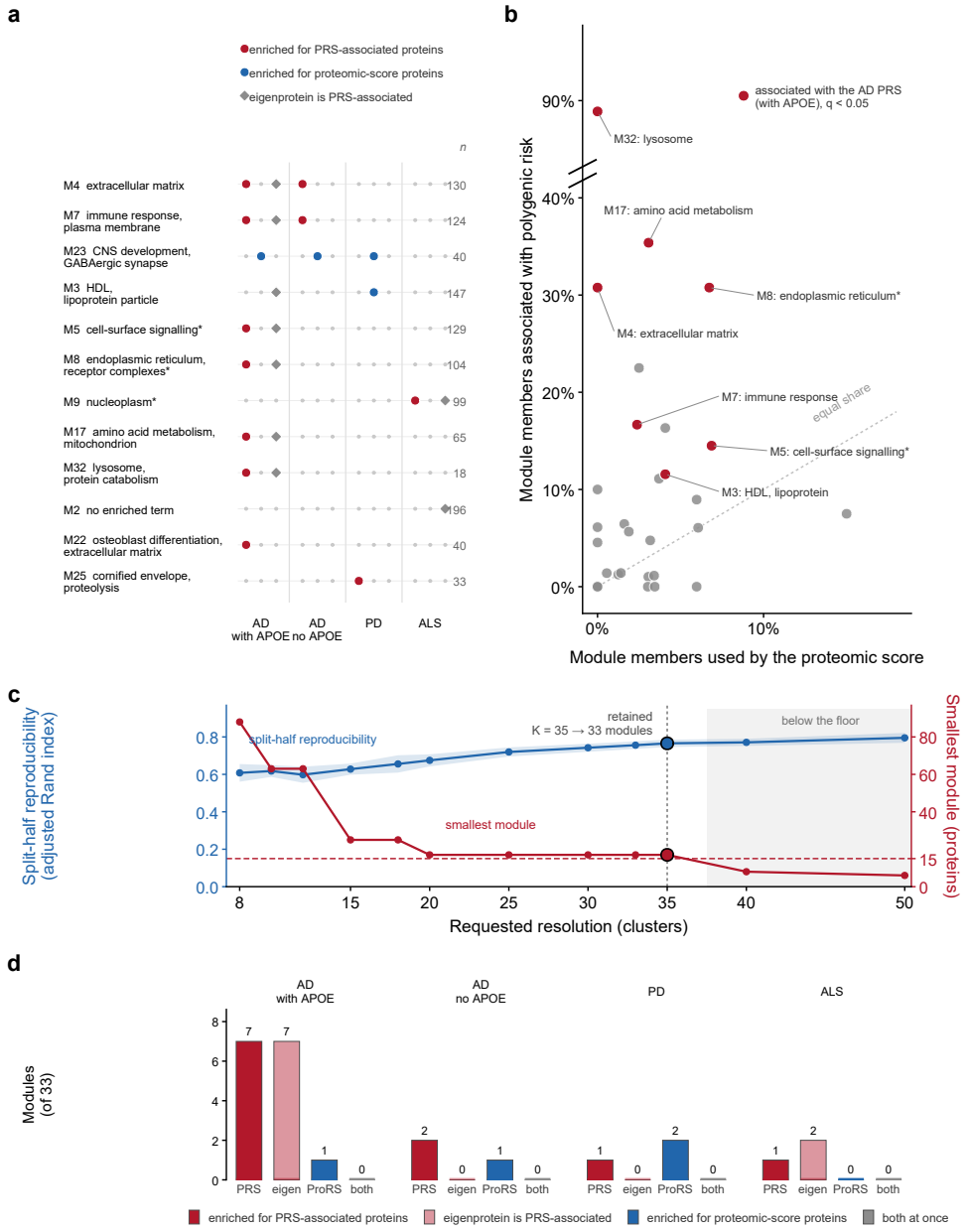


Fig. 2. *Module-resolution analysis. Modules are defined once in the 41,358 participants with no recorded AD diagnosis; the association axes are evaluated in each disease's own analysis population. (a) The 12 modules significant on at least one of the three axes, of the 33 tested. An asterisk marks a module named by its most enriched gene-ontology term where that term does not clear $q < 0.05$. (b) Per module, fraction of ProRS-selected proteins against the fraction of AD with APOE PRS-associated proteins, all 33 modules, filled red where the eigenprotein itself is associated with the AD PRS. (c) Split-half reproducibility and smallest module size across twelve resolutions, with the dashed line marking the floor of 15 proteins per module. (d) Barplot showing the counts of how many modules have eigenprotein associated with PRS, is enriched for ProRS-selected proteins (in $t \geq 7/15$ folds), is enriched for PRS-associated proteins, and is enriched for both ProRS and PRS proteins.*

surface signalling ($q = 4\times10^{-23}$), lysosomal hydrolases ($q = 3\times10^{-23}$), sarcomere ($q = 6\times10^{-10}$) and lipoprotein particles ($q = 2 \times 10^{-3}$) (Fig. 2a). Each module was analyzed for whether its membership was enriched for PRS-associated proteins, whether it was enriched for ProRS-selected proteins (selected in $t \geq 7/15$ folds), and whether its eigenprotein was associated with PRS (Fig. 2a). No module was enriched for both the ProRS-selected and the PRS-associated set in any disease, suggesting that network-level convergence of the two sets remains limited (Fig. 2d). This holds at $K = 40$; at $K = 30$ one module became significant for AD with *APOE* (ProRS enrichment $q = 0.049$ against 0.058 at $K = 35$), reflecting weaker correction over fewer modules rather than new biology. M32 and M23 are unchanged at both resolutions (Supplementary Table S9).

For PD, while 2 modules were enriched for ProRS-selected proteins, 1 was enriched for PRS-associated proteins, and no module's eigenprotein was PRS-associated. Importantly, because only 18 proteins in total met the PRS-association threshold for PD, module-level overlap tests are severely underpowered. Most modules contain 0 or 1 PRS-associated protein, so the absence of significant dual enrichment may reflect limited statistical power rather than definitive evidence of biological divergence at the module level. This contrasts with AD with APOE, where 7 modules were enriched for PRS-associated proteins and 7 module eigenproteins were themselves associated, dominated by the 18-protein lysosomal module M32, 89% of whose members are individually PRS-associated. This enrichment is absent for AD without *APOE* (PRS enrichment $q = 1.0$, eigenprotein $q = 0.058$), which is consistent with a concentration of the *APOE* signal in this module. *APOE* itself belongs to M14, and M32 holds 16 of the 238 PRS-associated proteins. None of the 18 members is selected by the ProRS at $t \geq 7/15$ folds. The two sets of seven overlap in six modules, M4, M5, M7, M8, M17 and M32, which are enriched for PRS-associated members and also associated as a module eigenprotein with PRS, but this pattern dissociates once *APOE* is removed from the score.

Although LASSO may shrink correlated proteins and thus potentially reduce the number of ProRS, PRS protein overlap, their separation here does not depend on the penalty. Refitting every ProRS by elastic net enlarged the selected sets by 17% to 48%, yet the overlap grew no faster than chance, the observed-to-expected ratio for AD with *APOE* moving from 2.76 to 2.22 while the other three models remained indistinguishable from chance, and no module was enriched for both sets under either penalty (Supplementary Figs. S2 and S3, Supplementary Table S10).

3.4. *Module definitions transfer across cohort, and coherence tracks which ones*

Upon defining $K = 33$ modules in UK Biobank, two ways to define module construction in *All of Us* at $K = 33$ were tested. First, modules were built *de novo* in *All of Us* on 2,819 shared proteins in 9,211 disease-free participants gave an adjusted Rand index of 0.582 against the UK Biobank partition. Read against the internal ceiling of 0.686, the median over ten random split-half rebuilds within *All of Us* (range 0.628 to 0.779), that recovers 84.8% of the attainable agreement (Fig. 3a).

Under the independent partitions, co-location was also tested to determine whether the

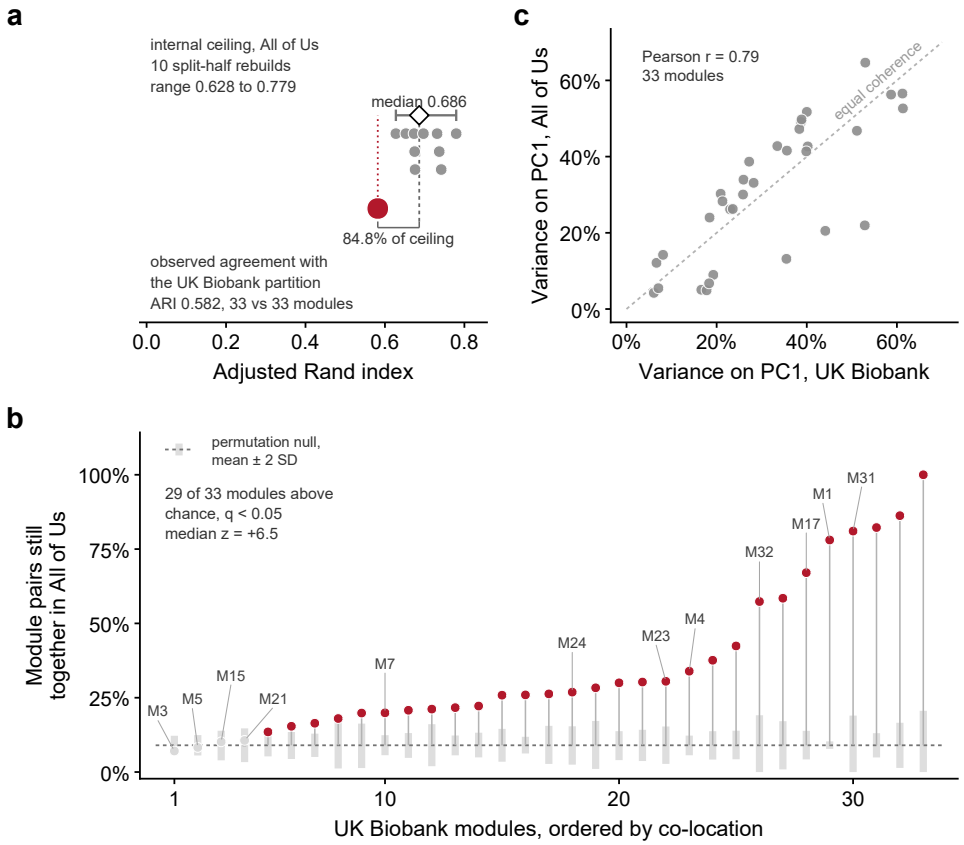


Fig. 3. *Transfer of the UK Biobank module definition to All of Us. All of Us partitioned* ***independently*** *from UK Biobank at* $K = 33$*: (a) Partition agreement between the two cohorts with the ten split-half rebuilds drawn individually and module counts annotated; (b) Per module, the fraction of within-module protein pairs still assigned together by the de novo All of Us partition, against a size-matched permutation null from the same 2,819 shared proteins. All of Us partitioned* ***identically*** *as UK Biobank: (c) Variance on the first principal component, membership held at the UK Biobank definition, measured once in each cohort.*

members of one UK Biobank module are still partitioned together in *All of Us*, and 29 of 33 modules beat a size-matched permutation null at $q < 0.05$, with a median z of +6.5 (Fig. 3b). The lysosomal module M32 keeps 57.4% of its member pairs together against a chance level near 9%, a z of +9.57. M3, the HDL and lipoprotein module, is one of the four failures at $z = -1.19$ (Supplementary Table S11).

The second approach applied UK Biobank's $K = 33$ partition unchanged to *All of Us* and asked whether the modules are equally coherent there. Comparing the variance on the first principal component of the same 33 modules across the two cohorts gave a Pearson correlation of 0.79. Coherence is therefore stable across these two Olink-assayed cohorts. Since both use the same assay platform, this does not establish that modules are free of technical structure, but it does show that internal consistency is reproducible rather than being particular to one sample. (Fig. 3c).

4. Discussion

Proteomic substitution for polygenic risk is never complete, and its size is not explained by heritability alone. Three scores whose liability-scale heritability agrees within one standard error give absorption estimates spanning most of the available range, from substantial and bounded away from zero in AD to estimates in PD that are compatible with little or no absorption (95% CI 0.00–0.38), ordered by whether the score carries a large-effect locus whose product is itself measured in plasma.[32] APOE is heavily implicated in our findings: the absorption gap, the number of PRS-associated proteins and the single lysosomal module all depend on inclusion of the *APOE* region in the score.

M32 is enriched for lysosomal proteins, holding 18 proteins across four enzyme classes. A plasma lysosomal signature of patients carrying *APOE* $\varepsilon 4$ has been reported,[6] and lysosomal *ABCA1* has been placed downstream of cholesterol accumulation in *APOE4* models, which offers a candidate explanation that our associational data do not test.[33] The association of M32 with the score for AD with *APOE* is therefore the expected one and, because it is absent without *APOE*, is consistent with a concentration of the *APOE* signal in this module; however, the proteomic score selects none of its 18 members.

M23 was consistently concentrated for ProRS-selected proteins across AD and PD. It is annotated to central nervous system development and the GABAergic synapse, and several members are established fluid markers of neurodegeneration, including NEFL, GFAP, SYT1 and NPTXR.[34–36] AD and PD were associated with this module through different proteins. AD selected BCAN, GFAP, NPTXR, SYT1 and VGF, PD selected F3, LRTM2, NCAN, PTPRN2, SCG2 and SLITRK1, and NEFL was the only member both scores use. This agreement between AD and PD presents evidence that module-level analysis may capture shared biology that individual proteins cannot.

This study defined two protein sets, those selected by LASSO in the ProRS model and those significantly associated with PRS, and showed the two are largely disjoint. However, the L_1 penalty has no grouping effect.[37,38] Among correlated predictors LASSO retains a subset and shrinks the rest to zero; which member survives varies across resamples, so a protein both PRS-associated and predictive could be dropped for a correlated partner, deflating the overlap.[37,38] The elastic-net refit rules this out. When the ProRS was refit with elastic net, which keeps correlated proteins together instead of dropping some of them, the selected sets of proteins grew in size, but their overlap with the PRS-associated proteins did not rise relative to chance, suggesting that the small overlap observed should not be attributed to the mechanism of LASSO dropping correlated proteins. The two sets may therefore be largely different groups of proteins that carry complementary information, a possibility our data do not directly test. CR1, for example, is shared by both sets in AD, and external Mendelian randomization work has supported its causal role.[39,40]

4.1. *Limitations*

Finally, every estimate rests on very small polygenic signal once *APOE* is removed, where the scores explain under 1% of liability and the three remaining absorption intervals reach lower bounds of 0.12, 0.01 and 0.00. Larger discovery GWAS would resolve this, and phenotype

definition compounds the problem. The FTD model is uninterpretable at this GWAS size and was excluded rather than reported as null, and ALS cases were scored as a single trait without *C9orf72* or *SOD1* stratification, treating a heterogeneous liability as homogeneous. Genotype-stratified replication would separate these. Modules were built without plate adjustment, so some module covariance may be technical. Preservation in *All of Us*, which was processed independently, argues against a cohort-specific artifact, but shared Olink platform structure cannot be excluded, particularly for M2, which has no enriched annotation. Adjusting for age and age squared does not let the relationship between the PRS and the outcome change with age, so a single baseline PRS from a case-control GWAS may distort absorption if genetic effects vary with age, and age-resolved summary statistics were not available. Prevalent cases enter every ProRS training fold, so medication such as cholinesterase inhibitors or levodopa could contribute to the score. They are a minority of cases, and only 5 prevalent AD cases were on a cholinesterase inhibitor at baseline, but we did not examine treatments in PD or ALS or refit without prevalent cases. The number of disease cases in *All of Us* is small, so the disease-specific absorption estimates cannot be interpreted as strongly replicated. The minimal overlap between PRS-associated and ProRS-selected proteins may also reflect a comparison of marginal univariate regressions with conditional LASSO selection rather than distinct biology. Cross-cohort module concordance (ARI = 0.582) should be read in light of k-nearest-neighbor imputation within a European subset in UK Biobank versus mean imputation across a multi-ancestry sample in *All of Us*, and fully harmonized, ancestry-stratified comparisons await larger *All of Us* releases. All PRS-protein and PRS-module relationships are associations in observational data and do not establish that genetic risk acts through these proteins.

Acknowledgments

This work was supported by the National Institutes of Health [R01 AG071470] and [U01 AG068057].

Supplementary Data

The supplementary files are available at https://github.com/czbw/Absorption-and-Coexpression-Modules-Show-Where-PRS-and-ProRS-Diverge-in-Neurodegenerative-Diseases

References

1. B. B. Sun *et al.*, Plasma proteomic associations with genetics and health in the UK Biobank, *Nature* **622**, 329 (2023).
2. E. Assarsson *et al.*, Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability, *PLoS ONE* **9**, p. e95192 (2014).
3. J. Woerner *et al.*, Large-scale evaluation of proteomic and polygenic risk scores reveals complementary contributions to incident disease prediction, *medRxiv* (2025), Preprint.
4. A. Shvetcov *et al.*, APOE ε4 carriers share immune-related proteomic changes across neurodegenerative diseases, *Nature Medicine* **31**, 2590 (2025).
5. K. A. Walker *et al.*, Proteomic analysis of APOEε4 carriers implicates lipid metabolism, complement and lymphocyte signaling in cognitive resilience, *Molecular Neurodegeneration* **19**, p. 81 (2024).

6. N. Amin *et al.*, Interplay between age, APOE $\varepsilon 4$ and the metabolome in plasma and brain in Alzheimer's disease, *Translational Psychiatry* **15**, p. 460 (2025).
7. P. Langfelder and S. Horvath, WGCNA: an R package for weighted correlation network analysis, *BMC Bioinformatics* **9**, p. 559 (2008).
8. E. C. B. Johnson *et al.*, Large-scale deep multi-layer analysis of Alzheimer's disease brain reveals strong proteomic disease-related changes not observed at the RNA level, *Nature Neuroscience* **25**, 213 (2022).
9. E. B. Dammer *et al.*, Multi-platform proteomic analysis of Alzheimer's disease cerebrospinal fluid and plasma reveals network biomarkers associated with proteostasis and the matrisome, *Alzheimer's Research & Therapy* **14**, p. 174 (2022).
10. R. Saloner *et al.*, Large-scale network analysis of the cerebrospinal fluid proteome identifies molecular signatures of frontotemporal lobar degeneration, *Nature Aging* **5**, 1143 (2025).
11. V. Emilsson *et al.*, Co-regulatory networks of human serum proteins link genetics to disease, *Science* **361**, 769 (2018).
12. C. Bycroft *et al.*, The UK Biobank resource with deep phenotyping and genomic data, *Nature* **562**, 203 (2018).
13. C. Sudlow *et al.*, UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age, *PLoS Medicine* **12**, p. e1001779 (2015).
14. The All of Us Research Program Investigators, The "All of Us" Research Program, *New England Journal of Medicine* **381**, 668 (2019).
15. The All of Us Research Program Genomics Investigators, Genomic data in the All of Us Research Program, *Nature* **627**, 340 (2024).
16. B. W. Kunkle *et al.*, Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing, *Nature Genetics* **51**, 414 (2019).
17. M. I. Kurki *et al.*, FinnGen provides genetic insights from a well-phenotyped isolated population, *Nature* **613**, 508 (2023).
18. W. van Rheenen *et al.*, Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology, *Nature Genetics* **53**, 1636 (2021).
19. R. Ferrari *et al.*, Frontotemporal dementia and its subtypes: a genome-wide association study, *The Lancet Neurology* **13**, 686 (2014).
20. T. Ge *et al.*, Polygenic prediction via Bayesian regression and continuous shrinkage priors, *Nature Communications* **10**, p. 1776 (2019).
21. C. C. Chang *et al.*, Second-generation PLINK: rising to the challenge of larger and richer datasets, *GigaScience* **4**, p. 7 (2015).
22. G. Leonenko *et al.*, Identifying individuals with high risk of Alzheimer's disease using polygenic risk scores, *Nature Communications* **12**, p. 4506 (2021).
23. B. K. Bulik-Sullivan *et al.*, LD Score regression distinguishes confounding from polygenicity in genome-wide association studies, *Nature Genetics* **47**, 291 (2015).
24. J. Friedman, T. Hastie and R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software* **33**, 1 (2010).
25. N. J. D. Nagelkerke, A note on a general definition of the coefficient of determination, *Biometrika* **78**, 691 (1991).
26. A. Gelman and H. Stern, The difference between "significant" and "not significant" is not itself statistically significant, *The American Statistician* **60**, 328 (2006).
27. E. R. DeLong, D. M. DeLong and D. L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, *Biometrics* **44**, 837 (1988).
28. M. C. Zody *et al.*, Evolutionary toggling of the MAPT 17q21.31 inversion region, *Nature Genetics*

40, 1076 (2008).
29. J. Trowsdale and J. C. Knight, Major histocompatibility complex genomics and human disease, *Annual Review of Genomics and Human Genetics* **14**, 301 (2013).
30. P. Langfelder *et al.*, Is my network module preserved and reproducible?, *PLoS Computational Biology* **7**, p. e1001057 (2011).
31. M. A. Nalls *et al.*, Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies, *The Lancet Neurology* **18**, 1091 (2019).
32. M. Uhlén *et al.*, Tissue-based map of the human proteome, *Science* **347**, p. 1260419 (2015).
33. S. Wang *et al.*, Cellular senescence induced by cholesterol accumulation is mediated by lysosomal ABCA1 in APOE4 and AD, *Molecular Neurodegeneration* **20**, p. 15 (2025).
34. M. Khalil, C. E. Teunissen, M. Otto *et al.*, Neurofilaments as biomarkers in neurological disorders, *Nature Reviews Neurology* **14**, 577 (2018).
35. L. Heimfarth, F. R. S. Passos, B. S. Monteiro *et al.*, Serum glial fibrillary acidic protein is a body fluid biomarker, a valuable prognostic for neurological disease, *International Immunopharmacology* **107**, p. 108624 (2022).
36. E. Camporesi, J. Nilsson, A. Brinkmalm *et al.*, Fluid biomarkers for synaptic dysfunction and loss, *Biomarker Insights* **15**, p. 1177271920950319 (2020).
37. R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 267 (1996).
38. H. Zou and T. Hastie, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301 (2005).
39. L. Belbasis *et al.*, Mendelian randomization identifies proteins involved in neurodegenerative diseases, *Brain* **148**, 2412 (2025).
40. H. Zhan *et al.*, Biomarker identification for Alzheimer's disease through integration of comprehensive Mendelian randomization and proteomics data, *Journal of Translational Medicine* **23**, p. 278 (2025).

COMPASS: Component-Wise Inference of Shared and Gene-Specific Perturbation Response

Huan Liang[1] and Rohit Singh[1,2,†]
[1]*Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA*
[2]*Department of Cell Biology, Duke University, Durham, NC, USA*
[†]*E-mail: rohit.singh@duke.edu*

Predicting how a genetic perturbation reshapes a cell's transcriptome is a central goal of computational biology. Previous studies report that the mean response across training perturbations rivals specialized models on standard accuracy metrics, even though it cannot distinguish which perturbation occurred. Across 2,270 CRISPRi perturbations measured in each of six cell lines, we show that this apparent paradox reflects a conserved organization of perturbation responses. Perturbations span a continuum from responses strongly aligned with the mean to more targeted responses that depart from it. Crucially, a perturbation's position along this continuum is conserved across cell lines (Kendall's $W = 0.59$) and predictable from STRING protein-interaction embeddings ($R^2 = 0.35$). We formalize this structure with COMPASS, an interpretable linear model that decomposes each response into shared and gene-specific components and estimates them separately. The shared-response component is modeled as a cell-line-wide response scaled by a perturbation-specific coefficient. This coefficient is strongly conserved across cell lines. The residual gene-specific component—which is moderately conserved across cell lines—recovers pathway-level programs. COMPASS outperforms scGPT, CPA, GEARS, GenePert, and State in both response accuracy (de-biased Pearson delta 0.34 vs. ≤ 0.32) and perturbation discrimination (cosine PDS gain 0.23 vs. ≤ 0.08). These results recast perturbation prediction across cellular contexts as component-wise inference, with each component estimated from the evidence best suited to it.

Keywords: Perturb-seq; genetic perturbation prediction; mean baseline; gene embeddings; response decomposition; interpretable models.

1. Introduction

One of the long-standing puzzles in perturbation prediction is the strength of simple baselines. For gene knockdowns measured by Perturb-seq, multiple benchmarking efforts[1,2] report that the mean response across training perturbations in the same cell line consistently rivals or outperforms purpose-built models such as GEARS and CPA, single-cell foundation models such as scGPT, and architectures trained across diverse experimental contexts.[3–6] Systema further showed that much of the apparent performance of these models arises from a systematic expression shift shared across perturbations, rather than from recovery of the response specific

© 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.

to the perturbed gene.[7]

While these observations can be read as a failure of model architectures, they also admit structural explanations. The first is about metrics: the mean baseline is *accurate*, recovering much of the measured perturbation effect, but it is ineffective at *discriminating* one perturbation from another. While early studies prioritized accuracy, discrimination is necessary for a complete assessment. The second is about biology: distinct knockdowns may converge on a common response direction because they perturb overlapping cellular programs. Such a shared response would explain the mean baseline's high accuracy and provide a scaffold for isolating perturbation-specific effects.

If such shared-response structure exists, individual perturbations should differ systematically in how strongly they engage the shared response. Knockdowns of core growth machinery are natural candidates for broad, stereotyped responses in which large numbers of genes shift together. Perturbations of specific regulatory machinery may instead produce responses that depart from the shared direction. To test this hypothesis, we characterize each perturbation p in cell line c using its response magnitude $m_{cp} \in \mathbb{R}$ and its alignment $a_{cp} \in \mathbb{R}$ with the cell line's mean perturbation direction. We explored how m and a are distributed and whether this organization is conserved across cell lines.

Across 2,270 perturbations shared by six cell lines from three independent CRISPRi studies,[8–10] we find that magnitude and alignment trace a similar continuum in every cell line. Strikingly, the same perturbations tend to occupy similar positions across studies, and their representations built from prior knowledge, particularly STRING network embeddings, predict those positions. This conservation motivates the decomposition $z_{cp} = \beta_p u_c + g_p + \varepsilon_{cp}$, where $u_c \in \mathbb{R}^G$ is the cell line's shared-response direction, $\beta_p \in \mathbb{R}$ measures how strongly perturbation p engages that direction, $g_p \in \mathbb{R}^G$ is the gene-specific residual response, and $\varepsilon_{cp} \in \mathbb{R}^G$ is the deviation specific to perturbation p in cell line c. Across cell lines, we compute $\bar{\beta}_p$ and g_p by averaging the per-cell-line coefficients and residuals, respectively. This converts the empirical observation of a perturbation continuum into two quantities with different conservation properties: $\bar{\beta}_p$ is strongly conserved and transferable across cell lines, while g_p is less well conserved but still retains meaningful cross-cell structure. Both quantities also carry substantial independent biological information.

The two components can therefore be estimated from different evidence and interpreted independently. This motivates COMPASS, an interpretable linear model that estimates these components under within-cell-line and cross-cell-line regimes and combines them to predict unseen perturbation responses. Across six cell lines, COMPASS substantially outperforms direct perturbation-response models and approaches the accuracy of the strongest fixed-representation predictor, TabICL, while matching or exceeding its cross-cell discrimination.

2. Data and Exploratory Analysis

2.1. *Data and evaluation contexts*

We analyzed six pooled CRISPRi Perturb-seq datasets from three independent studies: K562 and RPE1 from Replogle *et al.*, HepG2 and Jurkat from Nadig *et al.*, and HCT116 and HEK293T from X-Atlas/Orion.[8–10] After applying each study's guide-assignment and cell-

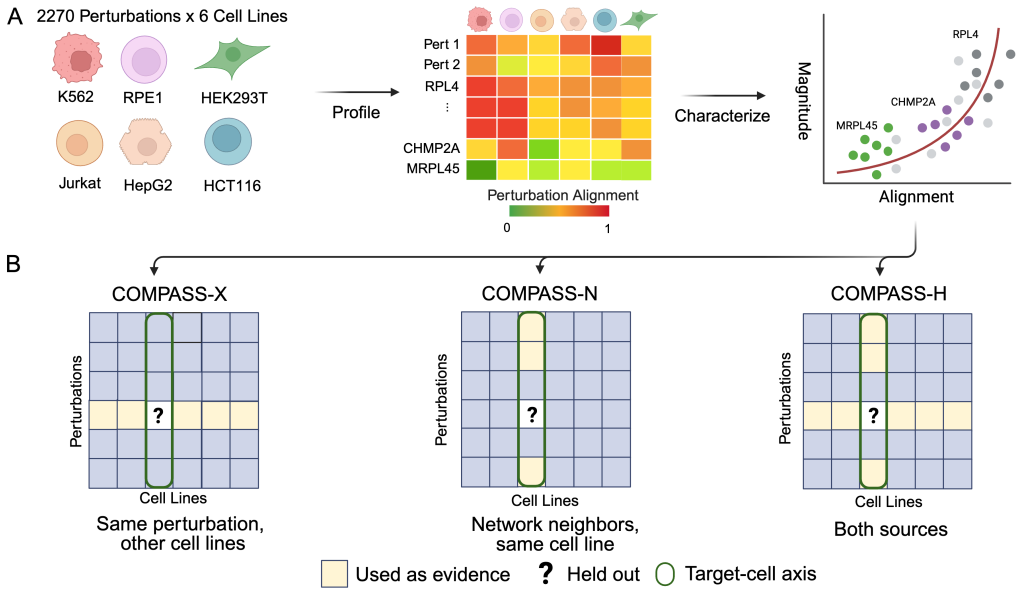


Fig. 1. **Study overview. (A)** We profiled 2,270 CRISPRi perturbations shared across six cell lines and characterized each perturbation by its response magnitude and its alignment with the cell-line average response. **(B)** COMPASS predicts unseen perturbations under three evidence regimes: COMPASS-X uses the same perturbation in other cell lines, COMPASS-N uses network-neighbor perturbations in the target cell line, and COMPASS-H combines both.

quality filters, we intersected the targets measured in all six lines to obtain 2,270 shared perturbations, predominantly essential genes from the Cancer Dependency Map project.[11,12] The shared perturbation set therefore supports matched cross-cell analysis but is enriched for genes with strong fitness effects, a limitation we return to in the Discussion. Counts were library-size normalized to 10,000 per cell and log-transformed as $\log(1 + x)$.

Predictions were scored on panels of 2,000 highly variable genes. We used two kinds of panels: cell-line-specific panels, selected independently for each line, and dataset-specific panels, shared across cell lines within a study group. Because the Replogle/Nadig (four cell lines) and Orion (two lines) dataset-specific panels overlap in only 282 readout genes, we kept our analysis dataset-specific; cross-cell quantities were estimated only within these groups. We evaluated predictions in two settings that correspond to different experimental situations. In the *within-cell-line* setting, a model observes a subset of perturbations in the query cell line and predicts held-out perturbations in the same line. This corresponds to extending a partially completed screen. In the *cross-cell-line* setting, the query perturbation has been measured in other cell lines but is held out from the target line. This corresponds to transferring knowledge from existing screens to a new cellular context. Extensive data processing details can be found in Supplementary Sections 6.1 and 6.2.

2.2. *Evaluation metrics: accuracy and discrimination*

A perturbation predictor must do two separable things: recover what the response looks like (*accuracy*) and distinguish one perturbation's response from another (*discrimination*). Measuring them requires distinct metrics.

For perturbation p, cell line c, and readout gene $g \in \mathcal{G}$ with $|\mathcal{G}| = G$, we define the pseudobulk perturbation effect

$$z_{cp,g} = \bar{x}_{cp,g} - \bar{x}_{c0,g}, \tag{1}$$

where $\bar{x}_{c0,g}$ is the mean among non-targeting controls. A predictive model returns a vector $\widehat{z}_{cp} \in \mathbb{R}^G$ for each held-out perturbation.

We measure accuracy with the **de-biased Pearson delta**, which quantifies agreement across readout genes between the predicted and measured perturbation-effect vectors. Following Nicol *et al.*, we form the predicted and observed effects using disjoint halves of the non-targeting controls, preventing their shared control reference from inflating the correlation.[13] We measure discrimination with the **cosine perturbation discrimination score (PDS) gain**: PDS tests whether each prediction is closer to its matched measured response than to those of other perturbations. We use cosine similarity, as ℓ_1/ℓ_2 is affected by magnitude of predicted vectors.[14] A score of 0.50 corresponds to chance, and we report the gain above chance.

We also computed DE overlap@N, which measures overlap between the top-ranked differential genes in the predicted and measured responses, with N set by the number of measured genes passing FDR < 0.05. Because DE overlap ranks methods nearly identically to the de-biased Pearson delta across all methods and cell lines (Supplementary Table A1), we report it in the supplement as a confirmatory accuracy measure. All formal metric definitions appear in Section 6.3.

2.3. *The mean baseline captures the shared response but cannot discriminate perturbations*

The training mean recovers much of the measured response but cannot distinguish which perturbation produced it. Across five random 80/20 splits per cell line, assigning every held-out perturbation the average training response matches or exceeds GEARS and scGPT on de-biased Pearson delta in every cell line, reaching 0.09–0.55 compared with 0.03–0.43 for GEARS (Supplementary Table A2). By construction, the same prediction is returned for every test perturbation, so its cosine PDS remains at chance (0.50). GEARS and scGPT also remain near chance, at gains of 0.01 and 0.00, despite producing perturbation-conditioned outputs. Their response-level agreement therefore comes primarily from structure shared across perturbations. All models and the training mean are described in Section 6.4.

GenePert outperforms the other baselines through a structure consistent with the decomposition we formalize in Section 3. The method fits a regularized regression from a gene embedding to the perturbation effect within a cell line and achieves the strongest response correlation among the within-cell baselines (0.32 on average, compared with 0.28 for the training mean). Its fitted intercept is the training mean, so each prediction is the shared response plus a shrunken gene-dependent correction. This gene-dependent correction raises its mean PDS gain to 0.08, although its discrimination remains modest relative to its response correlation.

Together, these results separate two prediction tasks: the mean's high accuracy shows that many perturbations share a dominant expression pattern, while its zero PDS gain shows that the perturbation-specific signal resides in the variation around the shared response. We next

characterize each response by both its overall strength and its agreement with the shared direction.

2.4. *Magnitude and alignment characterize perturbation responses*

For each perturbation p in cell line c, we computed a signed Wilcoxon rank-sum statistic for every readout gene by comparing perturbed cells with non-targeting controls from the same experiment. The resulting profile $w_{cp} \in \mathbb{R}^G$ summarizes the direction and consistency of the response. With $\bar{w}_c = |\mathcal{P}|^{-1} \sum_{p \in \mathcal{P}} w_{cp}$ denoting the mean profile over the 2,270 shared perturbations, we define

$$m_{cp} = \|w_{cp}\|_2 , \qquad a_{cp} = \mathrm{corr}\,(w_{cp}, \bar{w}_c) . \tag{2}$$

Magnitude measures the overall response strength, while alignment measures how closely the response follows the cell line's shared perturbation direction.

Magnitude and alignment are positively correlated in all six cell lines (Fig. 2), with Spearman correlations ranging from $\rho = 0.41$ in HEK293T to 0.82 in HepG2. As elaborated in Section 2.6, high-magnitude, high-alignment perturbations elicit broad, similar responses, suggesting extensive disruption of the cellular system. At the other end are perturbations that elicit specific responses that are poorly aligned with the mean.

This correlation is partly geometric: high-magnitude responses contribute disproportionately to $\bar{w}_c$ and therefore tend to align with the direction they collectively define, whereas the alignment of weak responses is estimated less reliably. Recomputing the mean without each perturbation, or using the median response, leaves the correlation essentially unchanged (Supplementary Table A3). Because magnitude and alignment are naturally coupled, they define a continuum rather than independent axes, motivating a single coordinate along the dominant trajectory (Section 2.5). This structure is robust to the choice of readout panel: recomputing the analysis with 1,000- and 5,000-highly variable gene panels preserves the continuum, cross-cell concordance, and response regimes (Supplementary Fig. A1 and Table A4), and to raising the minimum number of cells per perturbation (Supplementary Table A6).

2.5. *A one-dimensional coordinate orders the magnitude–alignment continuum*

To reduce this continuum to a single quantity, we standardized m_{cp} and a_{cp} within each cell line and computed a one-dimensional Isomap coordinate, $\boldsymbol{s_{cp}}$, oriented so that larger values of s_{cp} correspond to the high-magnitude, high-alignment arm (Section 6.5). This ordering is conserved across cell lines and studies. Across the fifteen cell-line pairs, Pearson correlations average $r = 0.50$ for alignment, $r = 0.38$ for magnitude, and $r = 0.50$ for s_{cp}, and every pairwise comparison is positive; Kendall's coefficients of concordance across all six lines are $W_a = 0.587$, $W_m = 0.533$, and $W_s = 0.588$ for alignment, magnitude, and s_{cp}, respectively (Fig. 3). These values are robust to raising the minimum number of cells per perturbation (Supplementary Table A6). Perturbations with high (low) s_{cp} in one cell line therefore tend to rank highly (lowly) in the others. This conservation persists across study-specific processing pipelines and environmental factors: the strongest concordance in s_{cp} occurs for the cross-study RPE1–HepG2

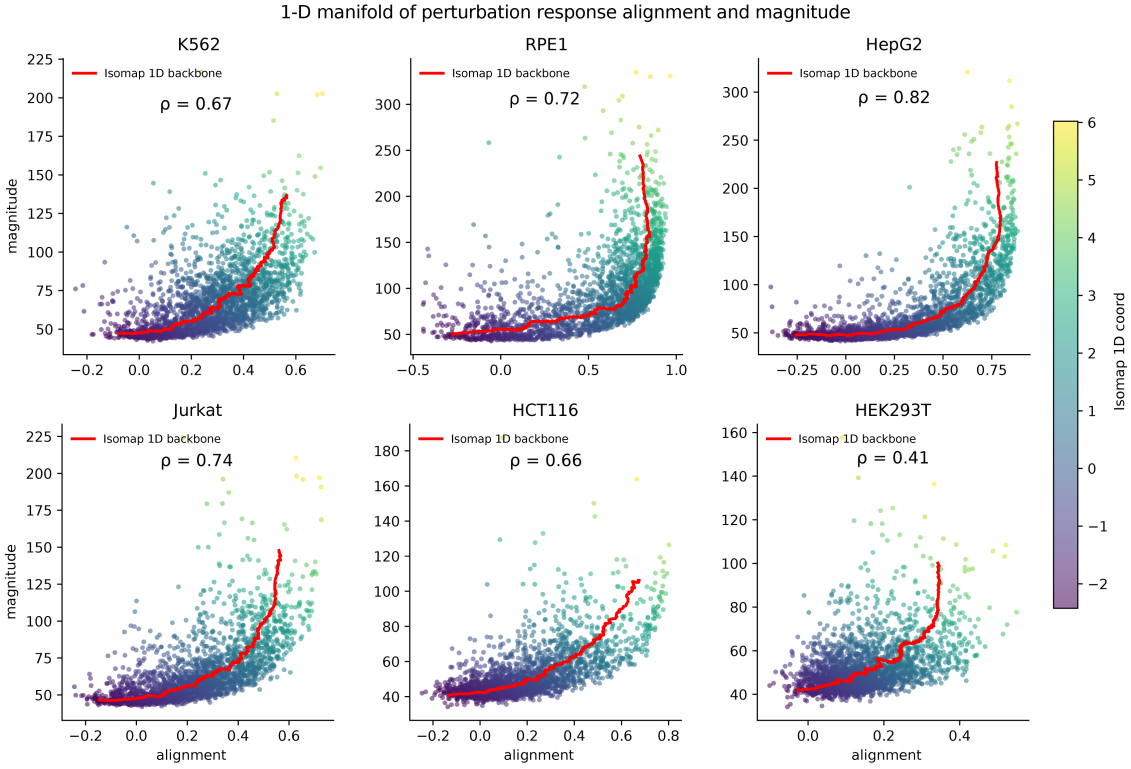


Fig. 2. **Magnitude and alignment trace a common response organization across six cell lines.** Response magnitude m_{cp} against alignment a_{cp} for the 2,270 shared perturbations. Cell-specific Spearman correlations between magnitude and alignment are marked on each plot. Points are coloured by the coordinate s_{cp} of Section 2.5.

pair ($r = 0.66$), and the next strongest, HepG2–HCT116 ($r = 0.62$), is also cross-study, while K562 and RPE1, the two Replogle lines, are among the least concordant ($r = 0.41$, eleventh of fifteen).

2.6. *The two ends of the continuum are biologically distinct*

The conserved ordering separates perturbations with distinct biological roles. We averaged the cell-specific coordinates to $\bar{s}(p) = |\mathcal{C}|^{-1} \sum_c s_{cp}$, ranked the 2,270 perturbations, and tested the top and bottom 30% for Gene Ontology overrepresentation (Supplementary Table A7). The high-magnitude, high-alignment end is dominated by core growth machinery, including cytoplasmic translation, rRNA processing, spliceosomal mRNA processing, RNA polymerase II preinitiation, and DNA replication. Representative genes include RPL4, RPS13, SF3B3, PRPF31, MCM2, MCM3, MCM6, GINS1, and GINS4. The low-$\bar{s}$ tail is enriched for regulatory and organelle-specific functions, including mitochondrial translation, chromatin remodeling, transcriptional coactivation, protein folding, and regulation of DNA-templated transcription, with genes such as MRPL45, MRPS7, KDM6A, BRD4, NCOA4, and PFDN1. The same terms remain significant when the tails are defined at 20%, 25%, or 30%; the full sensitivity analysis appears in Supplementary Table A8. These enrichments support the biological distinction underlying the continuum: core growth machinery is concentrated among strong, shared-

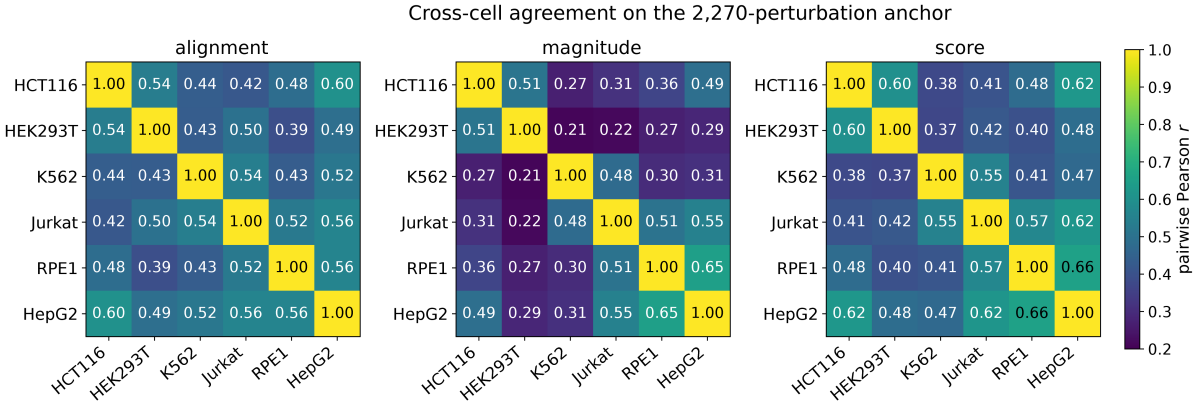


Fig. 3. **The response organization is conserved across cell lines and studies.** Pairwise cross-cell correlations of alignment, magnitude, and the coordinate s_{cp} across the 2,270 shared perturbations; all fifteen pairs are positive for each measurement.

response-dominated perturbations, whereas regulatory and organelle-specific functions are concentrated among less aligned responses.

2.7. *Gene representations predict conserved response position*

Because the same perturbations occupy similar positions across cell lines, we asked whether cell-agnostic prior knowledge about the targeted gene could predict that position. We trained a gradient-boosted regressor to predict the cross-cell means $\bar{m}(p)$, $\bar{a}(p)$, and $\bar{s}(p)$ from gene representations derived from protein-interaction networks, knowledge graphs, transcriptomic data, Gene Ontology annotations, protein sequences, and language models, evaluating every representation under identical perturbation-level five-fold splits, prediction targets, model class, and hyperparameters (Section 6.6).

STRING predicts a perturbation's position more accurately than the other representations (Supplementary Fig. A2). Consistent with its strong cross-cell agreement, alignment is the most predictable of the three quantities, with cross-validated $R^2 = 0.39$. Magnitude and $\bar{s}(p)$ are also well-predicted, albeit less strongly, at $R^2 = 0.24$ and 0.35, respectively. Graph-neural, knowledge-graph, Gene Ontology, and language-model representations also recover meaningful signal, while somewhat surprisingly, transcriptomic and protein-sequence representations perform less well.

3. Methods

3.1. *COMPASS: a shared-plus-specific decomposition of the perturbation response*

The response structure characterized above motivates a direct decomposition in expression space. For the pseudobulk effect vectors z_{cp} defined in Eq. (1), let $\bar{z}_c = |\mathcal{P}|^{-1} \sum_{p \in \mathcal{P}} z_{cp} \in \mathbb{R}^G$ denote the mean perturbation response in cell line c, and let $u_c = \bar{z}_c / \|\bar{z}_c\|_2$ be the unit vector along that mean. We decompose each response as

$$z_{cp} = \beta_{cp}\, u_c + r_{cp}, \qquad \beta_{cp} = \langle z_{cp}, u_c \rangle, \qquad r_{cp} = g_p + \varepsilon_{cp}, \tag{3}$$

Held-out cell line	K562	RPE1	HepG2	Jurkat	Mean
shared-response coefficient $\widehat{\beta}_{Cp} \to \beta_{Cp}$	0.61	0.70	0.71	0.62	**0.66**
gene-specific component $\widehat{g}_p \to r_{Cp}$	0.22	0.22	0.21	0.21	**0.22**

Table 1. **The shared-response coefficient transfers more strongly than the gene-specific response.** Each of the four Replogle/Nadig cell lines is held out in turn, and both components are estimated from the other three. The coefficient is a scalar per perturbation, compared by Pearson correlation across the 2,270 perturbations; the gene-specific response is a vector, compared by Pearson correlation across genes within each perturbation and averaged. A perturbation-shuffled null gives ≈ 0 for both.

Here $\beta_{cp} \in \mathbb{R}$ is the shared-response coefficient, computed by projecting z_{cp} onto u_c, and $\beta_{cp}u_c \in \mathbb{R}^G$ is the corresponding shared-response component. The vector $r_{cp} \in \mathbb{R}^G$ is the gene-specific response remaining after this component is removed; it is orthogonal to u_c by construction. Across cell lines with a common readout space, g_p denotes the conserved mean of this gene-specific response and ε_{cp} its cell-line-specific deviation. A single cell line cannot identify g_p and ε_{cp} separately, so within-cell estimators target r_{cp} directly. Here and below, *gene-specific* means specific to the perturbed gene p, rather than to an individual readout gene.

The shared-response coefficient β_{cp} is the expression-space counterpart of the empirical coordinate s_{cp}: both order perturbations by how strongly their responses follow the cell line's shared response. β_{cp} is slightly better conserved across cell lines than s_{cp} itself: per-cell-line estimated values correlate at mean pairwise Pearson 0.55 among the four Replogle/Nadig lines and 0.56 between the two Orion lines (Supplementary Table A9). The two orderings also agree closely: the Pearson correlation between the (Replogle/Nadig) cross-cell mean $\bar{\beta}_p = |\mathcal{C}|^{-1} \sum_c \beta_{cp}$ and $\bar{s}(p)$ is 0.88. Ranking by $\bar{\beta}_p$ yields similar gene function enrichments (Supplementary Table A10). Because $\bar{\beta}_p$ can be estimated from source cell lines and summarizes a perturbation's shared-response engagement in a single portable scalar, it can be used as an input in cross-cell prediction settings.

To measure how well these components transfer across cellular contexts, we performed leave-one-cell-line-out assessments on the four Replogle/Nadig lines, writing C for the held-out line. The estimate $\widehat{\beta}_{Cp}$, formed by averaging over the remaining three lines, predicts the held-out shared-response coefficient β_{Cp} with mean Pearson correlation 0.66, while the corresponding estimate $\widehat{g}_p$ matches the held-out gene-specific response r_{Cp} at mean Pearson 0.22 (Table 1). The shared-response coefficient thus transfers considerably more reliably than the gene-specific residual. With only one source line available, as for the Orion pair, transfer remains predictive but is clearly weaker than from three, 0.56 and 0.09 for the two components (Supplementary Table A11).

Together, these analyses identify two components with complementary conservation properties: the shared-response coefficient $\bar{\beta}_p$ is strongly conserved across cell lines and transfers reliably, while the gene-specific response, though more cell-line-specific, still retains meaningful cross-cell-line structure.

3.2. *Estimating COMPASS under different information regimes*

We now consider predicting z_{Cp} for an unmeasured perturbation p in a target cell line C. COMPASS assumes that the target shared-response direction u_C can be estimated from a calibration set of perturbations measured in C; the low-data analysis below shows that this estimate stabilizes within a few tens of perturbations (also see Discussion). The prediction task then reduces to estimating β_{Cp} and r_{Cp}. COMPASS provides three estimators for these quantities: a cross-cell estimator that carries both components over from other cell lines, a within-cell estimator that infers them from STRING neighbors of p measured in C, and a hybrid that combines both sources with component-specific weights (Supplementary Algorithms 1 to 3). In each case, the final prediction is assembled as

$$\widehat{z}_{Cp} = \widehat{\beta}_{Cp}\, u_C + \widehat{r}_{Cp}. \tag{4}$$

COMPASS-X: the perturbation is measured in other cell lines. When p is available in a set of source cell lines $\mathcal{S}_p = \{d \neq C : p \in \mathcal{P}_d\}$, we estimate its shared-response projection and conserved gene-specific component by

$$\widehat{\beta}^{\mathrm{X}}_{Cp} = \frac{1}{|\mathcal{S}_p|} \sum_{d \in \mathcal{S}_p} \beta_{dp}, \qquad \widehat{g}^{\mathrm{X}}_{p} = \frac{1}{|\mathcal{S}_p|} \sum_{d \in \mathcal{S}_p} r_{dp}. \tag{5}$$

Since the target-specific deviation ε_{Cp} is unobserved, we approximate $\widehat{r}^{\mathrm{X}}_{Cp} = \widehat{g}^{\mathrm{X}}_{p}$. We emphasize that this differs from simply averaging the raw source response vectors z_{dp} and transferring them directly. COMPASS-X first decomposes each source-line response into its shared-response projection and gene-specific component, then transfers these separately, applying the averaged coefficient $\widehat{\beta}^{\mathrm{X}}_{Cp}$ to the target-cell direction u_C rather than retaining the source-cell directions. Critically, this decomposed transfer outperforms direct source averaging, as we show below.

COMPASS-N: related perturbations are measured in the target cell. When the target line contains a sufficiently large perturbation set, or when p has not been screened elsewhere, we estimate both components from STRING neighbors measured in C. With $\mathcal{N}_k(p)$ denoting the k nearest training perturbations,

$$\widehat{\beta}^{\mathrm{N}}_{Cp} = \frac{1}{k_b} \sum_{q \in \mathcal{N}_{k_b}(p)} \beta_{Cq}, \qquad \widehat{r}^{\mathrm{N}}_{Cp} = \frac{1}{k_r} \sum_{q \in \mathcal{N}_{k_r}(p)} r_{Cq}. \tag{6}$$

The neighborhood sizes k_b and k_r can be selected separately on validation data (Section 6.7); we recommend default values of $k_b = 5$ and $k_r = 10$.

COMPASS-H: both sources of information are available. When both sources of information are available, we combine the β and r estimates from COMPASS-X and COMPASS-N. Because the shared-response coefficient and gene-specific response have different transfer properties, COMPASS-H assigns them separate mixing weights:

$$\begin{aligned} \widehat{\beta}^{\mathrm{H}}_{Cp} &= \lambda_b \widehat{\beta}^{\mathrm{X}}_{Cp} + (1-\lambda_b)\widehat{\beta}^{\mathrm{N}}_{Cp} \\ \widehat{r}^{\mathrm{H}}_{Cp} &= \lambda_r \widehat{r}^{\mathrm{X}}_{Cp} + (1-\lambda_r)\widehat{r}^{\mathrm{N}}_{Cp} \end{aligned} \tag{7}$$

We set $\lambda_b = 1$ because the cross-cell estimate comes from the same perturbation in other cell lines and better preserves its magnitude along the shared response direction; a sensitivity

analysis confirms that this choice gives the strongest discrimination (Supplementary Fig. A3). We select λ_r jointly with k_r on the validation split, allowing the gene-specific estimate to draw on target-cell neighbors when the conserved component alone is insufficient. As defaults, we recommend $\lambda_b = 1, \lambda_r = 0.5$.

		Accuracy: De-biased Pearson delta							Discrimination: Cosine PDS gain						
		K562	RPE1	HepG2	Jurkat	HCT116	HEK293T	Mean	K562	RPE1	HepG2	Jurkat	HCT116	HEK293T	Mean
Cross-cell	Source average	0.28	0.37	0.35	0.26	0.10	0.08	0.24	**0.25**	**0.32**	**0.24**	0.23	0.18	0.16	**0.23**
	STATE	0.14	0.44	0.31	0.17	0.10	0.08	0.21	0.18	0.04	0.04	0.11	0.07	0.06	0.08
	TabICL	**0.41**	**0.63**	**0.46**	**0.41**	**0.28**	**0.20**	**0.40**	0.24	0.18	0.20	**0.27**	0.15	0.16	0.20
	COMPASS-X	0.33	0.43	0.40	0.32	0.12	0.12	0.29	**0.25**	0.30	0.22	0.24	0.18	**0.19**	**0.23**
	COMPASS-H	0.36	0.51	0.43	0.35	0.22	0.16	0.34	0.24	0.28	0.21	0.24	**0.23**	**0.19**	**0.23**
Within-cell	Training mean	0.26	0.55	0.36	0.24	0.19	0.09	0.28	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	CPA	0.05	0.06	0.06	0.07	0.01	0.02	0.05	0.01	0.00	0.01	0.02	0.01	0.00	0.01
	GEARS	0.13	0.43	0.25	0.13	0.06	0.03	0.17	0.00	0.01	0.03	0.03	−0.01	0.02	0.01
	scGPT	0.18	0.55	0.34	0.19	0.16	0.06	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	GenePert	0.32	0.57	0.38	0.28	0.23	0.13	0.32	0.11	0.05	0.06	0.09	0.08	0.07	0.08
	TabICL	**0.39**	**0.61**	**0.43**	**0.35**	**0.28**	**0.19**	**0.37**	**0.21**	**0.14**	**0.15**	**0.19**	**0.16**	**0.15**	**0.16**
	COMPASS-N	0.34	0.57	0.39	0.30	0.25	0.16	0.34	0.19	0.13	0.13	0.18	0.15	0.12	0.15

Table 2. **Cross-context prediction across the six cell lines (random splits).** De-biased Pearson delta and cosine PDS gain, averaged over five 80/20 partitions per cell line. TabICL is run with all six prior-knowledge modalities as its authors do. **Bold** marks the best value in each column and metric, underline the second best. Variability across partitions and paired significance tests against the COMPASS variants are given in Supplementary Table A12.

4. Experiments

4.1. *COMPASS is competitive with the strongest baseline using a linear, interpretable model*

To benchmark the predictive performance of our approach, we compare the COMPASS suite against three families of existing models (Table 2): *structural baselines* (training mean, source average), *perturbation-conditioned deep learning models* (CPA, GEARS, scGPT, STATE), and *embedding-based regressors* (GenePert[15] and TabICL[16,17]). Our adaptation of TabICL follows Palla *et al.*'s work. Implementation details for each are given in Section 6.4. Compared to these, the COMPASS suite exploits the response structure characterized in Section 3.1: each prediction is a shared-response component plus a gene-specific residual, with both terms estimated from within-cell (COMPASS-N), cross-cell (COMPASS-X) or both (COMPASS-H). All models are scored on the same test perturbations, and any tuned hyperparameters are selected on validation perturbations (Section 6.9).

The COMPASS suite, specifically the hybrid model COMPASS-H, matches or outperforms every non-TabICL model on both accuracy (de-biased Pearson delta) and discrimination (cosine PDS gain) (Table 2). COMPASS-X uses the same inputs as the source average but decomposes them first, and is more accurate at comparable discrimination. This suggests that transferring information from other cell lines via our linear decomposition is more effective than using it directly. Adding within-cell information, i.e. COMPASS-H, further improves performance, while the source averaging baseline is unable to accommodate such additional information.

COMPASS substantially outperforms the direct perturbation-response deep learning models evaluated here. In within-cell contexts, COMPASS-N leads the deep learning approaches scGPT,

GEARS, and CPA. As shown in Section 2.3, several of these models also underperform simple averaging baselines on response accuracy.

GenePert and TabICL are regression-based frameworks that use one and six sets of gene embeddings, respectively, to predict response. GenePert works in a within-cell setting and uses regression; TabICL works in a hybrid setting and uses pre-trained tabular foundation models. We argue they are both best thought of as powerful variants of the training mean or source averaging baseline, as they both learn the overall response and add gene-specific tilts. Interestingly, both methods are stronger on accuracy than discrimination, analogous to the training mean.

TabICL is overall more accurate than COMPASS variants but slightly less discriminative; GenePert is outperformed by COMPASS on both metrics. TabICL leads COMPASS-H on de-biased Pearson delta by 0.06 cross-cell (0.40 vs. 0.34) and COMPASS-N by 0.03 within-cell (0.37 vs. 0.34). On cosine PDS gain, however, COMPASS-X and COMPASS-H match or exceed TabICL cross-cell (0.23 vs. 0.20), and COMPASS-N is within 0.01 within-cell (0.15 vs. 0.16). COMPASS is therefore competitive with a much heavier model, using STRING embeddings only as a distance metric to select neighbors rather than as regression inputs, and its decomposition additionally produces components that are interpretable in isolation (Section 4.2). For both COMPASS and TabICL, accuracy and discrimination decrease as test perturbations become more isolated from training perturbations, based on STRING embedding space (Supplementary Fig. A4).

COMPASS variants reflect the tension between accuracy and discrimination COMPASS-X is the stronger discriminator (PDS gain 0.23 vs. 0.15) because its perturbation-specific information comes from measurements of the same perturbation in other cell lines. COMPASS-N is the stronger accuracy model (de-biased Pearson 0.34 vs. 0.29) because averaging over STRING neighbors in the target line recovers the local shape of the response, but this neighbor-averaging cannot effectively separate a perturbation from its neighbors. COMPASS-N is therefore the appropriate estimator when the target-cell perturbation set is large but the target perturbation has not been measured elsewhere, while COMPASS-X is the appropriate estimator when the same perturbation has been measured in a related cell line but the target perturbation set is sparse (Supplementary Fig. A5). COMPASS-H estimates each component from the source that recovers it better, matching the stronger of the two on both axes at once.

For a new cell line, COMPASS is more data-efficient than TabICL When investigating a new cell line or tissue, a practical question is how many perturbations are needed before data from other cell lines can be effectively used. To assess this, we included all available cross-cell perturbations and increased the number of within-cell training perturbations from 10 to 130. Across ten seeds in each of the six cell lines (Fig. 4), COMPASS-X converges quickly: its de-biased Pearson delta rises from 0.253 to 0.288 across the range and its PDS gain stays flat near 0.24. This is because the target cell only enters through the shared response u_C, which converges within a few tens of perturbations. In contrast, TabICL starts well below COMPASS-X, and only overtakes it on accuracy at approximately 80 profiled perturbations; it does not outperform on discrimination even with maximal data in this scenario.

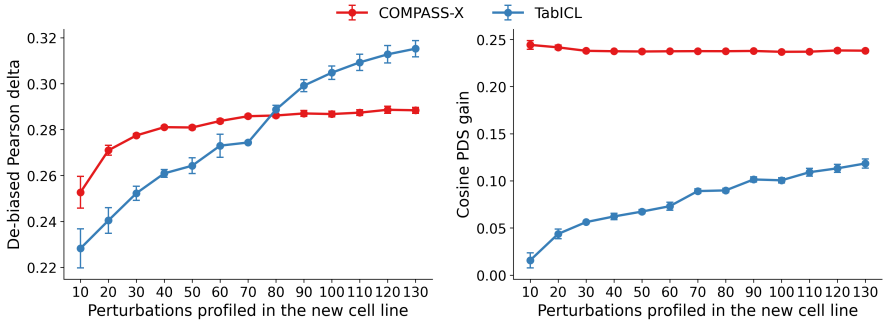


Fig. 4. **Cross-context transfer versus in-context learning as the profiling budget grows.** De-biased Pearson delta (left) and cosine PDS gain (right) against the number of perturbations profiled in a new cell line, with the remainder held out for test.

4.2. *The gene-specific response g_p encodes pathway-level biology*

The gene-specific component contains coherent biological programs rather than unstructured variation left after removing the shared response (Section 6.10). Averaging r_{cp} across the four Replogle/Nadig cell lines gives a 2,270 × 2,000 matrix of conserved components g_p. Clustering their normalized directions groups perturbations targeting common biological machinery, including mitochondrial translation, ribosome biogenesis, rRNA processing, mitosis, DNA replication, RNA splicing, and proteasome function (Supplementary Table A13). All fifteen clusters in the highest-coefficient tertile are enriched for perturbations targeting a coherent pathway or complex (Supplementary Table A15). The organization persists among perturbations that engage the shared response weakly, where clusters recover mitochondrial translation, N-glycosylation, and mitochondrial organization programs (Supplementary Table A14).

Individual gene-specific components also recover expected downstream responses. The TFAM component is dominated by reduced expression of mitochondrially encoded oxidative-phosphorylation transcripts. Proteasome-subunit perturbations, including PSMB5, repress cytoplasmic translation while inducing vacuolar and protein-stress programs; the PSMB5 component alone is enriched for unfolded-protein ($q = 8 \times 10^{-5}$) and heat-shock ($q = 3 \times 10^{-5}$) responses among its fifteen most induced genes. None of these analyses uses pathway labels during fitting. The gene-specific component therefore captures both relationships among perturbed genes and the downstream programs that distinguish their responses.

5. Discussion

The key conceptual advancement of our work is to efficiently disentangle gene-specific effects from shared responses in single-cell perturbation data. This separation both clarifies which pathways a perturbation engages and improves prediction of unmeasured perturbations. Across 2,270 CRISPRi knockdowns, response magnitude and alignment define a continuum from shared-response-dominated perturbations to those eliciting more gene-specific responses. The ordering of perturbations in the continuum is conserved across cell lines and studies. COMPASS formalizes this organization as $z_{cp} = \beta_{cp} u_c + r_{cp}$, with $r_{cp} = g_p + \varepsilon_{cp}$ across cell lines sharing a readout space. Averaging the shared-response coefficient for the same perturbation across source cell lines yields the portable perturbation-level quantity $\bar{\beta}_p$, which predicts how strongly that

perturbation will engage the shared response in a new cellular context. The shared-response coefficient transfers substantially more strongly than the gene-specific response g_p, which retains coherent pathway-level structure. Concurrent work by Molina and Zhang[18] similarly decomposes shared and perturbation-specific response structure, but our work additionally shows that the perturbation-specific coefficient β_{cp} on the shared response is itself portable across cell lines, enabling COMPASS to reconstruct the shared response as $\bar{\beta}_p u_c$, contributing to its efficiency and interpretability (Section 6.12).

The decomposition also explains why accuracy and discrimination diverge in recent perturbation-prediction benchmarks.[1,2,7] Accuracy metrics reward recovery of the response shared across many perturbations, whereas discrimination rewards capturing what makes each perturbation's response distinct. The COMPASS variants expose a corresponding tension between sources of evidence. In their corresponding regimes, COMPASS-X and COMPASS-N outperform scGPT, CPA, GEARS, GenePert, and STATE on both metrics. COMPASS-H combines the strengths of COMPASS-X and COMPASS-N and approaches TabICL's higher accuracy, exceeds its cross-cell discrimination, and is also more data and parameter-efficient than TabICL.

Our work suggests several directions for future work. As others have noted, perturbation predictors should be evaluated along two distinct axes, response accuracy and perturbation discrimination. Our results further suggest stratifying both metrics along the magnitude–alignment continuum, and model development should similarly treat the two components separately: cross-cell measurements and cell-agnostic gene knowledge may be particularly effective for estimating $\bar{\beta}_p$, whereas target-cell data may be needed to refine r_{cp}. The continuum also suggests a strategy for experimental design. A relatively small set of high β_p perturbations can establish the shared-response direction in a new cell type, allowing more of the profiling budget to be directed toward low-alignment perturbations whose responses carry distinctive and less predictable biology.

There are limitations to this work, including the modest number of cells per perturbation which results in uncertain estimation of the components (Supplementary Table A5). The current perturbation set is enriched for essential genes, so testing functionally diverse perturbations will establish whether the same shared-plus-specific organization persists beyond stress-dominated responses. Extending the analysis to primary cells and additional cellular contexts will similarly test how well $\bar{\beta}_p$ and g_p transfer beyond immortalized and cancer-derived cell lines and to out-of-distribution pathways. Finally, COMPASS currently requires perturbations in the target cell line to estimate u_c; learning this direction from unperturbed cells, for example through single-cell foundation model embeddings, could enable prediction in entirely new cellular contexts without a perturbation calibration set.

Code, data and supplemental materials availability Code, processed response matrices, split definitions, and scripts for reproducing the figures and tables is released at `https://github.com/rohitsinghlab/compass`. Source Perturb-seq datasets remain available from their respective studies. Supplemental sections, figures and tables can be found at `https://singhlab.net/compass/`.

References

1. 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).
2. Z. Wei, Y. Wang, Q. Liu *et al.*, Benchmarking algorithms for generalizable single-cell perturbation response prediction, *Nature Methods* (2025).
3. M. Lotfollahi, A. Susmelj, C. De Donno, Y. Ji, I. L. Ibarra, F. A. Wolf, N. Yakubova and F. J. Theis, Predicting cellular responses to complex perturbations in high-throughput screens, *Molecular Systems Biology* **19**, p. e11517 (2023).
4. Y. Roohani, K. Huang and J. Leskovec, Predicting transcriptional outcomes of novel multigene perturbations with GEARS, *Nature Biotechnology* **42**, 927 (2024).
5. H. Cui, C. Wang, H. Maan, K. Pang, F. Luo and B. Wang, scGPT: toward building a foundation model for single-cell multi-omics using generative AI, *Nature Methods* **21**, 1470 (2024).
6. A. K. Adduri, D. Gautam *et al.*, Predicting cellular responses to perturbation across diverse contexts with State, *bioRxiv* (2025).
7. R. Viñas Torné *et al.*, Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation, *Nature Biotechnology* (2025).
8. 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 *et al.*, Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq, *Cell* **185**, 2559 (2022).
9. A. Nadig, J. M. Replogle, A. N. Pogson, M. Murthy, S. A. McCarroll, J. S. Weissman, E. B. Robinson and L. J. O'Connor, Transcriptome-wide analysis of differential expression in perturbation atlases, *Nature Genetics* **57**, 1228 (2025).
10. A. C. Huang, T.-H. S. Hsieh, J. Zhu, J. Michuda, A. Teng, S. Kim, E. M. Rumsey, S. K. Lam, I. Anigbogu, P. Wright *et al.*, X-atlas/Orion: Genome-wide Perturb-seq datasets via a scalable Fix-Cryopreserve platform for training dose-dependent biological foundation models, *bioRxiv* , 2025 (2025).
11. A. Tsherniak, F. Vazquez, P. G. Montgomery, B. A. Weir, G. Kryukov, G. S. Cowley, S. Gill, W. F. Harrington, S. Pantel, J. M. Krill-Burger *et al.*, Defining a cancer dependency map, *Cell* **170**, 564 (2017).
12. J. M. Dempster, J. Rossen, M. Kazachkova, J. Pan, G. Kugener, D. E. Root and A. Tsherniak, Extracting biological insights from the Project Achilles genome-scale CRISPR screens in cancer cell lines, *bioRxiv* , p. 720243 (2019).
13. P. B. Nicol *et al.*, Spurious correlation inflates performance in single-cell perturbation prediction, *bioRxiv* (2026).
14. Q. Liu, Q. Zhang, J. Du, S. Zhao and J. Wang, Effects of distance metrics and scaling on the perturbation discrimination score, *arXiv* (2025).
15. Y. Chen and J. Zou, GenePert: Leveraging GenePT embeddings for gene perturbation prediction, *bioRxiv* (2024).
16. J. Qu, D. HolzmÃžller, G. Varoquaux and M. L. Morvan, Tabiclv2: A better, faster, scalable, and open tabular foundation model, *arXiv preprint arXiv:2602.11139* (2026).
17. G. Palla, A. Hillsley, Y.-J. Kim and L. A. Royer, Tabular foundation models are competitive cellular perturbation predictors across biological scales, *bioRxiv* , 2026 (2026).
18. A. Molina and X. Zhang, Perturbation response decomposition enables biologically aligned generalization to unseen perturbations and cellular contexts, *bioRxiv* , 2026 (2026).
19. E. Cole, G.-J. Huizing, S. Addagudi, N. Ho, E. Hasanaj, M. Kuijs, T. Johnstone, M. Carilli, A. Davi, C. Ellington *et al.*, Foundation models improve perturbation response prediction, *BioRxiv* , 2026 (2026).
20. D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A. L. Gable, T. Fang,

N. T. Doncheva, S. Pyysalo *et al.*, The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest, *Nucleic Acids Research* **51**, D638 (2023).
21. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander *et al.*, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proceedings of the national academy of sciences* **102**, 15545 (2005).
22. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.*, Gene ontology: tool for the unification of biology, *Nature Genetics* **25**, 25 (2000).

Asking the Right QUESTIONS: Network Spillover Reveals Temporal Signal in Gene Co-expression Networks

Carly A. Bobak

Research Computing and Data in Information, Technology & Consulting, & Biomedical Data Science, Dartmouth College
Hanover, NH 03784, USA
†*E-mail: carly.a.bobak@dartmouth.edu*

Jaini Shah

Geisel School of Medicine, Dartmouth College
Lebanon, NH 03755, USA
E-mail: jaini.h.shah@dartmouth.edu

A. James O'Malley

The Dartmouth Institute for Health Policy and Clinical Practice & Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College
Lebanon, NH 03755, USA
E-mail: james.omalley@dartmouth.edu

Gene co-expression networks group genes into modules that approximate biological programs, but standard analyses summarize each module by its own activity and ignore what the module's network neighbors contribute. We adapt network spillover, a construct developed to quantify indirect (exogenous peer or spillover) effects that falls under the causal inference concept of interference in social networks, to gene co-expression networks. Treating modules as units and the co-expression graph as the conduit of influence, we decompose a module's relationship to an outcome into a direct term, its own activity, and a spillover term, the exposure it receives from other modules across between-module edges, and we test whether cross-module spillover improves the predictive accuracy of an outcome beyond that provided from its own direct activity. Applied to a longitudinal cynomolgus macaque model of *Mycobacterium tuberculosis* infection, with lung FDG avidity as a measure of disease severity, cross-module spillover of a given module into myeloid modules during the peak-response window predicts severity beyond the modules' own activity, is stable under resampling and cross-validation, and yields a platelet module whose own activity is uninformative while its spillover is not. The unsupervised network additionally recovers a coherent map of blood cell lineages and immune programs. The associations are correlational. We provide the framework as an open-access R package, QUESTIONS, and discuss extensions to time-varying and single-sample networks.

Keywords: Gene co-expression networks; network spillover; interference; disease severity; tuberculosis.

© 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.

1. Introduction

Biological function arises not from genes in isolation but from their coordinated activity.[1] Gene co-expression networks have become a standard tool for representing this coordination, grouping genes into modules that approximate pathways and cell-type programs and relating those modules to phenotypes.[2,3] When expression is profiled longitudinally, across development, circadian cycles, disease progression, or response to a perturbation, the coordination itself changes over time: modules gain and lose importance, and the relationships among genes rewire as the biological context shifts.[4,5] Capturing how network structure evolves, rather than collapsing a time course into a single static picture, is widely recognized as necessary for understanding dynamic regulatory processes, since a network built from all time points pooled together obscures precisely the transient structure that distinguishes one phase of a response from the next.[4]

Most co-expression analyses stop at the level of individual modules, relating each module's own activity to phenotype but rarely asking whether the information carried between modules is itself informative. Most temporal transcriptomic analysis still operates one gene at a time, tracking expression trajectories while ignoring the coordinated structure that co-expression is meant to capture.[5] Where network structure is modeled over time, the dominant strategies pursue a different goal than ours. One line of work estimates a separate network at each time point, for example through time-varying graphical models, and asks how edges appear or disappear;[6,7] a second, network propagation, diffuses a signal such as a set of seed genes across a fixed molecular network to amplify and prioritize associated genes, and has become a mainstay of disease-gene discovery;[8,9] a third, differential and single-sample co-expression analysis, quantifies how much the wiring around a gene or module changes between conditions or in an individual patient, and can relate that rewiring to clinical outcomes.[10,11] These approaches establish that network structure carries phenotypic information beyond mean expression. What none of them provides is a way to ask, for a given module, whether the activity of its network neighbors carries information about an outcome over and above the information in the module's own activity, which is the question we take up here.

The above question has a direct precedent outside biology. In the analysis of interference in social networks, spillover (or peer-exposure) methods quantify how a unit's outcome is shaped not only by its own treatment but by the treatments of the units it is connected to.[12] A concrete illustration comes from a stepped-wedge trial of a video-game intervention designed to increase physicians' Advance Care Planning (ACP) conversations with hospitalized patients.[13,14] Physicians were connected through a shared-patient network constructed from insurance billing records. Because physicians who had not yet received the intervention still collaborated with those who had, the standard trial analysis risked conflating the intervention's direct effect with the effect of proximity to already-treated peers. Modeling each physician's network exposure separately from their own treatment status showed that peer exposure predicted ACP billing in its own right, and that omitting it biased, and in some specifications reversed, the estimated direct effect. Other areas of social science or medicine in which peer-effects have been modeled include Medicare fraud[15] or extreme spending in home health care,[15] polypharmacy in older people,[16] the adoption of medical technologies such as

robotic surgery,[17] hospital adoption of implantable cardiac defibrillators,[18] and physician willingness to participate in research studies.[19] The same logic underlying each of these examples applies here: a module's neighbors in a co-expression network may carry information about a phenotype that the module's own activity does not.

In recent work extending stepped-wedge cluster-randomized trials to account for interference, we represented a physician collaboration network as a weighted adjacency matrix, row-normalized it to a transition matrix $\mathbf{W}$ whose entry $(\mathbf{W})_{ij}$ is the share of influence acting upon actor i from actor j (by construction of the network, $W_{ii} = 0$), and defined each actor's network exposure as the neighbor-weighted sum of other actors' treatment, $\boldsymbol{e} = \mathbf{W}\,\boldsymbol{d}$.[13,14] Including this exposure term alongside the direct-treatment term in an outcome model separates an actor's own effect from the peer-to-peer influence reaching it through the network. Ignoring such spillover biases the estimated direct effect when interference is present.[13] In the following, our construction is deliberately generic: it requires only a weighted network in which the units are nodes, a per-unit quantity to propagate, and an outcome.

We adapt the above statistical approach used to account for spillover effects in human and institutional/organizational studies in which the actors have the potential to influence one another such that one actors outcome might be affected by another actor's treatment ("intervening on the untreated") to gene co-expression networks. We treat co-expression modules as the units and a gene network as the conduit of influence, applying the same construction to a distinct biological question. This decomposes each module's relationship to an outcome into a direct term (the module's own activity) and a spillover term (the exposure it receives from other modules through between-module edges). We then ask whether cross-module spillover predicts the outcome beyond direct activity. We apply the method to a longitudinal cynomolgus macaque model of *Mycobacterium tuberculosis* infection in which whole-blood transcriptomes are paired with lung FDG avidity, an imaging measure of disease severity.[20] We provide the framework as an R package, QUESTIONS.

2. The QUESTIONS framework

2.1. *Direct and cross-module spillover on a co-expression network*

Let $\mathbf{A}$ be the symmetric weighted adjacency matrix of a gene co-expression network on p genes, partitioned into modules by a community assignment $m(\cdot)$. A biological network carries signal between modules, not only within them. We therefore treat each module as a unit of analysis whose outcome has two quantities: its own state (the direct term) and a neighbor-weighted summary of the states of the modules it connects to (the exposure, or spillover, term). The weights come from an (optionally) row-normalized network.[12–14,21] We row-normalize the network so that a module's spillover reflects the proportion of its connections that are active rather than their absolute number. This suits correlation-based edge weights and prevents large modules from registering greater spillover merely as an artifact of their size. Retaining the unnormalized adjacency is a defensible alternative under which influence is treated as cumulative rather than proportional, at the cost of a spillover coefficient whose magnitude is entangled with module connectivity.

To make the spillover term carry only cross-module influence, we first row-normalize the

full adjacency matrix, $P_{gh} = A_{gh} / \sum_{h'} A_{gh'}$, so that each row of $\mathbf{P}$ is a gene's distribution of influence over all of its neighbors, and then retain only the between-module entries,

$$W_{gh} = P_{gh} \mathbb{1}\{\, m(g) \neq m(h) \,\}. \tag{1}$$

We do not renormalize after this step, so the row sum of $\mathbf{W}$ is the fraction of gene g's influence that leaves its own module. This has a useful consequence: a gene wired predominantly to its own module contributes little cross-module exposure, because most of its normalized weight was assigned to within-module neighbors that are then removed. Restricting to between-module weights prevents the spillover term from re-expressing a module's own internal activity, so that direct and spillover terms measure distinct quantities by construction.

For a unit of observation i (a sample, subject, or, in the case study below, an animal), let $\boldsymbol{d}_i \in \mathbb{R}^p$ be a gene-level effect vector, for example each gene's departure from a reference state (defined below). The gene-level spillover reaching each gene is the neighbor-weighted departure $\boldsymbol{e}_i = \mathbf{W}\,\boldsymbol{d}_i$, exactly paralleling the social-network exposure $\boldsymbol{e} = \mathbf{W}\boldsymbol{d}$.[13] We then summarize both quantities per module.

Let $G_M \subseteq \{1, \ldots, p\}$ index the genes of module M, and let $\mathbf{X}_i \in \mathbb{R}^{|G_M|}$ denote unit i's standardized expression of those genes. The module eigengene[22] is the first principal component, $\mathrm{eig}_{iM} = \mathbf{v}_M^\top \mathbf{X}_i$, where $\mathbf{v}_M$ is the leading loading vector of the standardized module-expression matrix, sign-aligned so that $\mathbf{v}_M$ correlates positively with mean module expression. We define a departure operator relative to unit i's own baseline samples $\mathcal{B}_i$,

$$\Delta(z_i) \;=\; z_i - \frac{1}{|\mathcal{B}_i|} \sum_{b \in \mathcal{B}_i} z_b, \tag{2}$$

which subtracts a unit's mean over its baseline samples. The direct term for module M is the eigengene expressed as a departure, $\mathrm{direct}_{iM} = \Delta(\mathrm{eig}_{iM})$, and the gene-level effect vector propagated above is the per-gene departure, $(\boldsymbol{d}_i)_g = \Delta(x_{ig})$ for gene g. The spillover term is the mean of the gene-level spillover over the module's genes,

$$\mathrm{spill}_{iM} \;=\; \frac{1}{|G_M|} \sum_{g \in G_M} (\boldsymbol{e}_i)_g. \tag{3}$$

In words, the direct term is how much module M itself departed from its reference, and the spillover term is how much departure reached M from the rest of the network.

2.2. *Outcome model*

Given an outcome y_i, the framework fits, for each module M,

$$y_i = \beta_0 + \beta_{\mathrm{dir}}\, \mathrm{direct}_{iM} + \beta_{\mathrm{spill}}\, \mathrm{spill}_{iM} + \varepsilon_i, \tag{4}$$

with the direct and spillover terms standardized so that β_{spill} is interpretable per standard deviation sized change of the spillover predictor. The estimand of interest is β_{spill}: whether a module's cross-module spillover predicts the outcome beyond the module's own activity. The link function is chosen to suit the outcome (identity, logistic, log, or otherwise). Because the underlying network is an undirected co-expression graph, in which edges may arise through a selective process such as preferential attachment[23] rather than random assignment, β_{spill} is an association rather than a causal effect.

2.3. *Inference for small samples*

Given the small sample sizes typical of omics studies, we pair the outcome model with distribution-free and resampling inference to interpret β_{spill} from the full (spillover) model relative to the reduced (direct-only) model. The partial spillover effect is tested by a Freedman–Lane permutation scheme, which permutes the residuals of the direct-only model and refits, yielding a distribution-free test of β_{spill} that conditions correctly on the direct term;[24,25] when many module × condition combinations are tested, permutation p-values are adjusted by Benjamini–Hochberg over a pre-specified candidate family.[26] Coefficient stability is summarized by bootstrapping units to test for sign-consistency and to estimate confidence intervals of β_{spill};[27] added predictive value is summarized by repeated k-fold cross-validation, as the gain in out-of-sample R^2 from adding the spillover term to a direct-only model; and leave-one-unit-out refitting is used as an influence check.

2.4. *Implementation*

The framework is implemented in the R package QUESTIONS (Quantifying Expression Spillover Trajectories in Omic Network Series), which constructs the between-module weight matrix, computes each module's direct and spillover effects, and runs the outcome model with the validation battery above. It is freely available at `https://github.com/CarlyBobak/QUESTIONS`.

3. Case study: tuberculosis disease severity

3.1. *Data and preprocessing*

We applied the framework to GSE84152, a longitudinal study of *Mycobacterium tuberculosis* infection in cynomolgus macaques in which whole-blood transcriptomes were profiled on an Illumina HumanHT-12 v4 array (human gene symbols) at serial time points from two pre-infection baselines through six months post-infection, with per-animal lung FDG avidity measured by PET-CT as a surrogate for disease severity.[20] The outcome is total lung FDG avidity; because it is strongly right-skewed we modeled it on the $\log(1+x)$ scale, standard for total-glycolytic PET summaries and robust to a few high-burden animals. Expression was corrected for batch structure across processing runs with the empirical-Bayes ComBat method.[28] Each per-animal, per-gene departure was computed relative to that animal's own pre-infection baseline, so the propagated quantity is a within-animal response and between-animal baseline differences are removed. The analysis included $n = 37$ animals. Nine time points were profiled per animal (two pre-infection baselines and seven post-infection: days 7, 10, 20, 30, 42, and 56, and month 3). Pooling the per-animal, per-time-point profiles that passed quality control yields the 337 windowed samples used to build the network.

3.2. *Consensus network construction*

To focus the network on genes that respond to infection rather than on constitutive covariation, we restricted the analysis to infection-responsive genes identified by a moderated F-test for any change across any time points, with within-animal correlation accommodated by

treating animal as a blocking factor via `duplicateCorrelation` and using an empirical-Bayes variance estimator;[29–31] genes passing a false-discovery threshold of 0.01 were retained (3,829 genes). Over these genes we built a single consensus network, aggregated across 337 windowed samples so that it reflects coordination sustained across the time course rather than the structure of any one time point. Gene pairs were joined by a signed mutual k-nearest-neighbor rule ($k = 50$): an edge is kept only when each gene lies in the other's k-neighborhood.[32] The reciprocal criterion prunes one-sided ties, yielding a sparse, modular graph that preserves the between-module bridges the spillover analysis requires.

3.3. *Module identification*

Communities were detected with Infomap, a flow-based method that partitions the graph by compressing the description of a random walk over it;[33] each community defines a module, and each module is summarized by its eigengene as in Section 2. The resulting network is strongly modular (modularity 0.761, within-module edge fraction 0.83).

3.4. *Pathway enrichment and module annotation*

Modules were annotated by over-representation analysis against curated gene-set collections, using the responsive-gene network as background so that enrichment reflects structure within the analyzed set.[34,35] Because a single immune-centric collection under-annotates non-immune modules and a naive top-hit-per-collection rule is dominated by large generic terms, we combined a cell-type call from the Blood Transcription Modules[36] with a specificity-scored (term frequency–inverse document frequency, TF–IDF) pooling of hits across each of GO biological process, Reactome, WikiPathways, and MSigDB Hallmark.[37] Enrichment is considered corroborating evidence of the network's biological coherence, not as independent statistical validation of the spillover results, since the same modules carry both.

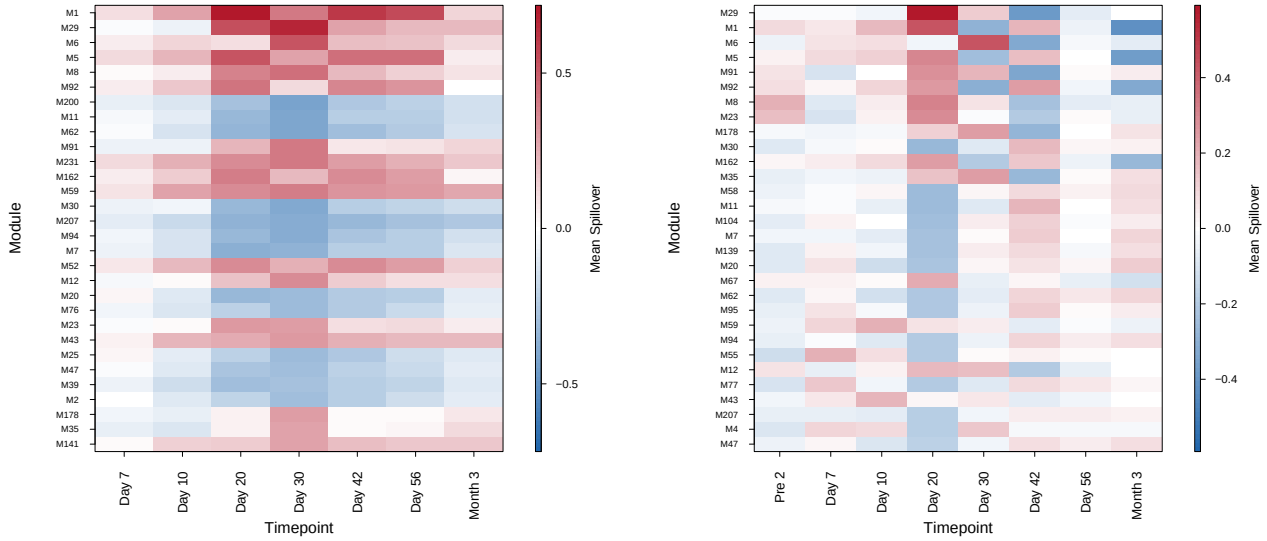
3.5. *Candidate selection and testing*

Applying the outcome model of Section 2, candidate module $\times$ time-point combinations were restricted, before examining the outcome, to the modules with the most pronounced descriptive spillover dynamics (Figure 1). This outcome-independent pre-selection (the "candidate family") fixes the set of tests before any outcome is examined. The 15 modules with the most pronounced descriptive spillover dynamics, each evaluated at the seven post-baseline time points, give the 105 tests over which Benjamini–Hochberg correction is applied. The direct and spillover terms were standardized, the model fit with an identity link on the $\log(1 + x)$ outcome across $n = 37$ animals, and β_{spill} evaluated with the permutation, bootstrap, cross-validation, and leave-one-out procedures described in Section 2.

4. Results

4.1. *Cross-module spillover predicts severity beyond direct activity*

Across the pre-specified family of 105 module $\times$ time-point tests, spillover clears Benjamini–Hochberg adjustment for five combinations, all in myeloid or platelet modules and all in



(a) Spillover of baseline departure.

(b) Spillover of the timepoint increment.

Fig. 1: **Cross-module spillover dynamics over the infection time course, and the two questions spillover can ask.** Each cell is the mean cross-module spillover received by a module (row) at a time point (column): the between-module, neighbor-weighted departure defined in Section 2, averaged over animals and over the module's genes. Warm values indicate a module receiving positive cross-module influence, cool values the reverse. The panels differ only in how each gene's change is defined before it is spread across the network: relative to the pre-infection baseline in (a), or as the change from the preceding time point in (b). In (a) spillover reads as the cumulative cross-module influence that has accrued relative to the healthy state. In (b) spillover reads as the influence being exchanged at each interval. The two views pose different temporal questions of the same network, cumulative burden versus instantaneous exchange, and clarify that spillover is a lens applied to a chosen departure rather than a fixed quantity. Modules with the most pronounced spillover dynamics, selected independently of the outcome, defined the candidate set carried to the outcome model.

the peak-response window (Table 1). The strongest are cross-module spillover into the neutrophil/inflammatory module M8, at both day 42 and month 3, and into the platelet module M6 at day 56, each at adjusted $p = 0.017$; spillover into the monocyte module M91 follows at day 30 and day 42 (adjusted $p = 0.026$ and 0.052). All five carry a large positive standardized spillover coefficient (for example M8 at day 42, $\beta_{\text{spill}} = 8.14$, bootstrap 95% CI $[5.42, 11.67]$; M91 at day 30, 7.72, $[3.01, 11.60]$; M6 at day 56, 2.45, $[0.47, 3.98]$), are sign-stable in 100% of bootstrap resamples, are robust to leave-one-animal-out, and improve out-of-sample prediction (cross-validated ΔR^2 between +0.23 and +0.39). The spillover p-values are corrected across the full family whereas the direct-term p-values are reported unadjusted, so this comparison holds spillover to the stricter standard. The signal concentrates in the day-42 to month-3 window, coinciding with the peak transcriptional response reported for this cohort.[20]

Table 1: Cross-module spillover predicts lung FDG avidity beyond a module's own activity. Outcome is $\log(1 + \text{total lung FDG avidity})$; β_{spill} is the standardized spillover coefficient (per SD), with a bootstrap 95% confidence interval over animals. p_{perm} is a Freedman–Lane permutation p-value for the partial spillover effect; p_{BH} is Benjamini–Hochberg adjusted p-values across all 105 module × timepoint tests. ΔR^2_{CV} is the repeated 5-fold cross-validated gain in out-of-sample R^2 from adding spillover to a direct-only model. p_{direct} is the direct-term (module eigengene) p-value in the same model, reported unadjusted, so the spillover effect is held to a stricter multiple-testing standard than the direct term. All five combinations are sign-stable in 100% of bootstrap resamples and robust to leave-one-animal-out.

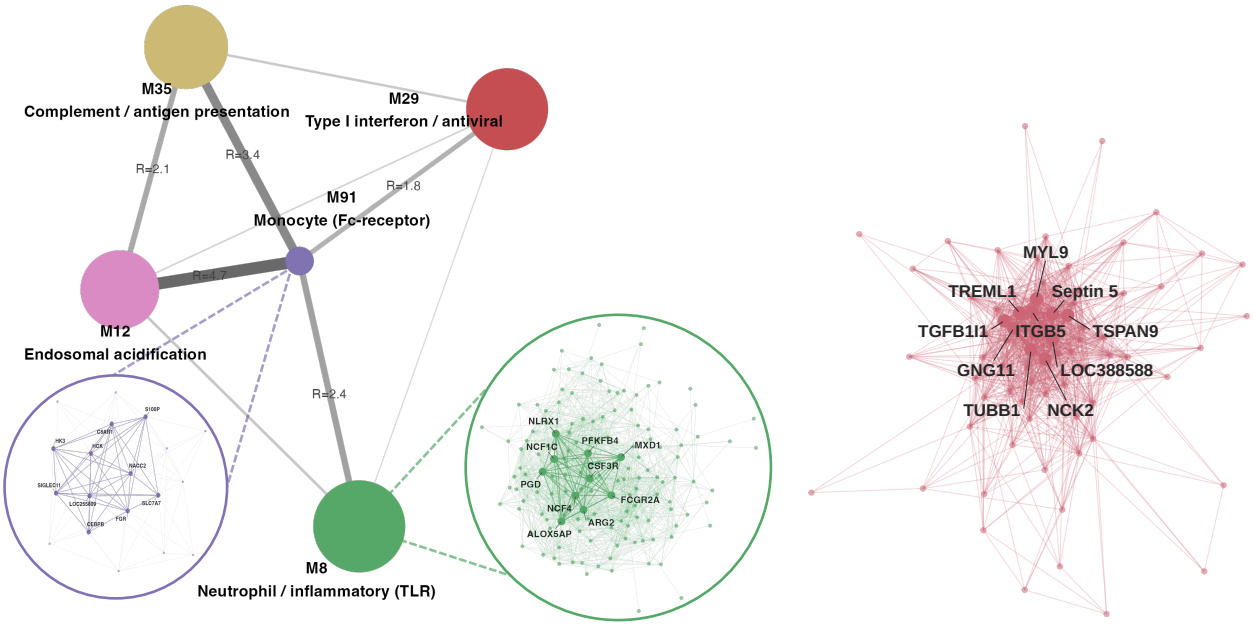
Module	Time	β_{spill}	95% CI	p_{perm}	p_{BH}	ΔR^2_{CV}	p_{direct}
M8 (neutrophil/inflammatory)	D42	8.14	[5.42, 11.67]	5×10^{-4}	0.017	+0.389	2.9×10^{-4}
M8 (neutrophil/inflammatory)	M3	7.27	[3.82, 11.18]	5×10^{-4}	0.017	+0.296	3.1×10^{-3}
M6 (platelet/coagulation)	D56	2.45	[0.47, 3.98]	5×10^{-4}	0.017	+0.227	0.118
M91 (monocyte, Fc-receptor)	D30	7.72	[3.01, 11.60]	1×10^{-3}	0.026	+0.230	5.8×10^{-3}
M91 (monocyte, Fc-receptor)	D42	6.90	[3.18, 11.25]	2×10^{-3}	0.052	+0.230	4.1×10^{-3}

4.2. *Direct activity and spillover capture distinct information*

The platelet module M6 best illustrates what the spillover value. Its cross-module spillover is among the strongest predictors of severity (day 56, adjusted $p = 0.017$; Figure 2b), yet its own activity is uninformative (direct-term $p = 0.118$). Thus, this framework allows a module whose activity carries no outcome signal to still receive outcome-relevant spillover through the network. The interferon module M29 shows the opposite pattern. Its spillover never clears correction (best adjusted $p = 0.62$), while its own activity is strongly associated with severity (direct-term $p = 2.8 \times 10^{-5}$ at day 42). M6 carries outcome information through its neighbors and M29 through its own activity, which shows that the two terms capture different information rather than restating each other. A binary high/low FDG outcome recovered essentially no spillover signal, consistent with dichotomization discarding graded severity.

4.3. *The unsupervised network recovers coherent blood biology*

The network recovers a coherent map of the major hematopoietic lineages and immune programs, erythroid, translation, neutrophil, interferon, complement, T, B, NK, monocyte, plasma, and platelet modules, with no cell-type input; the full network is shown in Figure 3). The outcome-associated modules sit where TB immunopathology would place them: M8 and M91 in the myeloid-inflammatory compartment active during the peak-response window (Figure 2a), and M6 in platelet/coagulation biology (Figure 2b). The myeloid core is densely intercoupled while M6 is peripheral, coupling to the myeloid modules at near-zero degree-adjusted strength, so the M6 spillover signal does not travel an identifiable myeloid route. Spillover reaches a module over all of its between-module edges regardless of the confidence tier of the neighbor. M6's signal therefore arrives through its individually weak, non-myeloid connections rather than any strong link to the myeloid core, which is why no single coupling in Figure 2 accounts for it.



(a) Myeloid-core meta-network. (b) Platelet module M6.

Fig. 2: **Network structure behind the spillover result.** (a) Meta-network of the myeloid core: each node is a module, sized by gene count, and each edge a degree-adjusted coupling strength R between modules, with only above-chance couplings drawn and edge width scaled to R. The monocyte module M91 sits at the hub, coupling most strongly to the endosomal module M12 ($R = 4.7$) and linking the complement (M35), interferon (M29), and neutrophil (M8) modules. Gene-level zoom insets show the internal hub structure of M8 and M12. Spillover into M8 and M91 during the peak-response window predicts disease severity. (b) The platelet/coagulation module M6 at gene level, with platelet-specific hubs (MYL9, TREML1, ITGB5, TUBB1, GNG11). In contrast to the densely inter-coupled myeloid core, M6 is peripheral: its degree-adjusted coupling to the myeloid modules is near zero. Its spillover is associated with severity even though its own activity is not (Section 4)

5. Discussion

We have shown that network spillover, a construct developed in the analysis of interference in social networks, transfers to gene co-expression networks and captures temporal biological signals that a module's own activity does not. The framework separates a module's own activity from the activity that reaches it across between-module edges, and tests whether the latter predicts an outcome beyond the former. In a longitudinal model of tuberculosis this decomposition identified cross-module spillover into myeloid modules as predictive of disease severity, stably and out of sample, and produced a clean case, the platelet module, in which the module's own activity was uninformative while its spillover was not.

The platelet module M6 invites an analogy to interference in randomized trials, where a control unit that acquires exposure through its treated peers can appear affected despite

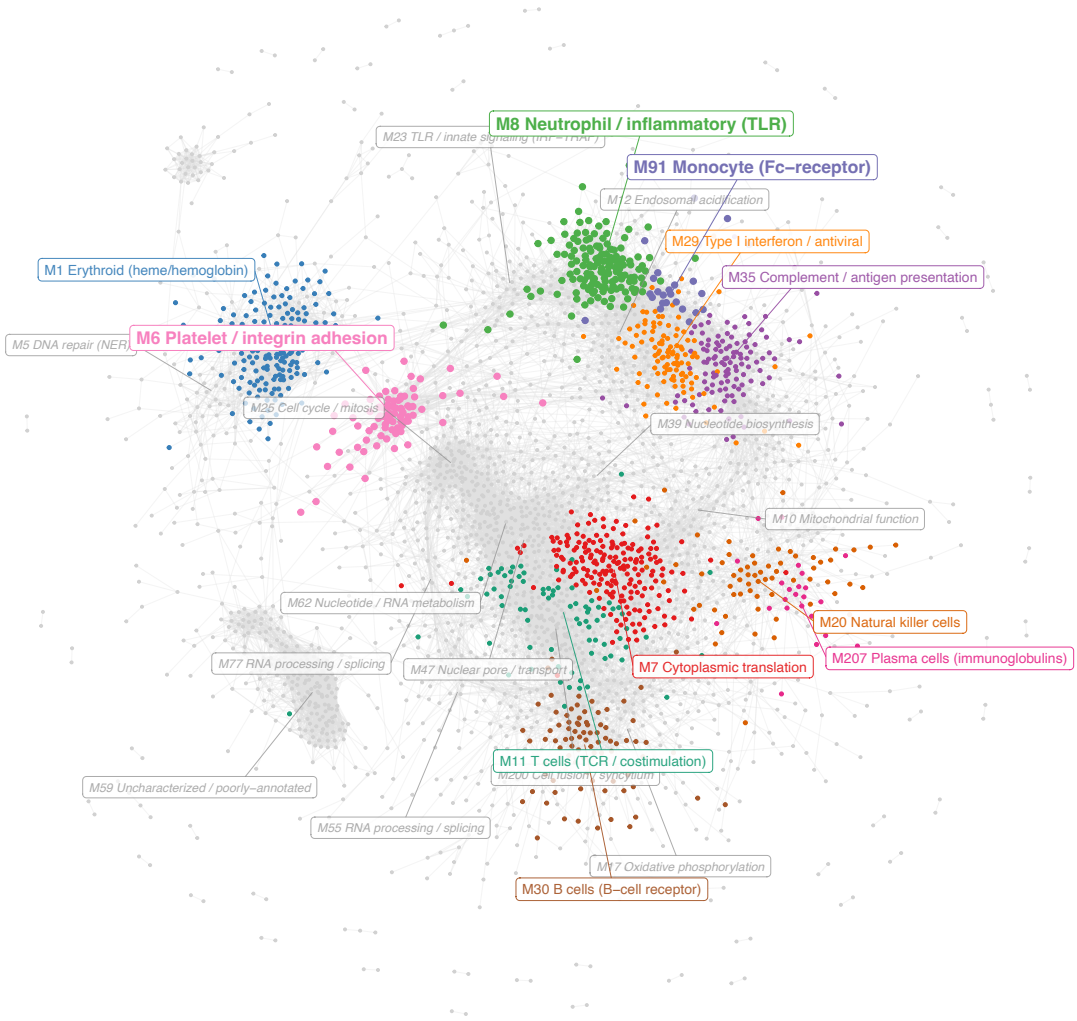


Fig. 3: **The full infection-responsive co-expression network.** All non-isolated genes, laid out by a force-directed algorithm and colored by module (high- and medium-confidence modules only; tentative modules in gray italic, unannotated modules unlabeled). Isolated singletons are omitted. The network recovers the major hematopoietic lineages and immune programs as distinct communities with no cell-type input; the outcome-associated modules M6, M8, and M91 are emphasized.

no direct assignment. Here, M6's own activity is uninformative about severity, yet outcome-relevant signal reaches it across the network. The analogy is partial. In a trial, contamination is a nuisance that biases the treatment-effect estimate and is adjusted away. Here, the neighbor-borne signal is the quantity of interest.

This paper fills a gap left by existing approaches to network structure in biology. Network propagation diffuses a seed signal over a fixed network to amplify and prioritize genes,[8,9] and differential and single-sample co-expression methods quantify how wiring changes between conditions or individuals and relates that change to outcomes;[10] both establishing that network

structure carries phenotypic information. What the spillover decomposition adds is the explicit separation of a module's own signal from its neighbors' contribution within a single outcome model, so that the neighbor contribution can be tested on its own terms. The two descriptive views of spillover in Figure 1, one relative to a pre-infection baseline and one as a timepoint increment, show that the same construction answers different temporal questions depending on the departure that is propagated.

Although the framework is the focus, the biological hits are plausible and literature-consistent, which supports the method's validity. The modules whose spillover tracked severity are the myeloid neutrophil and monocyte programs, and the network recovered the canonical type I interferon signature as a distinct module, all of which are central to the blood transcriptional response to *M. tuberculosis* and to the source cohort's own findings on FDG-defined severity.[20] The platelet module is likewise coherent with a growing literature on platelet involvement in tuberculosis immunopathology.[38] We emphasize that these associations are correlational and that a full immunological interpretation is beyond the scope of this paper.

Several limitations bound these conclusions. The most important is that every result is conditional on the inferred network. Spillover is defined over the between-module edges, so a mis-estimated network, wrong edges, missed edges, or an unstable community partition, propagates directly into the spillover term and into every downstream test. Co-expression networks are sensitive to construction choices such as the similarity measure, sparsification, and community-detection algorithm, and our use of a single fixed consensus network conditions the entire analysis on one such choice. A different but equally defensible network could shift which modules and bridges carry the signal. Second, and common to longitudinal omics, the sample size is small ($n = 37$ animals from a single cohort with no external replication), which is why inference rests on permutation, bootstrap, cross-validation, and leave-one-out rather than on parametric p-values, and why candidate modules were pre-specified independently of the outcome. The surviving associations should nonetheless be read as strong internal evidence awaiting replication rather than as established effects. Third, the enrichment that anchors the biological reading corroborates the network's coherence but does not independently validate the spillover results, since the same modules carry both. Fourth, the direction of influence is not identified. The network is undirected co-expression, and although temporal ordering offers a directionality proxy, spillover between modules does not establish that one module drives another, so the results are associational. Fifth, the module summaries (eigengenes) and the mean-over-genes spillover summary discards within-module heterogeneity, and results can depend on the resolution at which communities are defined. Finally, the platform is an older microarray with deprecated probe annotations, which limits annotation of some modules, and the outcome is a single imaging surrogate of severity rather than a clinical endpoint.

The clearest extensions concern the network itself. Our analysis propagates per-animal departures over one network fixed across the time course, which assumes the wiring is stable even as expression changes; a natural next step is to let the network vary over time, so that spillover is computed on timepoint-specific or window-specific networks and can capture rewiring rather than only changing node states. A complementary direction operates at the level of individuals. Single-sample network inference, in particular LIONESS, estimates a

network for each sample by measuring how the aggregate network shifts when that sample is left out,[39] so that spillover could be computed on per-sample networks and related to per-sample outcomes, turning the present fixed-network analysis into one that also captures individual network heterogeneity. Because such per-sample networks yield an individual-level spillover profile, this is also the natural route to patient stratification and precision-medicine applications of the framework. Both directions would let the framework ask not only whether a module's neighbors carry outcome information, but whether the connections themselves reorganize over time or across individuals in outcome-relevant ways. Further extensions include incorporating lagged spillover terms to model influence that acts across, rather than within, time points, and applying the same decomposition to other network types and omics layers, since the construction requires only a weighted network over units, a per-unit quantity to propagate, and an outcome.

6. Conclusion

Network spillover offers a principled way to ask a question that standard co-expression analysis leaves unaddressed: whether the activity of a module's network neighbors carries information about an outcome beyond the module's own activity. By adapting the direct-versus-spillover decomposition from the analysis of interference in social networks, and pairing it with small-sample inference, the QUESTIONS framework makes that question testable on temporal omics data. In a longitudinal model of tuberculosis it identified cross-module spillover into myeloid modules as predictive of disease severity and recovered coherent blood biology without supervision, including a platelet module whose spillover predicted severity while its own activity did not. The approach is general, requiring only a weighted network over units, a per-unit quantity to propagate, and an outcome, and it extends naturally to time-varying and single-sample networks. As longitudinal and multi-omic studies grow, network spillover effect modeling and associated statistical tests provide a compact, interpretable lens on how influence travels through biological networks.

Use of large language models

The network-spillover methodology adapted here was developed in the social-network setting without the use of large language models.[13,14] In preparing this adaptation to gene co-expression networks, the authors used an LLM-based assistant in a supporting role: to help draft and revise manuscript text, to assist in writing and debugging R-package code, and to help prepare figures and documentation. All methodological, modeling, and scope decisions were made by the authors, who ran every analysis, verified all code and numerical results against the underlying data, and take full responsibility for the content. The assistant did not generate data or produce results autonomously.

References

1. L. H. Hartwell, J. J. Hopfield, S. Leibler and A. W. Murray, From molecular to modular cell biology, *Nature* **402**, C47 (December 1999).

2. P. Langfelder and S. Horvath, WGCNA: An R Package for Weighted Correlation Network Analysis, *BMC Bioinformatics* **9**, p. 559 (2008).
3. S. van Dam, U. Võsa, A. van der Graaf, L. Franke and J. P. de Magalhães, Gene Co-expression Analysis for Functional Classification and Gene–Disease Predictions, *Briefings in Bioinformatics* **19**, 575 (2018).
4. K. Greenham and C. R. McClung, Time to Build on Good Design: Resolving the Temporal Dynamics of Gene Regulatory Networks, *Proceedings of the National Academy of Sciences* **115**, 6325 (2018).
5. M. Toolabinejad and others, A Comprehensive Survey of Gene Co-expression Network Analysis: Methods, Tools, Challenges, and Future Directions, *Network Modeling Analysis in Health Informatics and Bioinformatics* **14** (2025).
6. M. Kuismin and M. J. Sillanpää, tvsfglasso: Time-varying Scale-free Graphical Lasso for Network Estimation from Time-series Data, *PLOS Computational Biology* **21**, p. e1013710 (2025).
7. S. Ozminkowski, L. Hou, D. R. Jacobs and H. Jiang, Dynamic Co-Expression Network Estimation via Multivariate Mixed-Effects Models (2026).
8. L. Cowen, T. Ideker, B. J. Raphael and R. Sharan, Network Propagation: A Universal Amplifier of Genetic Associations, *Nature Reviews Genetics* **18**, 551 (2017).
9. G. Visonà, E. Bouzigon, F. Demenais and G. Schweikert, Network Propagation for GWAS Analysis: A Practical Guide to Leveraging Molecular Networks for Disease Gene Discovery, *Briefings in Bioinformatics* **25**, p. bbae014 (2024).
10. L. Wei, Y. Xin, M. Pu and Y. Zhang, Patient-specific Analysis of Co-expression to Measure Biological Network Rewiring in Individuals, *Life Science Alliance* **7**, p. e202302253 (2024).
11. D. M. Gysi, T. de Miranda Fragoso, F. Zebardast, W. Bertoli, V. Busskamp, E. Almaas and K. Nowick, Whole Transcriptomic Network Analysis Using Co-expression Differential Network Analysis (CoDiNA), *PLOS ONE* **15**, p. e0240523 (2020).
12. M. G. Hudgens and M. E. Halloran, Toward Causal Inference With Interference, *Journal of the American Statistical Association* **103**, 832 (2008).
13. A. J. O'Malley, C. A. Bobak and A. E. Barnato, Causal Inference for Intervention Spillover in a Stepped-Wedge Cluster-randomized Trial: Lessons from a Physician Network, *Social Networks* **86**, 431 (2026).
14. C. A. Bobak, M. A. Murphy, A. E. Barnato and A. J. O'Malley, Gaming the System: Evaluating Spillover in a Video Game Intervention for Advance Care Planning using Physician Social Networks, *Health Services and Outcomes Research Methodology* (2026).
15. A. J. O'Malley, T. A. Bubolz and J. S. Skinner, The diffusion of health care fraud: A bipartite network analysis, *Social Science & Medicine* **327**, p. 115927 (June 2023).
16. X. Ran, N. E. Morden, E. Meara, E. L. Moen, D. N. Rockmore and A. J. O'Malley, Exploiting relationship directionality to enhance statistical modeling of peer-influence across social networks, *Statistics in Medicine* **43**, 4073 (September 2024).
17. G. Chen and A. J. O'Malley, Developing and Comparing Four Families of Bayesian Network Autocorrelation Models for Binary Outcomes: Estimating Peer Effects Involving Adoption of Medical Technologies, *Biometrical Journal. Biometrische Zeitschrift* **67**, p. e70030 (February 2025).
18. A. J. O'Malley, E. L. Moen, J. P. W. Bynum, A. M. Austin and J. S. Skinner, Modeling peer effect modification by network strength: The diffusion of implantable cardioverter defibrillators in the US hospital network, *Statistics in Medicine* **39**, 1125 (April 2020).
19. C. A. Bobak, D. Mohan, M. A. Murphy, A. E. Barnato and A. J. O'Malley, Constructing within and between hospital physician social networks for modeling physician research participation, *BMC Medical Research Methodology* **23**, p. 253 (October 2023).
20. H. P. Gideon, J. A. Skinner, N. Baldwin, J. L. Flynn and P. L. Lin, Early Whole Blood Transcrip-

tional Signatures Are Associated with Severity of Lung Inflammation in Cynomolgus Macaques with Mycobacterium tuberculosis Infection, *The Journal of Immunology* **197**, 4817 (2016).
21. P. M. Aronow and C. Samii, Estimating Average Causal Effects Under General Interference, with Application to a Social Network Experiment, *The Annals of Applied Statistics* **11**, 1912 (2017).
22. P. Langfelder and S. Horvath, Eigengene Networks for Studying the Relationships Between Co-expression Modules, *BMC Systems Biology* **1**, p. 54 (2007).
23. A.-L. Barabasi and R. Albert, Emergence of Scaling in Random Networks, *Science* **286**, 509 (October 1999).
24. D. Freedman and D. Lane, A Nonstochastic Interpretation of Reported Significance Levels, *Journal of Business & Economic Statistics* **1**, 292 (1983).
25. A. M. Winkler, G. R. Ridgway, M. A. Webster, S. M. Smith and T. E. Nichols, Permutation Inference for the General Linear Model, *NeuroImage* **92**, 381 (2014).
26. Y. Benjamini and Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society: Series B* **57**, 289 (1995).
27. B. Efron, Bootstrap Methods: Another Look at the Jackknife, *The Annals of Statistics* **7**, 1 (1979).
28. W. E. Johnson, C. Li and A. Rabinovic, Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods, *Biostatistics* **8**, 118 (2007).
29. M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi and G. K. Smyth, limma Powers Differential Expression Analyses for RNA-sequencing and Microarray Studies, *Nucleic Acids Research* **43**, p. e47 (2015).
30. G. K. Smyth, Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments, *Statistical Applications in Genetics and Molecular Biology* **3**, p. Article 3 (2004).
31. G. K. Smyth, J. Michaud and H. S. Scott, Use of Within-array Replicate Spots for Assessing Differential Expression in Microarray Experiments, *Bioinformatics* **21**, 2067 (2005).
32. M. R. Brito, E. L. Chávez, A. J. Quiroz and J. E. Yukich, Connectivity of the Mutual k-nearest-neighbor Graph in Clustering and Outlier Detection, *Statistics & Probability Letters* **35**, 33 (1997).
33. M. Rosvall and C. T. Bergstrom, Maps of Random Walks on Complex Networks Reveal Community Structure, *Proceedings of the National Academy of Sciences* **105**, 1118 (2008).
34. U. Raudvere, L. Kolberg, I. Kuzmin, T. Arak, P. Adler, H. Peterson and J. Vilo, g:Profiler: A Web Server for Functional Enrichment Analysis and Conversions of Gene Lists (2019 update), *Nucleic Acids Research* **47**, W191 (2019).
35. T. Wu, E. Hu, S. Xu, M. Chen, P. Guo, Z. Dai, T. Feng, L. Zhou, W. Tang, L. Zhan, X. Fu, S. Liu, X. Bo and G. Yu, clusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data, *The Innovation* **2**, p. 100141 (2021).
36. S. Li, N. Rouphael, S. Duraisingham and others, Molecular Signatures of Antibody Responses Derived from a Systems Biology Study of Five Human Vaccines, *Nature Immunology* **15**, 195 (2014).
37. A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov and P. Tamayo, The Molecular Signatures Database Hallmark Gene Set Collection, *Cell Systems* **1**, 417 (2015).
38. D. E. Kirwan, D. L. W. Chong and J. S. Friedland, Platelet Activation and the Immune Response to Tuberculosis, *Frontiers in Immunology* **12** (May 2021).
39. M. L. Kuijjer, M. G. Tung, G. Yuan, J. Quackenbush and K. Glass, Estimating Sample-Specific Regulatory Networks, *iScience* **14**, 226 (2019).

PhysioMap: an ontology-grounded causal knowledge graph of human physiology

Robert Hoehndorf
King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
E-mail: robert.hoehndorf@kaust.edu.sa

Paul N. Schofield
Dept of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, UK

Georgios V. Gkoutos
Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK

Physiological mechanisms connect molecular changes to measurements such as blood pressure and plasma metabolite concentrations. We developed PhysioMap to represent these connections across biological scales. PhysioMap defines physiological traits through ontologies and distinguishes causal influence, production, constitution, quantitative identity, and modulation. A versioned projection maps entailed ontology patterns to typed constraints on quantitative structural causal models. A derivative-sign abstraction supports qualitative prediction of intervention responses at a stable equilibrium with feedback. The resource contains 1,699 traits and 2,387 typed relations. Expert review assessed 83 relations across all five types. In a rare metabolic disease evaluation against the Human Phenotype Ontology annotation corpus, the solver returned 175 directions across 870 downstream gene–phenotype pairs. Of these directions, 171 agreed with the reference, giving 97.7% agreement at 20.1% coverage. We compared forward predictions and inverse lesion rankings with shortest signed paths and signed diffusion on the same reference. PhysioMap connects physiological trait definitions to executable intervention queries and supports the assembly of models from explicitly typed physiological knowledge.

Keywords: systems biology; computational physiology; causal knowledge graph; ontology; structural causal model; qualitative abstraction; comparative statics; feedback; abduction.

1. Introduction

Physiological interventions can affect several organs through feedback. For example, inhibiting angiotensin-converting enzyme (ACE) reduces its activity in the renin–angiotensin–aldosterone system, which connects angiotensin II to vascular resistance, renal sodium handling, and blood pressure. A model must combine these dependencies to predict the direction of the blood-pressure response. A measured trait also needs a precise meaning: plasma angiotensin II concentration specifies a chemical, a quantity, and its compartment.

We define a physiological trait as a quality borne by an entity or process in a specified biological context. Forward prediction takes an intervention on one or more traits and computes which other traits increase or decrease after the system reaches a new equilibrium.

© 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.

Abduction takes observed trait directions and ranks interventions that explain them. Rare metabolic diseases supply an application: an HFE-associated defect reduces hepcidin, which alters ferroportin-dependent iron handling and measurable iron traits. Here, a lesion variable denotes the primary activity or physiological function affected by a gene-associated disease mechanism.

Directed resources such as SIGNOR, OmniPath, INDRA, Reactome, and KEGG provide regulatory, signaling, and reaction knowledge.[1–5] Network propagation spreads information through their connections,[6,7] while sign-consistency and optimization methods find network explanations for measured molecular changes.[8,9] Predicting an intervention response in a feedback system additionally requires a rule for combining coupled mechanisms and opposing contributions. Parameterized differential equations supply that rule, and Physiome, CellML, BioModels, and HumMod provide reusable quantitative models.[10–13] Their assembly across organs requires compatible variables, equations, and parameters.

Earlier physiomaps represent physiological processes and participants, and Chalkboard traces qualitative perturbations through their networks.[14,15] Toonsi et al. connect knowledge graphs to causal and probabilistic semantics and use graph-guided mediation to analyze drug–disease effects in population data.[16] PhysioMap develops these ideas for quantitative physiological traits, relation-specific structural constraints, and qualitative equilibrium responses in cyclic systems.

Entity–quality ontology patterns define traits using anatomy, chemicals, proteins, processes, and qualities.[17–19] Web Ontology Language (OWL) reasoning classifies these definitions and entails relations among them. To compute interventions, these logical relations require an explicit causal interpretation.[20] Structural causal models (SCMs) supply such an interpretation: a perfect intervention replaces the equation for the intervened variable and retains the other equations.[21,22] Extensions to cyclic systems accommodate physiological feedback.[23,24] The representation must also distinguish causal influence from production, constitution, quantitative identity, and modulation, because these relations constrain different mathematical objects.

We collected physiological statements from textbooks, quantitative models, curated resources, and expert curation; classified their relation types; and encoded them in an OWL knowledge base. A versioned projection transforms entailed patterns into a typed causal knowledge graph (CKG) that constrains a class of quantitative SCMs. Derivative signs provide a qualitative abstraction of those constraints. We implemented a feedback-aware solver for first-order responses and a separate modulation query. We evaluate forward prediction of directional phenotypes and abduction of single-lesion hypotheses in rare metabolic disease, and use focused semantic checks to assess how the typed representation handles intervention.

Data and code availability. The knowledge base, projection registry, software, evaluation inputs, and expert review are available at `https://github.com/bio-ontology-research-group/physiomap/`. Supplementary material is hosted at `https://github.com/bio-ontology-research-group/physiomap/releases/download/v1.1.2/supplement.pdf`. An interactive map and interven-

tion interface is available at `https://bio2vec.net/physiomap/`. Software uses BSD 3-Clause; the map, benchmark data, and documentation use CC BY 4.0.

2. Methods

2.1. *Ontology and trait semantics*

We represent PhysioMap as an OWL 2 terminological knowledge base (TBox) $\mathcal{K}$ with standard OWL semantics.[25] Its upper level distinguishes continuants, processes, and traits. Imported modules provide classes for chemicals, proteins, cells, anatomy, processes, and qualities from Chemical Entities of Biological Interest (ChEBI), the Protein Ontology (PR), the Cell Ontology (CL), Uberon, the Gene Ontology (GO), and the ontology of phenotypic qualities (PATO).[17,26–30]

We use the entity–quality phenotype pattern to define a trait as a quality borne by an entity or process in a biological context.[18,19] A named trait T_i has the necessary conditions

$$\mathcal{K} \models T_i \sqsubseteq \mathsf{Trait} \sqcap \exists\, \mathsf{hasPart}.\big(E_i \sqcap \exists\, \kappa_i.C_i \sqcap \exists\, \mathsf{hasQuality}.Q_i^E\big), \qquad \kappa_i = \begin{cases} \mathsf{contextPartOf} & E_i \text{ continuant} \\ \mathsf{occursIn} & E_i \text{ process,} \end{cases} \tag{1}$$

where E_i is the characterized entity or process, C_i its recorded context, and Q_i^E an entity-specific subclass of the PATO quality Q_i. We omit the context conjunct when no context is recorded. The relation $\mathsf{contextPartOf}$ localizes a continuant, whereas $\mathsf{occursIn}$ localizes a process and propagates along parthood ($\mathsf{occursIn} \circ \mathsf{partOf} \sqsubseteq \mathsf{occursIn}$). The PATO hierarchy determines whether a quality requires a continuant or process bearer.

Parthood connects trait definitions across scales. For composable qualities, the reasoner groups traits by their bearers or their bearers' parts. For example, plasma volume, erythrocyte volume, and blood volume share the group *volume of blood or its parts*. Quantitative composition is specified separately through the identities and constitutive relations below. Extensive quantities such as volume admit these grouping rules; intensive quantities such as pressure require their own measurement model.

Causal axioms describe a type-level capacity. We encode that capacity with the collection pattern for biological function.[31] Each trait class T has an associated collection T^{all}, with $T \sqsubseteq \exists\, \mathsf{memberOf}.T^{\text{all}}$ and $T^{\text{all}} \sqsubseteq \exists\, \mathsf{hasMember}.T$. A causal axiom has the form $T_t^{\text{all}} \sqsubseteq \exists\, \mathsf{hasMember}.(\exists\, \mathsf{causedBy}.T_s)$: some member of the target collection is caused by an instance of the source. These existential axioms remain in the OWL 2 EL profile. Production and modulation use corresponding patterns, with modulation qualifying the causal influence whose slope it changes. Suppl. Sec. S1.2 gives the complete collection axioms. The projection supplies their interventional interpretation.

2.2. *Projection from ontology to a causal knowledge graph*

We next project the OWL knowledge base into a causal knowledge graph (CKG), that is, a knowledge graph with an explicitly interpreted causal edge type.[16] Let V be its named trait classes and $E_c \subseteq V \times V$ its causal-influence edges. The CKG also contains the typed production, constitution, quantitative-identity, and modulation relation instances defined below. A binary

relation instance $T_s \to T_t$ labeled r is present whenever $\mathcal{K}$ entails a relational pattern that the finite, versioned registry Π registers for relation type r. Each pattern $\rho \in \Pi$ defines an arity, direction, relation type, OWL class-expression premise φ_ρ, and the relation instance produced when that premise is entailed:

$$\mathcal{K} \models \varphi_\rho(T_s, T_t) \quad \Longrightarrow \quad T_s \xrightarrow{\rho} T_t. \tag{2}$$

Higher-arity patterns analogously produce typed hyperedges. The causal pattern, for example, has $\varphi_\rho(T_s, T_t) = \big(T_t^{\text{all}} \sqsubseteq \exists\, \mathsf{hasMember}.(\exists\, \mathsf{causedBy}.T_s)\big)$, so the entailment holding of ACE activity and angiotensin II yields the causal edge between those two traits.

Projection follows entailments.[32,33]

2.3. *Structural causal model semantics for typed relations*

The projection targets a class $\mathfrak{M}(\mathcal{K}, \Pi)$ of quantitative structural causal models. Trait T_i denotes an endogenous random variable X_i, and x_i a realized value. Exogenous random variables U_i enter the differentiable dynamics $\dot{x}_i = F_i(x_i, x_{\text{pa}(i)}, U_i)$. We assume a locally unique equilibrium x^* with a Hurwitz Jacobian and local self-regulation $\partial F_i / \partial x_i < 0$. A perfect intervention $\text{do}(x_A = \theta_A)$ replaces the mechanism or defining equation of each x_a by $x_a = \theta_a$, removes its incoming functional dependencies, and leaves all other equations unchanged.[21,34,35] Relation typing matters because it determines whether an intervention replaces a causal mechanism, a constitutive constraint, or a quantitative identity.

To define a direct effect, we let x_t re-equilibrate after an intervention on x_s while keeping every other variable at its reference value. Differentiating the target equilibrium equation and using negative self-regulation gives the direct-effect sign identity:

$$D_{st} := \left.\frac{dx_t^*}{d\theta}\right|_{\text{do}(x_s=\theta,\, x_j = x_j^*\ (j \neq s,t))} = -\,\frac{\partial F_t/\partial x_s}{\partial F_t/\partial x_t}, \qquad \text{sign}(D_{st}) = \text{sign}\left(\frac{\partial F_t}{\partial x_s}\right). \tag{3}$$

For a multivariate function, σ_j denotes the sign of its partial derivative with respect to argument j. Table 1 gives the quantitative constraint for each first-order relation type, its derivative-sign abstraction, and a PhysioMap example.

Table 1. Structural causal model semantics, derivative-sign abstractions, and examples for the first-order relation types. D_{st} is the direct effect in Eq. (3).

Relation	Quantitative constraint and derivative-sign abstraction	PhysioMap example
Causal influence	The target mechanism F_t depends non-constantly on x_s; $\sigma_{st} = \mathrm{sign}(\partial F_t/\partial x_s) = \mathrm{sign}(D_{st})$.	More plasma ACE directly raises plasma angiotensin II: $D_{\mathrm{ACE,AngII}} > 0$.
Production	$F_t = R_t + \epsilon_{pt} q_{pt}(x_p)$, with $q'_{pt} > 0$ and $\epsilon_{pt} \in \{-1, +1\}$; $\sigma_{pt} = \epsilon_{pt} = \mathrm{sign}(D_{pt})$.	A higher adrenal aldosterone-secretion rate raises plasma aldosterone.
Constitution	A whole is determined by its constituents: $X_w = C_w(X_{p_1}, \ldots, X_{p_r})$, with $\mathrm{sign}(\partial C_w/\partial x_{p_j}) = \sigma_j$ for $j = 1, \ldots, r$.	Adipose-tissue mass contributes positively to total body weight.
Quantitative identity	A named random variable is exactly a function of its arguments: $X_q = Q_q(X_{a_1}, \ldots, X_{a_\ell})$, with $\mathrm{sign}(\partial Q_q/\partial x_{a_j}) = \sigma_j$ for $j = 1, \ldots, \ell$.	Mean arterial pressure = cardiac output × total peripheral resistance; both partial derivatives are positive.

Sign annotations belong to a derived abstraction of the quantitative constraints. The abstract value $\sigma = ?$ denotes non-constant direct dependence with an unresolved derivative sign. Constitution and quantitative identity are directed: an intervention on the result replaces its defining equation while retaining the equations of its arguments.

A modulation $(m; (s, t), \mu)$, where μ is the mixed-derivative sign, is a (node, causal-influence edge) tuple in $\mathcal{H} \subseteq V \times E_c$. Its semantics is a paired intervention on modulator and source,

$$\widetilde{F}_t(\theta, \eta) := F_t\,|_{\mathrm{do}(x_s=\theta,\, x_m=\eta)}, \qquad \mathrm{sign}\left(\frac{\partial^2 \widetilde{F}_t}{\partial\eta\,\partial\theta}\right) = \mu. \tag{4}$$

With other arguments held fixed, the sign states whether increasing x_m makes the mechanism slope $\partial F_t/\partial x_s$ more or less positive. For example, cortisol increases the positive slope of the peripheral-resistance mechanism with respect to norepinephrine. Modulation is retained in $\mathcal{H}$ and supports a separate gain query. The first-order experiments use causal influence, production, quantitative identity, and constitution. A first-order effect of a modulator enters through an independently authored causal influence. The gain query combines the modulator's response with the mixed-derivative sign to determine how the signed slope changes. Suppl. Secs. S9.2 and S9.3 specify the second-order and population interpretations.

2.4. *First-order derivative-sign abstraction and feedback-aware prediction*

The quantitative SCM supplies the semantics; the PhysioMap solver reads only its first-order derivative-sign abstraction. The direct effects above keep the rest of the system fixed. Forward prediction instead asks for the total effect after every other variable re-equilibrates. At the

assumed equilibrium $F(x^*, \theta) = 0$, the total effect is the comparative static

$$\frac{dx^*}{d\theta} = -\left(\frac{\partial F}{\partial x}\right)^{-1} \frac{\partial F}{\partial \theta}. \tag{5}$$

The inverse Jacobian couples all variables in a feedback component. The signs of its entries may determine a response direction, or unavailable derivative magnitudes may leave it unresolved.[23,24,36] The PhysioMap solver returns `?` in the latter case.

The relation types constrain the SCM differently. The PhysioMap solver projects the first-order derivative constraints for causal influence, production, and quantitative identity into a signed dependency graph $G_\partial = (V, E_\partial)$, whose signed adjacency pattern defines one signed Jacobian. The PhysioMap solver represents constitution through a separate constitutive dependency graph $G_\kappa = (V, C)$ on the same vertices, with $(p_j, w) \in C$ labeled σ_j for each constituent in Table 1; adipose-tissue mass and lean-body mass each constitute total body weight. The constitutive dependency graph is acyclic because scales are strictly hierarchical, so no trait constitutes one of its own constituents, and determination is transferred along it in topological order by

$$\operatorname{sign}\left(\frac{dx_w^*}{d\theta}\right) = \bigoplus_{(p_j, w) \in C} \sigma_j \cdot \operatorname{sign}\left(\frac{dx_{p_j}^*}{d\theta}\right), \tag{6}$$

where $\cdot$ and $\bigoplus$ are the sign product and parallel combination defined below, and a constituted variable that also carries a causal sign combines the two the same way. Determination proceeds from constituents to the whole. The signed dependency graph can contain cycles, so computing downstream effects must account for feedback.

After intervention, we remove the incoming dependencies of the target from G_∂, yielding $G_\partial^{\mathrm{do}(k)}$, and decompose this graph into strongly connected components and process their acyclic condensation graph in topological order. A singleton uses sign algebra over $\{+, -, 0, ?\}$ with $+ \cdot + = - \cdot - = +$, $+ \cdot - = -$, and 0 annihilating products. An unknown factor absorbs a non-zero factor. The sign of a chain is the product of its edge signs. For parallel contributions, 0 is neutral, agreeing non-zero signs retain that sign, and opposing or unknown contributions give $?$. For a feedback component S, let J_S be its Jacobian sign pattern and b_S its signed forcing. The implicit function theorem and Cramer's rule give

$$\operatorname{sign}\left(\frac{dx_i^*}{d\theta}\right) = \operatorname{sign}\left(\frac{\det J_S^{(i)}}{\det J_S}\right), \qquad J_S^{(i)} = J_S \text{ with column } i \text{ replaced by } -b_S. \tag{7}$$

For a component with m variables, differentiating its equilibrium equations gives $J_S z = -b_S$. Cramer's rule gives Eq. (7). A real Hurwitz matrix has determinant sign $(-1)^m$; real eigenvalues contribute negative factors and conjugate pairs contribute positive products. Sound sign addition and multiplication ensure that a definite numerator sign holds for every compatible stable realization. Applying these operations in component order extends the argument to upstream forcing.[37–39] Suppl. Sec. S12.2 gives the full proof.

The exact component algorithm handles up to 16 traits, bounding each memoized determinant expansion by 2^{16} subproblems. Larger components use a conservative fixed-point

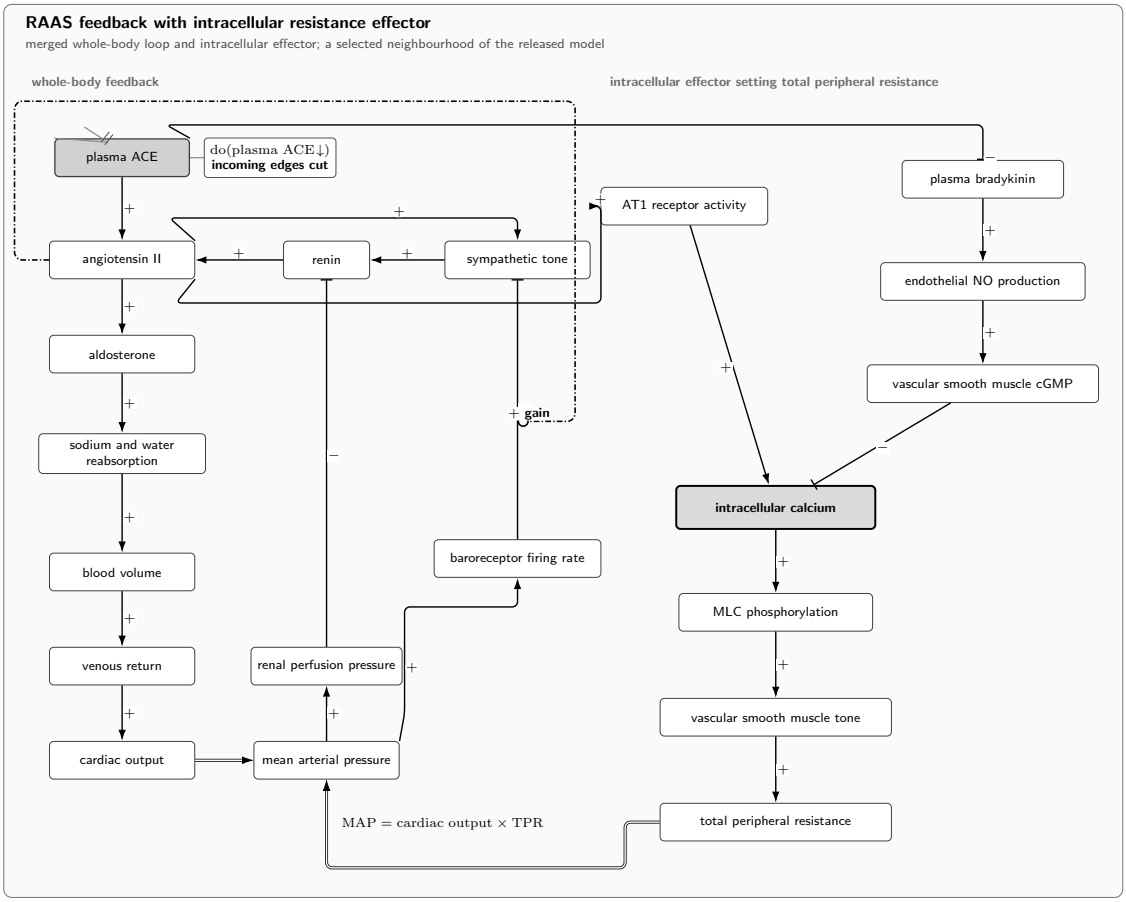


Fig. 1. A selected PhysioMap neighbourhood merges whole-body renin–angiotensin–aldosterone system (RAAS) feedback with an intracellular resistance effector. Pointed arrows are positive causal influences; bars are negative influences. These glyphs show the corresponding SCM derivative signs. The double line denotes the quantitative identity mean arterial pressure (MAP) = cardiac output × total peripheral resistance (TPR), and the dash-dot line denotes modulation of the inhibitory baroreceptor-to-sympathetic influence. The intervention on plasma angiotensin-converting enzyme removes incoming causal influences. AT1, angiotensin II type 1 receptor; NO, nitric oxide; cGMP, cyclic guanosine monophosphate; MLC, myosin light chain.

algorithm. Both algorithms return ? when their sign calculation remains unresolved. The guarantee is conditional on a locally unique stable equilibrium and negative self-regulation; the software takes these as model assumptions.

Figure 1 shows ACE inhibition perturbing volume and vascular-resistance pathways coupled by renal and baroreflex feedback. An intracellular arm connects receptor signals to smooth-muscle tone; typed relations distinguish causal influences, an identity, and modulation.

2.5. *Construction and content validation*

We assembled PhysioMap by direct expert curation, extraction from quantitative models and curated databases, and checked language-model proposals from textbooks. The cardiovascular

core used 17 Guyton BioModels modules and CellML integrators. Available equations supplied derivatives, while directional sources supplied constraints without complete functions. SIGNOR, the Adverse Outcome Pathway Wiki (AOP-Wiki), and direct literature curation provided additional content. Suppl. Sec. S15.1 lists the sources and acquisition metadata.

Following the entity–quality convention,[17–19] each candidate trait received an entity identifier, PATO determinable, optional context, stable label, and scale. We resolved identifiers against pinned releases, classified measurement kind as extensive, ratio, rate, or intensive, checked that bearer and quality categories matched, and reconciled duplicates. Only composable measurement kinds admit part-to-whole rules.[40,41] Unresolved traits remained explicitly primitive.

We assigned each relation to one of the five semantic types. Release checks covered schema, references, identifiers, bearers, constitution, relation typing, and regression against a frozen behavioral baseline. These checks assess formal consistency and preservation of expected behavior.

For cross-type content validation, P.N.S. reviewed a fixed-seed stratified sample of 83 v1.1.1 relations covering all five types. We interpreted `TRUE` as accepted, `FALSE?` as flagged for further review or investigation, and `FALSE` as rejected. We report the judgments for each sampled relation type. Suppl. Sec. S15.3 gives the protocol, sampling strata, and archive.

2.6. *Evidence and provenance model*

Evidence and provenance govern admission and document construction. Type, arguments, and context determine a relation's semantics, and its sign constrains a derivative. The solver reads relation types, arguments, contexts, and derivative signs.

For causal-influence candidates, the evidence model admits six classes. Perturbation, pharmacological intervention, genetic loss or gain of function, and Mendelian randomization form the interventional group because they use a manipulation or valid instrument. Mechanistic-model evidence derives a sign from a curated model, while curated-mechanistic evidence records a curated functional or mechanistic account; these form the mechanistic group. Admission requires evidence of the stated intervention or functional mechanism. Other relation types use type-specific evidence, and support awaiting assignment to a controlled evidence class remains unclassified. Each stable relation identifier links to support and a projection trace.

2.7. *Evaluation of prediction and abduction*

We evaluated forward prediction against the Human Phenotype Ontology (HPO) and its gene-to-phenotype annotation corpus (HPOA) from release 2026-02-16. For each gene entering the evaluation, we specified one or more primary lesion variables from gene-product identity and the disease mechanism. Suppl. Sec. S15.4 gives the mapping rule and links the complete versioned specification.

HPO phenotype classes follow an entity–quality (EQ) pattern, in which the entity E identifies the affected entity or process and the quality Q is drawn from PATO.[17,42] This pattern lets us identify a directional HPO class when Q states an increase or decrease relative to normal.

For example, hypoglycemia pairs blood glucose with decreased concentration (PATO:0001163). We encode the PATO distinction between increased quality (PATO:0002300) and decreased quality (PATO:0002301) as $s = \texttt{+}$ and $s = \texttt{−}$, respectively.

We propagated HPOA annotations only upward through the HPO subclass hierarchy. Let A_g be the HPO classes directly annotated to gene g, let HPO denote the HPO ontology, and let $M(a) = (X_i, s)$ map a directional HPO class a to PhysioMap variable X_i and direction s. The reference directions for X_i were

$$R_g(X_i) = \{s \in \{\texttt{+}, \texttt{−}\} \mid \exists h \in A_g \; \exists a : \mathsf{HPO} \models h \sqsubseteq a \;\wedge\; M(a) = (X_i, s)\}. \tag{8}$$

Therefore, a specific annotation h contributed the mapping of an ancestor a, never that of a descendant. We retained a gene–variable direction only when $R_g(X_i)$ was a singleton. We blocked propagation for two deoxycortisol terms whose asserted HPO ancestry would otherwise map them to cortisol, although neither term denotes cortisol (Suppl. Sec. S15.4). We did not score the intervened variable.

We scored all unambiguous downstream reference directions. To assess literature overlap between HPOA and PhysioMap, we repeated the analysis after excluding the 15 genes named in phenotype-directed curation fragments. Suppl. Sec. S15.4 gives the mapping sources and the overlap specification.

For each lesion, we applied the PhysioMap solver once and compared every downstream sign with the fixed reference pairs. Let C, W, and A denote correct, wrong, and abstaining pairs; missing and `?` predictions both count in A. We report precision $= C/(C + W)$ and coverage $= (C + W)/(C + W + A)$.

We compared two alternative inference rules on the same axioms and interventions. These rules differ only in how they combine directions, so the comparison evaluates the inference rules on shared physiological content. The shortest-path rule receives a lesion and an endpoint, selects one shortest directed path, and multiplies its edge signs. Signed diffusion receives a lesion and assigns a score to every downstream trait. Let S be the signed, column-normalized adjacency matrix, so a node with d explicitly signed outgoing arcs contributes $+1/d$ or $-1/d$ to each successor. We omit unknown-sign arcs before computing reachability and normalizing weights. Let e encode the signed intervention. We solved

$$r = (1 - \alpha)(I - \alpha S)^{-1} e, \tag{9}$$

taking $\mathrm{sign}(r_i)$ as the prediction and $|r_i|$ as its confidence. We set $\alpha = 0.85$ and swept every distinct confidence observed on the evaluation pairs to produce the precision–coverage curve.

We tested whether abstention concentrates shortest-path errors by permuting abstention labels within genes, conditional on each gene's path-error and abstention counts among path-answered pairs. The statistic counts errors assigned to abstained pairs; we report its exact one-sided upper-tail probability (Suppl. Sec. S15.4).

For the relation-typing ablation, we cumulatively added production, quantitative identity, and constitution to causal influence and repeated both tasks. Modulation uses a separate query. We checked typed intervention semantics with small models whose expected responses follow directly from their equations. The checks cover part-to-whole determination, production, quantitative identities, and the separate modulation query. Suppl. Sec. S15.7 gives the

models, expectations, and executable checks.

For abduction, we precomputed determinate phenotype signs for a fixed closed pool of single-lesion hypotheses and compared them with each observed directional profile. Agreement added one, contradiction subtracted one, and abstention scored zero. We ranked by net agreement, then fewer contradictions, then more agreements. Rank is one plus the number of strictly better candidates, so top-k, median rank, and mean reciprocal rank use the best position within a tie; unique top-1 requires no tie. We excluded the intervened trait and retained the true lesion in the pool.

Experiments used PhysioMap v1.1.1 (analysis v1.1.2). Suppl. Sec. S15.4 gives input checksums and file names; Suppl. Sec. S15.6 gives the reproduction commands.

3. Results

3.1. *Construction, projection, and expert validation of PhysioMap*

We built PhysioMap v1.1.1 as an OWL knowledge base containing 1,699 contextualized physiological traits and 2,270 causal-influence, 85 production, 4 constitution, 9 quantitative-identity, and 19 modulation axioms. Projection preserves evidence and provenance (Section 2.6).

Projection produced 2,387 typed relation instances as edges or hyperedges; the largest strongly connected component contains 213 traits.

The cross-type expert review accepted 69 of 83 sampled relations, flagged 12 for further investigation, and rejected 2 (Suppl. Table S1). The flagged category denotes an unresolved review judgment; both rejected relations were causal. These judgments identify content for correction or further investigation.

3.2. *Forward prediction of directional endophenotypes*

We tested how physiological dependencies recover the direction of measurements associated with rare metabolic disease. The HPO mapping provided downstream directions for 167 of 183 matched genes; the other 16 had a mapped direction only for the intervened variable. Across 870 downstream pairs, the solver returned 175 determinate directions: 171 agreed with the reference and 4 disagreed. Agreement was 97.7% at 20.1% coverage (Figure 2b).

Shortest paths returned 379 directions, with 292 agreements and 87 disagreements (77.0% agreement). Figure 2b compares signed diffusion across thresholds on the same reference and interventions. The overlap analysis retained 155 determinate predictions, of which 151 agreed with the reference.

A blinded audit of 60 sampled concordant pairs supported 57 reference directions and judged three context-dependent; all four discordant pairs supported PhysioMap (Suppl. Sec. S15.5). The primary result remains unadjusted at 97.7% agreement and 20.1% coverage.

PhysioMap and shortest paths returned identical directions on the 175 shared answered pairs, with 2.3% error. On the 204 additional pairs answered by shortest paths, PhysioMap abstained and the path error rate was 40.7%. The within-gene test gave $P = 2.05 \times 10^{-7}$: abstention concentrated cases difficult for path inference, explaining the precision–coverage trade-off.

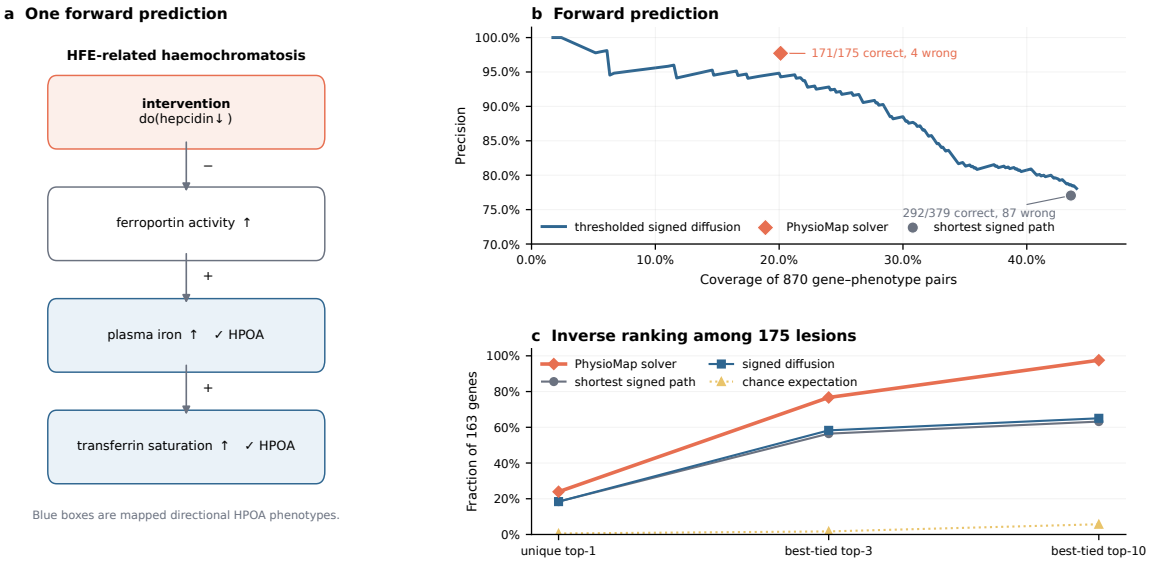


Fig. 2. Evaluation of prediction and abduction. (a) Relations supporting an HFE-associated hereditary haemochromatosis prediction; only downstream phenotypes enter scoring. (b) Precision versus coverage for thresholded signed diffusion, with operating points and exact counts for the PhysioMap solver and shortest signed path. (c) Abductive ranking within 175 known lesions for the three inference rules and chance expectation. Top-3 and top-10 use the best rank within ties, and a unique top-1 requires no tied candidate; mean reciprocal rank for the PhysioMap solver is 0.698.

3.3. *Abduction ranks candidate lesions from phenotype profiles*

The inverse benchmark excludes ARSA, CTNS, KYNU, and PKLR, which each have two primary lesion variables, leaving 163 genes. Each true lesion belongs to the closed pool of 175 hypotheses. Only the inference rule varies across comparisons (Table 2).

Table 2. Inverse lesion ranking over 175 candidates for 163 genes. Scoring, tie handling, and metrics are identical in every row; only the rule producing a candidate's predicted profile differs. Top-3 and top-10 use the best rank within ties, and a unique top-1 requires no tied candidate.

inference rule	unique top-1	top-3	top-10	MRR
PhysioMap solver	39	125	159	0.698
shortest signed path	30	92	103	0.528
signed diffusion	30	95	106	0.545
chance	1	3	9	0.033

The solver ranked the true lesion within the best-tied top three for 125 genes and uniquely first for 39 (Table 2).

Adding the first-order relation types gave no aggregate improvement in forward coverage or inverse ranking (Suppl. Table S3). The semantic checks verify typed intervention behavior

in small models; broader non-causal content would test its predictive contribution.

4. Discussion

PhysioMap's ontology defines physiological variables through shared entities, qualities, and contexts, providing a basis for interoperability across models and data sources. The same entity–quality pattern connects these variables to phenotype ontologies:[18,19] we used it to map directional HPO phenotypes to physiological changes. This connection allows phenotype annotations to inform both evaluation of predicted responses and inference of candidate lesions. The typed projection then gives the relations between variables their intervention semantics.

The solver extends qualitative inference in earlier physiomaps and Chalkboard[14,15] to responses coupled by cycles and feedback. Its derivative-sign abstraction supports predictions around a locally unique stable equilibrium with negative self-regulation. Opposing contributions can leave responses dependent on unknown derivative magnitudes. Quantitative systems-biology models can supply these magnitudes, resolve feedback responses, and assess stability numerically. PhysioMap could support model assembly from Physiome, CellML, and BioModels[10–12] by identifying shared variables and constraints on their dependencies; integration requires compatible units, contexts, and parameters. Regenerative mechanisms and transitions between equilibria require explicit dynamics, while coupled second-order modulation responses require further models and experiments.

Current forward coverage is 20.1% and depends on represented physiology and lesion mapping. Abstention concentrated cases with higher shortest-path error. PhysioMap and HPOA can share literature; the overlap analysis assesses phenotype-directed curation, while independent perturbation data would test transfer. Inverse ranking uses best-tied positions in a closed pool containing the true lesion; prospective open-set evaluation would test lesions outside that pool. Broader non-causal content would test the predictive contribution of the additional relation types. Content review used one co-author and unequal sampling rates; independent representative review would estimate release-wide quality. The reference audit concealed predictions and earlier judgments, but its co-author reviewer had previously assessed the four discordant pairs. Independent review would test reproducibility.

PhysioMap already contains a feedback component of 213 traits. Exact sign inference is restricted to small components; conservative inference on larger components leaves additional responses unresolved. Extending the map across human physiology will increase both curation demands and the complexity of dependencies across organs and scales. Reusable organ models could support expansion, but their shared variables and feedback require reconciliation. Ontology alignment can help identify these connections; unresolved context-specific conflicts currently yield unknown signs. Expansion should therefore be evaluated through predictive coverage, content quality, and computational cost.

By connecting phenotype observations to physiological mechanisms, PhysioMap provides a basis for reasoning about interventions and could guide the assembly of quantitative models across organs and scales.

References

1. L. Licata, P. Lo Surdo, M. Iannuccelli, A. Palma, E. Micarelli, L. Perfetto, D. Peluso, A. Calderone, L. Castagnoli and G. Cesareni, SIGNOR 2.0, the SIGnaling network open resource 2.0: 2019 update, *Nucleic Acids Research* **48**, D504 (2020).
2. D. Türei, A. Valdeolivas, L. Gul, N. Palacio-Escat, M. Klein, O. Ivanova, M. Ölbei, A. Gábor, F. Theis, D. Módos *et al.*, Integrated intra- and intercellular signaling knowledge for multicellular omics analysis, *Molecular Systems Biology* **17** (2021).
3. B. M. Gyori, J. A. Bachman, K. Subramanian, J. L. Muhlich, L. Galescu and P. K. Sorger, From word models to executable models of signaling networks using automated assembly, *Molecular Systems Biology* **13** (2017).
4. M. Milacic, D. Beavers, P. Conley, C. Gong, M. Gillespie, J. Griss, R. Haw, B. Jassal, L. Matthews, B. May *et al.*, The Reactome pathway knowledgebase 2024, *Nucleic Acids Research* **52**, p. D672–D678 (2024).
5. M. Kanehisa, M. Furumichi, Y. Sato, Y. Matsuura and M. Ishiguro-Watanabe, KEGG: biological systems database as a model of the real world, *Nucleic Acids Research* **53**, p. D672–D677 (2025).
6. A.-L. Barabási, N. Gulbahce and J. Loscalzo, Network medicine: a network-based approach to human disease, *Nature Reviews Genetics* **12**, p. 56–68 (2011).
7. L. Cowen, T. Ideker, B. J. Raphael and R. Sharan, Network propagation: a universal amplifier of genetic associations, *Nature Reviews Genetics* **18**, p. 551–562 (2017).
8. M. Gebser, T. Schaub, S. Thiele and P. Veber, Detecting inconsistencies in large biological networks with answer set programming, *Theory and Practice of Logic Programming* **11**, p. 323–360 (2011).
9. A. Liu, P. Trairatphisan, E. Gjerga, A. Didangelos, J. Barratt and J. Saez-Rodriguez, From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL, *npj Systems Biology and Applications* **5** (2019).
10. P. J. Hunter and T. K. Borg, Integration from proteins to organs: the Physiome project, *Nature Reviews Molecular Cell Biology* **4**, p. 237–243 (2003).
11. C. M. Lloyd, J. R. Lawson, P. J. Hunter and P. F. Nielsen, The CellML model repository, *Bioinformatics* **24**, p. 2122–2123 (2008).
12. R. S. Malik-Sheriff, M. Glont, T. V. N. Nguyen, K. Tiwari, M. G. Roberts, A. Xavier, M. T. Vu, J. Men, M. Maire, S. Kananathan *et al.*, BioModels: 15 years of sharing computational models in life science, *Nucleic Acids Research* **48**, D407 (2020).
13. R. L. Hester, A. J. Brown, L. Husband, R. Iliescu, D. Pruett, R. Summers and T. G. Coleman, HumMod: A modeling environment for the simulation of integrative human physiology, *Frontiers in Physiology* **2** (2011).
14. D. L. Cook, M. L. Neal, R. Hoehndorf, G. V. Gkoutos and J. H. Gennari, Representing physiological processes and their participants with PhysioMaps, *Journal of Biomedical Semantics* **4**, p. S2 (2013).
15. D. L. Cook, J. C. Wiley and J. H. Gennari, Chalkboard: ontology-based pathway modeling and qualitative inference of disease mechanisms, in *Pacific Symposium on Biocomputing 2007*, (World Scientific, 2007).
16. S. Toonsi, P. N. Schofield and R. Hoehndorf, Causal knowledge graph analysis identifies adverse drug effects, *Bioinformatics* **42** (2026).
17. G. V. Gkoutos, E. C. Green, A.-M. Mallon, J. M. Hancock and D. Davidson, Using ontologies to describe mouse phenotypes, *Genome Biology* **6** (2005).
18. R. Hoehndorf, A. Oellrich and D. Rebholz-Schuhmann, Interoperability between phenotype and anatomy ontologies, *Bioinformatics* **26**, 3112 (2010).
19. G. V. Gkoutos, P. N. Schofield and R. Hoehndorf, The anatomy of phenotype ontologies: prin-

ciples, properties and applications, *Briefings in Bioinformatics* **19**, p. 1008–1021 (2018).
20. A. Rector, R. Stevens and N. Drummond, What causes pneumonia? the case for a standard semantics for "may" in OWL, in *Proceedings of the 5th OWLED Workshop on OWL: Experiences and Directions*, , CEUR Workshop Proceedings Vol. 432 (RWTH Aachen University, 2008).
21. J. Pearl, *Causality: Models, Reasoning, and Inference* (Cambridge University Press, 2009).
22. J. Woodward, *Making Things Happen: A Theory of Causal Explanation* (Oxford University Press, 2003).
23. J. M. Mooij, D. Janzing and B. Schölkopf, From ordinary differential equations to structural causal models: the deterministic case, in *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI)*, (AUAI Press, 2013). arXiv:1304.7920.
24. S. Bongers, P. Forré, J. Peters and J. M. Mooij, Foundations of structural causal models with cycles and latent variables, *The Annals of Statistics* **49** (2021).
25. B. Motik, P. F. Patel-Schneider and B. Cuenca Grau, OWL 2 Web Ontology Language: Direct Semantics (Second Edition) W3C Recommendation, (2012).
26. J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes and C. Steinbeck, ChEBI in 2016: Improved services and an expanding collection of metabolites, *Nucleic Acids Research* **44**, D1214 (2016).
27. C. J. Mungall, C. Torniai, G. V. Gkoutos, S. E. Lewis and M. A. Haendel, Uberon, an integrative multi-species anatomy ontology, *Genome Biology* **13**, p. R5 (2012).
28. A. D. Diehl, T. F. Meehan, Y. M. Bradford, M. H. Brush, W. M. Dahdul, D. S. Dougall, Y. He, D. Osumi-Sutherland, A. Ruttenberg, S. Sarntivijai, C. E. Van Slyke, N. A. Vasilevsky, M. A. Haendel, J. A. Blake and C. J. Mungall, The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability, *Journal of Biomedical Semantics* **7**, p. 44 (2016).
29. The Gene Ontology Consortium, Gene ontology: tool for the unification of biology, *Nature Genetics* **25**, 25 (2000).
30. D. A. Natale, C. N. Arighi, J. A. Blake, C. J. Bult, K. R. Christie, J. Cowart, P. D'Eustachio, A. D. Diehl, H. J. Drabkin, O. Helfer *et al.*, Protein ontology: a controlled structured network of protein entities, *Nucleic Acids Research* **42**, D415 (2014).
31. R. Hoehndorf, L. Mencel, G. V. Gkoutos and P. N. Schofield, Large-scale reasoning over functions in biomedical ontologies, in *Formal Ontology in Information Systems: Proceedings of the 9th International Conference (FOIS 2016)*, , Frontiers in Artificial Intelligence and Applications Vol. 283 (IOS Press, 2016).
32. R. Hoehndorf, A. Oellrich, M. Dumontier, J. Kelso, D. Rebholz-Schuhmann and H. Herre, Relations as patterns: bridging the gap between OBO and OWL, *BMC Bioinformatics* **11**, p. 441 (2010).
33. M. A. Rodríguez-García and R. Hoehndorf, Inferring ontology graph structures using OWL reasoning, *BMC Bioinformatics* **19**, p. 7 (2018).
34. S. Beckers and J. Y. Halpern, Abstracting causal models, in *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, (AAAI Press, 2019).
35. P. K. Rubenstein, S. Weichwald, S. Bongers, J. M. Mooij, D. Janzing, M. Grosse-Wentrup and B. Schölkopf, Causal consistency of structural equation models, in *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*, (AUAI Press, 2017). arXiv:1707.00819.
36. P. A. Samuelson, *Foundations of Economic Analysis* (Harvard University Press, Cambridge, MA, 1947).
37. K. Lancaster, The scope of qualitative economics, *The Review of Economic Studies* **29**, p. 99 (1962).
38. J. Quirk and R. Ruppert, Qualitative economics and the stability of equilibrium, *The Review of Economic Studies* **32**, p. 311 (1965).
39. J. Maybee and J. Quirk, Qualitative problems in matrix theory, *SIAM Review* **11**, p. 30–51

(1969).
40. P. Suppes, A set of independent axioms for extensive quantities, *Portugaliae Mathematica* **10**, 163 (1951).
41. D. H. Krantz, R. D. Luce, P. Suppes and A. Tversky, *Foundations of Measurement, Volume I: Additive and Polynomial Representations* (Academic Press, New York, 1971).
42. S. Köhler, M. Gargano, N. Matentzoglu, L. C. Carmody, D. Lewis-Smith, N. A. Vasilevsky, D. Danis, G. Balagura, G. Baynam, A. M. Brower *et al.*, The human phenotype ontology in 2021, *Nucleic Acids Research* **49**, p. D1207–D1217 (2021).

KG-BIOPORTAL: GRAPH-NATIVE ACCESS TO BIOMEDICAL ONTOLOGIES FOR NETWORK AND SYSTEMS BIOLOGY

J. Harry Caufield*, Nomi L. Harris, and Christopher J. Mungall

*Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA *Email: jhc@lbl.gov*

Jennifer Vendetti, Michael V. Dorf, Alex Skrenchuk, and Mark A. Musen

Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA 94305, USA

BioPortal is the world's most comprehensive repository of biomedical ontologies, hosting 1,549 ontologies with more than 15 million classes and over 100 million cross-ontology mappings. Network-based analysis in biology and medicine increasingly relies on knowledge graphs (KGs) that combine experimental observations with curated knowledge, but the ontological knowledge on which such graphs depend has been difficult to obtain in graph-native form: it is spread across hundreds of separately maintained artifacts, serialized in different languages, and rarely accompanied by consistent node and edge typing or version-level provenance. We describe recent work that makes BioPortal knowledge-graph-friendly.

KG-Bioportal consists of an open-source workflow, the software that runs it, and the derived data collection it publishes. It converts BioPortal ontologies into standard KGX-format node and edge lists, categorizes entities using the Biolink Model, records the source and submission version of every ontology, and publishes an independently downloadable graph for each one. Where an ontology imports others, a second graph with the import closure merged in is published beside it. The pipeline runs monthly on public infrastructure; the current collection covers 1,184 ontologies, 8,259,694 nodes, and 15,924,223 edges. For every ontology the software could not convert, it reports why.

Because graphs are released per ontology rather than as a monolith, users can compose made-for-purpose graphs containing only necessary domains. The median published graph has 412 nodes and 88% have fewer than 10,000 edges. We describe complementary improvements that support graph-based work: a machine-readable release index, and agent skills for retrieving and combining graphs, and a Model Context Protocol (MCP) server that lets AI agents query BioPortal without fabricating term identifiers. Finally we present a browser that indexes BioPortal-derived graphs alongside 161 community knowledge graphs, and our design for extending BioPortal to accept, search, and serve knowledge graphs directly.

Keywords: knowledge graphs; biomedical ontologies; data integration; semantic interoperability; BioPortal; Biolink Model; network analysis.

© 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.

1. Introduction

A great deal of modern biomedical analysis is, in its roots, network analysis. Gene regulatory network inference, signaling pathway reconstruction, multi-omics integration, and causal modeling all share a conceptual lineage with more applied research tasks: defining a cohort from clinical notes, detecting a pharmacovigilance signal, harmonizing agricultural or environmental measurements, or indexing a literature corpus. In every case the analyst must be able to assert that an entity recorded in one place is the same entity as, or stands in a defined relationship to, an entity recorded in another. A transcript identifier, a drug trade name, or a phenotype extracted from a clinical note are referring expressions whose referents must be reconciled before any network built from them will mean anything. Ontologies (standardized, computable representations of the entities in a discipline) do this reconciling work[1]. They supply the identifiers, the definitions, the synonym sets, and the hierarchical and non-hierarchical relationships that let heterogeneous measurements be placed in a shared frame of reference.

Increasingly, that knowledge is consumed by software rather than read by people. Graph machine learning methods are now routine in biomedical discovery, and libraries such as GRAPE[2] operate on typed node and edge lists rather than on ontology documents, as do the AI-powered agentic systems now applied to annotation and curation. Both require the same thing: domain knowledge already in graph form, with consistent types and traceable provenance.

BioPortal has served as the bioinformatics community's principal repository of ontology knowledge for nearly two decades[3,4]. Launched by the National Center for Biomedical Ontology in 2005, BioPortal now houses 1,549 biomedical ontologies, 1,182 of them public, which together contain 15,293,440 classes and more than 101 million cross-ontology mappings, spanning molecular biology, clinical terminology, pharmacology, agriculture, environmental science, and the social and behavioral sciences. It is embedded in the routine operation of widely used research systems including REDCap[5] and the CEDAR Workbench[6]. BioPortal is deliberately open: any registered user may contribute an ontology. The repository includes all ontologies from the OBO Foundry along with a large selection of UMLS vocabularies and several hundred ontologies uploaded directly by their authors.

Despite this breadth, the knowledge BioPortal holds has historically been tricky to use in graph-based analysis. The obstacles are practical, and they are largely inherited from how ontologies are authored and published rather than introduced by the repository: BioPortal is faithful to what its contributors submit, and what they submit is not graphs. Ontologies arrive in several serializations, including OWL in any of its RDF encodings, OBO format, SKOS thesauri, and vocabularies derived from the UMLS, and a graph builder must handle each. For the OWL portion of the repository, much of the biologically interesting structure is carried in class axioms

rather than in simple triples, so a naive conversion loses exactly the non-hierarchical relationships that network analysis wants; the statement that a structure is part of some organ, for example, is written as an existential restriction inside a logical definition rather than as a direct edge. Node and edge types are inconsistent across sources, which frustrates any downstream method that expects a schema. Provenance is usually absent from the converted product, so a graph assembled six months ago is difficult to reproduce. The consequence is that most groups building biomedical KGs write and maintain their own ingest code for each ontology they need, an effort duplicated across the field and one that quietly encodes different conversion decisions into each graph.

2. KG-Bioportal: Rendering Ontologies as Knowledge Graphs

KG-Bioportal is three things under a single name. First, it is a *workflow*: a sequence of inventory, download, conversion, typing, and publication steps applied to every public BioPortal ontology. Second, it is the *software* that runs that workflow: a Python package with a command-line interface, executed on public infrastructure on a monthly schedule and runnable locally by anyone. Finally, it includes the *data collection* the workflow produces: one downloadable, independently versioned graph per ontology, published as release assets and indexed by a machine-readable manifest and a public browser. It is not a database, a query service, or a merge tool. Combining graphs is a separate operation that users perform with standard tooling (**Section 2.5**). KG-Bioportal's input is the BioPortal REST API. Its output, per ontology, is an archive containing node and edge tables. Each transformation is tracked in a publicly-accessible index.

2.1. *Design goals*

We designed KG-Bioportal around four principles, each responding to an issue described in the Introduction section.

Graph-native output. Every ontology is emitted as a pair of node and edge lists in the tab-delimited Knowledge Graph Exchange (KGX) format, a representation for property graphs already used across the KG-Hub ecosystem[7]. The output is directly loadable by common graph tooling without further parsing of OWL.

Modularity and composability. Rather than publishing a single merged KG, we publish one graph per ontology and leave composition to the user. Each ontology thus becomes a unit that can be selected, pinned to a version, cited, swapped out, and combined independently of the others.

Versioning and provenance. Each graph is identified by the BioPortal submission number (i.e., the identifier BioPortal uses to distinguish different uploads) of the ontology it came from, and every node and edge carries the source it was derived from.

Reproducibility and openness. The workflow is open source, runs unattended on a schedule on public infrastructure, and publishes both its products and the statistics describing them.

2.2. *The transformation pipeline*

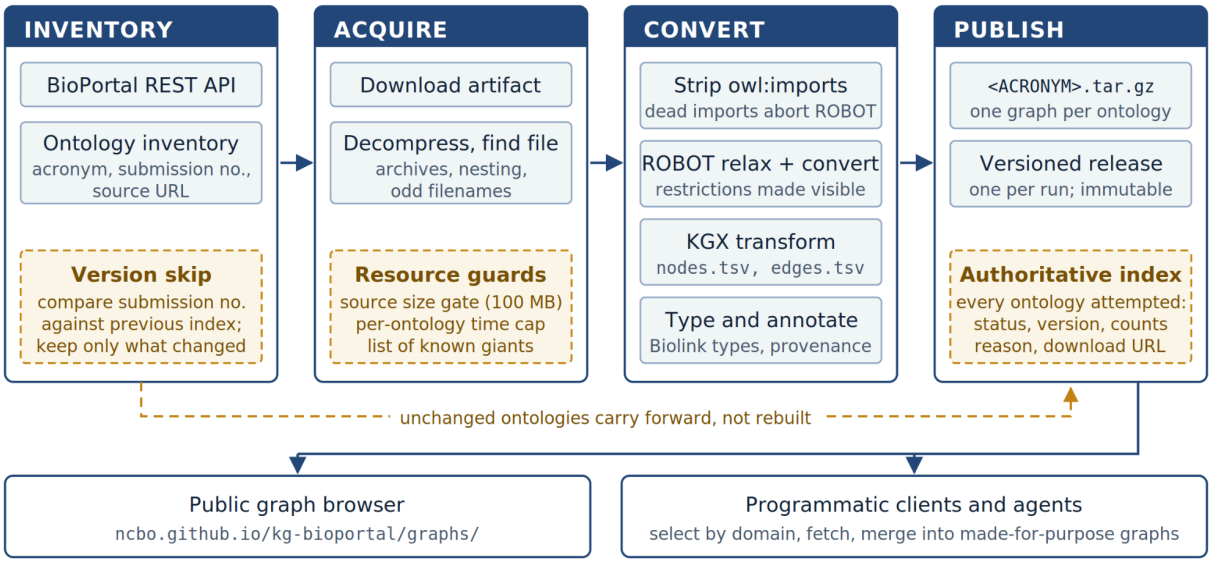


Fig. 1. The KG-Bioportal pipeline: inventory, acquisition, conversion, and publication, with the version-skip and resource-guard policies that bound each monthly run.

The pipeline workflow has four stages, illustrated in **Fig. 1** and described below. A scheduled run re-transforms only what changed and takes about ten minutes. A complete rebuild requires about two hours on twenty parallel runners.

The first stage retrieves the current inventory of ontologies via the BioPortal REST API, including each ontology's acronym, submission number, and download URL. Because each published graph records the submission it was built from, a scheduled run compares the new submission numbers against the previous index and re-transforms only the updated ontologies, carrying the rest forward untouched. This makes an unattended monthly build affordable.

The second stage downloads those ontologies. In practice this step carries most of the operational burden, because ontology files in an open repository are heterogeneous in ways that are difficult to anticipate: some arrive compressed, in several different archive formats; some archives contain a directory rather than a single file, or contain the ontology under an unexpected name; some submissions lack retrievable metadata altogether. The pipeline can detect and recover from these conditions so that one malformed ontology file will not halt the entire build process. Two resource guards apply here: a source-size gate (250 MB by default) and a per-ontology wall-clock cap, both configurable, which bound the cost of any single ontology on shared infrastructure.

The third stage performs the conversion. Each ontology is first processed with ROBOT[8], the standard command-line toolkit for ontology workflows. We apply ROBOT's *relax* operation, which does not add new relationships but makes existing ones visible to a converter that does no reasoning. Ontology authors commonly define a class with an equivalence axiom of the

genus-differentia form, stating that an X is a Y that is part of some Z; the existential restriction inside such a definition is invisible to a triple-level conversion. ‘Relax’ rewrites these into the weaker subclass form, asserting separately that X is a subclass of Y and that X is a subclass of the restriction *part of some Z*, and the subsequent conversion step renders the latter as an edge from X to Z labeled *part of*. Without this step a flattening conversion would yield little more than a subsumption tree, discarding the partonomic, regulatory, and participation relationships that encode most of the biology. The relaxed ontology is then transformed with KGX into node and edge tables, provenance fields recording the source ontology and submission are attached, and the resulting tables are written to a per-ontology archive.

Imports are handled in two ways, based on the OBO Foundry’s distinction between base and full ontology releases. Every ontology is published as a base graph such that the graph contains only what the ontology asserts and can be merged with other base graphs, avoiding duplicate nodes. When imports are declared and can be fetched, a full graph is published as well, with the import closure merged in. In the 2026-09-09 build, 263 of the 1,184 transformed ontologies declared imports, yielding 141 full graphs (120 failed due to unavailable imports; 2 of the full graphs exceeded the size cutoff). Nine ontologies contain only imports and are flagged accordingly in the KG-Bioportal index and pages.

KGX writes biolink:NamedThing on every node, because an OWL file says nothing about the Biolink Model[9]. The transform then assigns a more specific class wherever the graph's own evidence supports one, by four routes applied in order. A short table of seed terms lists classes whose Biolink meaning is not in doubt, such as GO's *biological_process* and MONDO's *disease*. Every class beneath a seed, along *subclass_of* or *skos:broader*, inherits its category. A category crosses an *exact_match* edge to a node that has none. Many ontologies hang off roots whose meaning is a fact about that ontology rather than about the term, such as an ICD chapter or an HGNC locus group. These roots were placed by reading each one and its subclasses in the published graph, one ontology at a time, with the assistance of an AI agent. Each placement is recorded in the repository with the reviewer, the date, the graph read, the subclass labels that served as evidence, and a confirmation field that stays empty until a maintainer has checked it. Roots that were read and refused are recorded with the reason. Nothing else is guessed: a node with no evidence stays *NamedThing*. Measured over the Sept. 9 build, the combined strategies assign a specific category to 52.2% of nodes. Every ontology's page in the browser reports the tally by route, with the disclosure that an AI agent assisted.

Products are published as release assets in the KG-Bioportal project repository. Hosting is organized as a set of releases plus one authoritative index: every run publishes its own release containing only what it built, and the index, republished in full each time, records for every ontology its status, name, version, submission number, node and edge counts, and a direct

download URL pointing at whichever release holds its artifact. The index is a small YAML file and is the intended entry point for programmatic users. Users who prefer to build graphs themselves (e.g., to pin a particular ontology version) can run the same pipeline locally against the BioPortal API.

2.3. *The resulting collection*

The collection is not one large graph stored in pieces, but a long tail of small, domain-specific graphs. As of the 2026-09-09 build, the collection comprises base graphs for 1,184 ontologies containing 8,259,694 nodes, representing ontology classes, and 15,924,223 edges, representing relationships among them. The size distribution matters: the median published graph has 412 nodes and 512 edges; 1,038 of the 1,184 graphs have fewer than 10,000 edges, and 1,139 have fewer than 100,000. The ten largest graphs hold 36% of all edges between them.

2.4. *What does not convert and why*

A build that quietly drops what it cannot handle may accomplish its purpose but would not support debugging. The index therefore carries an entry for every ontology the pipeline has attempted, successful or not, with a machine-readable reason for each failure, and the public browser (**Section 5**) gives every ontology graph a page whether or not an artifact exists behind it. **Table 1** provides details for the 2026-09-09 build. Of 1,303 ontologies attempted, 1,184 were converted. Of the 119 that were not, 31 were withheld by policy: 8 are available only under a license we do not hold, nearly all UMLS vocabularies, and are counted apart from failures because there is nothing to fix; 16 known giants, among them SNOMED CT, NCBI Taxonomy, RxNorm, MeSH, and NCIT, are skipped before download; and 7, including ChEBI, EFO, and FMA, exceed the 250 MB source-size gate. Of the 88 that failed, 78 failed on BioPortal's side, with no submission, no file, or no retrievable metadata, and 10 failed in conversion itself. The same transform code may be run locally, without the guards, on otherwise excluded ontologies.

Table 1. Outcome of the 2026-09-09 build for every ontology attempted.

Outcome	Reason recorded in the index	Ontologies
Converted	Base graph published (9 of these are import-only)	1,184
Withheld	License restricted (including UMLS vocabularies)	8
Withheld	Skipped before download due to file size limits	16
Withheld	Source file found to be over 250 MB limit	7
Failed	BioPortal side: no submission (27), no file (35), metadata not retrievable (16)	78
Failed	Conversion error: source unreadable by any parser (4), transform error (6)	10
Total attempted		1,303

The failure classes are not fixed. In the first public build, on 2026-08-02, conversion errors affected 189 ontologies, and the dominant cause was not malformed content but owl:imports. ROBOT resolves imports over the network when it loads an ontology, and a single dead, slow, or

access-restricted import URL aborts the entire conversion. Ontologies are often published with import targets that no longer resolve. The failure is silent for the author, whose local build environment has the imports cached. Stripping import declarations before invoking ROBOT rescued 121 of the 189 affected ontologies in that build, at the cost of leaving cross-ontology references as dangling edges that resolve when graphs are combined. Recording which stage lost each ontology, and what it said, then let the remaining errors be fixed one class at a time over the following month, and by the 2026-09-09 build transform errors were down to six. Import resolution proved an underappreciated obstacle to ontology reuse at scale: one change to one step moved roughly a tenth of the BioPortal inventory from unavailable to available. The cost of that change is measurable. In the 2026-09-09 base graphs, 5,265,221 of 15,924,223 edges (33%) touch a node the graph itself does not name: a term imported from another ontology, or an identifier the source refers to without defining. The share is 54% in the 263 graphs whose source declares imports and 28% in the rest, whose references were never imported to begin with. In the 141 ontologies for which a full graph was built, merging the import closure supplies a name for 36% of those edges' endpoints, which is the part attributable to stripping imports and the part the full graph repairs.

2.5. *Made-for-purpose graph assembly*

Publishing one graph per ontology is the feature that makes the collection more widely usable. We perform graph merges with the KGX cat-merge utility and document this as a recipe rather than implement it as a service. Node rows from the input graphs are unioned on the identifier column, so that two rows sharing a CURIE become a single node with their properties combined, and edge rows are concatenated and deduplicated on the subject-predicate-object triple. Nothing is collapsed on the basis of labels. Combining all graph-ready BioPortal ontologies this way is possible, and yields a giant merged KG of 8.3 million nodes and 15.9 million edges, but for most analyses it is the wrong object.

The difficulty a merged graph creates is not that it silently fuses distinct concepts, since identifier-keyed merging does not do that, but that it makes every sense of every term simultaneously present. For example, the label *nucleus* denotes the eukaryotic organelle in the Gene Ontology (`GO:0005634`), a neuroanatomical structure in the Zebrafish Anatomy and Development Ontology (`ZFA:0001665`), and the atomic nucleus in the Chemical Entities of Biological Interest (ChEBI) ontology (`CHEBI:33252`). All three terms survive the merge as separate nodes, and every operation that works from labels rather than from identifiers, including text annotation, lexical mapping, string search, and most embedding methods, must then arbitrate among them.

KG-Bioportal therefore supports selective assembly: a user selects the ontologies relevant to their specific question, retrieves those graphs by their download URLs from the index, and merges only those, by the procedure above. Graphs built this way typically contain fewer than 100,000 edges, two to three orders of magnitude smaller than the all-ontology graph. They are small enough to store in the local memory in most computing environments, to visualize, and to reason over interactively. Because the products conform to KGX, they can be easily combined with other KGX-format graphs such as those built under KG-Hub conventions[7] and those produced through the NCATS Translator project[10]. They can also be used as inputs for graph machine learning libraries such as GRAPE[2] for embedding and link prediction without an intervening translation step. KG-Bioportal intentionally restricts its scope to data, reserving analysis for existing graph libraries usable with this resource's common format. Selecting the ontologies that enter a graph is also a scoping decision recorded in the graph's construction rather than buried in an analysis script.

3. Supporting Infrastructure for Graph-based Bioinformatics

3.1. *Retrieving and composing graphs programmatically*

Systems biology workflows include steps that seem suitable for agentic approaches, such as assigning cell types to transcriptomic clusters, normalizing the columns of a sample sheet, or extracting entities from literature; however, generative AI has a well-documented tendency to fabricate plausible-looking ontology identifiers[11]. To avoid this, we developed *bioportal-mcp*, a server implementing the Model Context Protocol for AI agents that exposes BioPortal term search, property search, and usage analytics as tools an agent can call, so that the agent asks BioPortal what a term's identifier is instead of recalling or guessing it. This server is distributed through PyPI and can be added to an agent client with a single command. The server does not yet expose BioPortal's mapping service, though mappings may be obtained through the BioPortal REST API using the agent skills described below.

3.2. *Documenting procedures for reusable graph operations*

Documentation tells a reader what a resource contains but may not tell a piece of software what to do with it. An agent handed the KG-Bioportal index still has to rediscover which release holds an artifact and what columns the tables carry. Each detail is a place where a capable agent may produce confident nonsense that a human reviewer may not check.

We therefore distribute the procedural knowledge alongside the data, as three agent skills carried in the pipeline repository: *kg-bioportal-data* for locating and loading graphs, *kg-bioportal-merge* for combining them with cat-merge and reporting dangling cross-ontology edges, and *bioportal-for-kg-building* for using BioPortal's Annotator, Recommender, and mapping services as inputs. A fourth procedure belongs to the pipeline itself: a roots command that lays out

a graph's top-level classes with their subclass labels, which is the packet the category review of **Section 2.2** works from. Each carries helper scripts and data-model facts verified against the live collection, enabling agents to assemble made-for-purpose graphs. Agents are not compelled to use all data obtained from BioPortal (e.g., they are not required to use the provided mappings) but may instead work with graph data from multiple sources if specified.

4. Cross-domain Inference with Assembled Graphs

The value of rendering heterogeneous ontologies as composable graphs is best shown by example. Consider a clinician-researcher asking what adverse interactions might occur if a patient being treated for atrial fibrillation drinks grapefruit juice. A graph assembled from five ontologies answers the question along a single path. The Food Ontology supplies a term for raw grapefruit (`FOODON:03301702`). The Food Interactions with Drugs Evidence Ontology contributes a class describing an interaction between grapefruit and dronedarone (`FIDEO:000001707`). ChEBI provides the corresponding chemical entity (`CHEBI:50659`), conveniently annotated with the role "anti-arrhythmia drug," which both confirms the clinical context and makes the drug discoverable by mechanism rather than by name. A mapping links the ChEBI entry to its RxNorm equivalent (`RXNORM:233698`), where the drug acquires the trade names and dose forms that appear in prescribing records. The traversal from a food to a specific oral tablet crosses four independently maintained vocabularies; in the assembled graph it is a six-hop path, or longer if we extend the traversal to the Toxic Process Ontology (TXPO) (**Fig. 2**). The mapping that carries the path from ChEBI into RxNorm is an exact-label match, the kind of label-driven correspondence **Section 2.5** warns against. A lexical mapping like this is a candidate that a person or a rule accepts one at a time, as here, and not a basis on which a merge silently fuses nodes. Though neither ChEBI nor RxNorm graph transforms are in KG-Bioportal's hosted collection, their graphs may be built by running the same pipeline locally. This is the intended path for anything the hosted build does not carry.

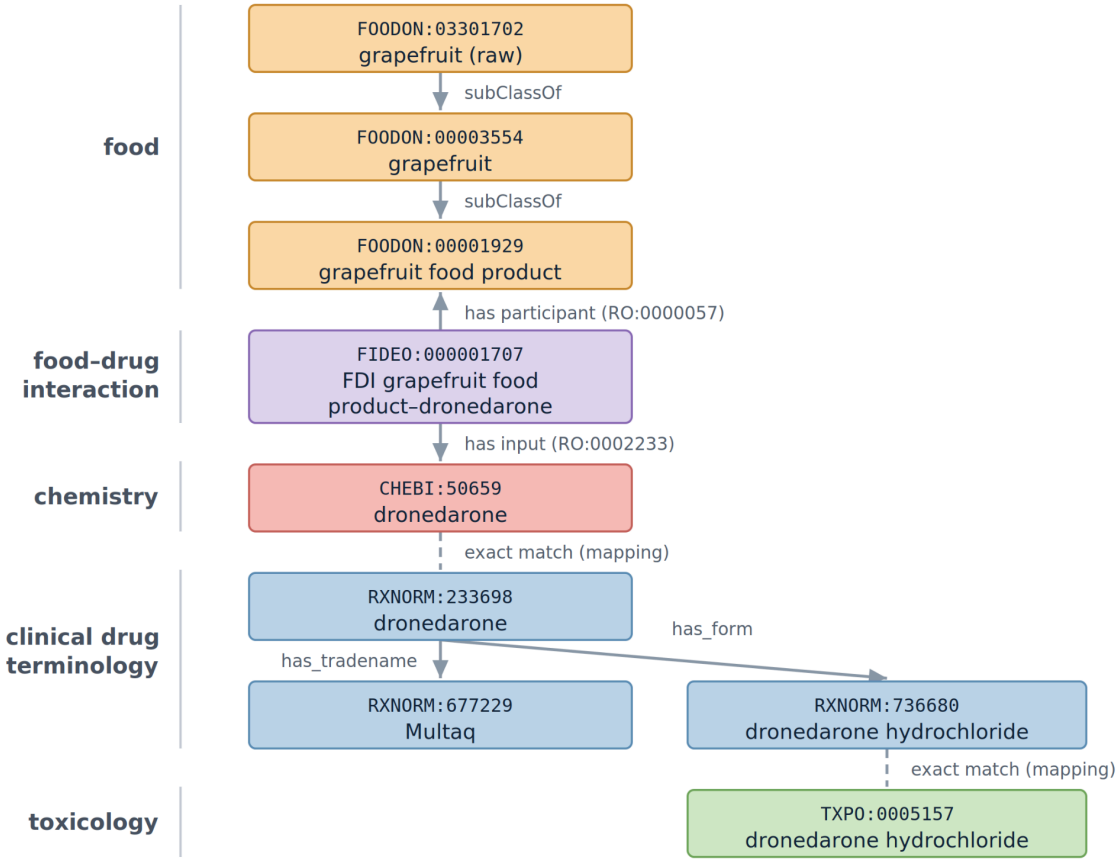


Fig. 2. A path through a made-for-purpose graph assembled from nine nodes from five BioPortal ontologies. Arrows between nodes represent typed relationships, along with their Relation Ontology (RO) identifiers where available. Dashed lines represent mappings obtained through BioPortal's LOOM method; all mappings shown here are exact label matches. Node color indicates the source ontology. The labels at left name the domain each source belongs to.

5. Toward a Repository for Knowledge Graphs

KG-Bioportal makes ontologies available as graphs. Biomedical KGs such as Hetionet[12], SPOKE[13], the Monarch Initiative KG[14], and PrimeKG[15] are central to work on drug repurposing, precision medicine, and genotype–phenotype inference, but they remain siloed, without a common entry point. The repository pattern that BioPortal established for ontologies has no counterpart for graphs, and the community has asked us to provide one.

BioPortal is built on RDF and stores its content in a triplestore, which gives us a graph database to build on but does not by itself solve the problem, because the biomedical KG community has largely moved toward labeled property graphs (LPGs). LPGs permit data to be attached to edges as well as nodes, so supporting them requires extending the BioPortal object model to associate information with edges. Once that is done, an edge asserting that imatinib

inhibits BCR-ABL1 can carry its interaction type, its supporting evidence, its source database, and its version, making the relationship not merely findable but interpretable.

We have deployed a first version of that entry point. It presents, in the familiar BioPortal browse idiom, a single filterable index over two populations that have until now been described in separate places: the BioPortal ontology entries produced by the pipeline and 161 community knowledge graphs, each with its domain, license, publications, and reported metrics, listed and faceted together. Every entry has a resource page, including the ontologies that failed to convert.

Storing graphs in the same backend as ontologies means that much existing BioPortal functionality extends to them at once. Nodes from submitted KGs become searchable exactly as ontology classes are, and may be returned by services such as the Annotator, which broadens the range of entities that BioPortal services can recognize. A user searching for quercetin would find the DrugCentral node (`drugcentral:3514`) even though no current BioPortal ontology contains it, and could follow its contraindication edges to Mondo kidney disease terms and its targets edges to Protein Ontology terms, with provenance carried on each edge. No comparable capability yet exists for KGs elsewhere.

6. Related Work

Several services besides BioPortal serve biomedical ontologies. EMBL-EBI's Ontology Lookup Service[16], the OBO Foundry[1], and Ontobee[17] all provide term-level search, resolution, and in some cases reasoning over ontologies. Each is designed for term retrieval rather than graph operations. Several projects have focused on rendering sets of OBO ontologies as graphs: Ubergraph materializes relaxed and reasoned OBO Foundry ontologies as one RDF graph[18]; the OBO Graphs JSON serialization[19] and semantic-sql[20] both give individual ontologies more graph-like or relational forms; and our own KG-OBO project served as a precursor to KG-Bioportal, as it performed OBO to KGX transformations across the full Foundry[21]. KG-Bioportal differs from each of these in scope and packaging. It covers nearly all of BioPortal rather than the OBO Foundry alone, involving a collection several times larger and containing a wider set of topics. KG-Bioportal also tracks the source and cause of all transformation errors.

KG-Bioportal is part of a broader ecosystem of tools and resources for working with biological and biomedical graph data. KNIT, for example, approaches reuse from the other direction, allowing users to explore an existing knowledge graph to decide which ontologies to reuse[22]. BioPortal's own Recommender tool serves the same purpose for free text. The KG-Bioportal graphs can then be assembled based on what the tool highlights. The knowledge graphs named in Section 5 also complement our work well.

Parts of the BioPortal infrastructure described in Sections 1 and 5 were reported in the 2025 BioPortal update. An early form of KG-Bioportal was presented at the 2025 Bioinformatics Open Source Conference. This paper is the first full description of the KG-Bioportal pipeline.

7. Discussion

The work described here rests on a simple observation: the biomedical community has invested enormous effort in building ontologies, and a second, largely separate effort in building knowledge graphs. The connection between the two has been left to individual groups to reconstruct repeatedly. Converting BioPortal's holdings into a consistent, typed, versioned graph collection turns a body of curated knowledge that was accessible but awkward into one that is directly usable by network methods. Running that conversion on a schedule, in public, with its failures on the record, yields a standing measurement of the biomedical ontology landscape’s graph-readiness.

Several limitations bound what we have achieved. Biolink categorization is only as good as the evidence in the source: 52.2% of nodes carry a category more specific than NamedThing, and the rest stay untyped rather than guessed. The ROBOT relax step and the conversion keep most but not all axiomatic structure. Measuring this on six OBO ontologies (GO, CL, HP, PATO, ENVO, and FOODON) by running the pipeline's own transform on their current sources and classifying every class axiom with a named subject (268,364 in all), we found that 99.0% of those axioms are a named superclass, a simple existential restriction, or an intersection of such. 99.97% of these are emitted as an edge. The remaining 1.0% (2,576 axioms, and 8.2% in ENVO, the most heavily axiomatized of the six) are nested or non-existential expressions: unions, complements, cardinality and value restrictions. KGX does not drop these outright but renders each as an edge to an unlabeled anonymous node. 1.3% of all edges in the collection are of that kind so the axiom is lost. Stripping owl:imports makes many more ontologies convertible but leaves cross-ontology references dangling until graphs are combined (**Section 2.4**), which is the right trade for a modular collection, and the full graphs exist for users who want a self-contained one.

Coverage is incomplete, for the reasons given in **Section 2.4**. The graphs inherit their sources' errors as faithfully as their content. A merged graph does not audit the ontologies that enter it, and a densely connected region is not a well-evidenced one. The provenance on every node and edge is helpful here, as a suspect assertion can be traced to the ontology and submission that made it. Finally, we have not yet shown that a graph composed from this collection serves a downstream task, such as link prediction, better than one built by a conventional OWL ingest. KG-Bioportal makes that comparison easy to set up and more reproducible. Use case comparisons and development of a full task-level benchmark will be performed in future work.

Some functionality will require additional consideration. A knowledge graph is not hierarchical and requires interface components able to display non-hierarchical structure. Services

such as the Recommender, which ranks ontologies by their fit to a body of text, need to be reconceived for graphs. Content overlap among graphs is expected and is not a problem to be resolved by merging: each graph can occupy its own space in the repository, retaining its own provenance and its own curatorial decisions, while shared entities become discoverable across graphs through BioPortal's mapping services. A user who finds a term in one graph should be able to see how other graphs represent it based on the evidence and context specific to each.

The trajectory we have set out points toward a repository in which ontologies and knowledge graphs are managed as two aspects of a single problem. Both define relationships among concepts; both are modular, versioned, community-maintained artifacts; both are consumed increasingly by software rather than by people. Treating them together lowers the barrier to entry for researchers who need reusable domain knowledge and, we hope, makes the graphs on which network-based biology increasingly depends as accessible and interoperable as the ontologies that underlie them.

BioPortal is freely available at https://bioportal.bioontology.org. KG-Bioportal is open source at https://github.com/ncbo/kg-bioportal, with prebuilt graphs and the collection index distributed through the repository's releases and browsable at https://ncbo.github.io/kg-bioportal/graphs/. The *bioportal-mcp* code is available at https://github.com/ncbo/bioportal-mcp and from PyPI. The software is released under a BSD license. Each graph is a derivative of one ontology and is redistributed under the ontology's own terms as recorded by BioPortal at its submission. Ontologies with licenses forbidding redistribution are not transformed (**Section 2.4**). Licenses for each transformed ontology are specified on KG-Bioportal's site and index except where no license has been specified.

Acknowledgments

This work has been supported in part by the National Institutes of Health [U24 GM143402, U54 HG004028, 5U01HG009453-03 (past support), 3OT2DB000009]; and by the Director, Office of Science, Office of Basic Energy Sciences of the US Department of Energy [Contract No. DE-AC0205CH11231 to JHC, NLH, and CJM]. We are grateful to Franz, Inc. (https://franz.com) for their generosity and technical support in enabling the integration of AllegroGraph with BioPortal, and to the BioPortal user and developer community, including the members of the OntoPortal Alliance, for their sustained contributions.

References

1. Jackson, R. *et al.* OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. *Database (Oxford)* **2021**, (2021).
2. Cappelletti, L. *et al.* GraPE: fast and scalable Graph Processing and Embedding. *Research Square* (2022) doi:10.21203/rs.3.rs-967145/v1.
3. Whetzel, P. L. *et al.* BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res.* **39**, W541-5 (2011).
4. Vendetti, J. *et al.* BioPortal: an open community resource for sharing, searching, and utilizing biomedical ontologies. *Nucleic Acids Res.* **53**, W84–W94 (2025).
5. Harris, P. A. *et al.* The REDCap consortium: Building an international community of software platform partners. *J. Biomed. Inform.* **95**, 103208 (2019).
6. Musen, M. A. *et al.* The center for expanded data annotation and retrieval. *J. Am. Med. Inform. Assoc.* **22**, 1148–1152 (2015).
7. Caufield, J. H. *et al.* KG-Hub-building and exchanging biological knowledge graphs. *Bioinformatics* **39**, btad418 (2023).
8. Jackson, R. C. *et al.* ROBOT: A tool for automating ontology workflows. *BMC Bioinformatics* **20**, 407 (2019).
9. Unni, D. R. *et al.* Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science. *Clin. Transl. Sci.* **15**, 1848–1855 (2022).
10. Fecho, K. *et al.* Announcing the biomedical data Translator: Initial public release. *Clin. Transl. Sci.* **18**, e70284 (2025).
11. Groza, T. *et al.* An evaluation of GPT models for phenotype concept recognition. *BMC Med. Inform. Decis. Mak.* **24**, 30 (2024).
12. Himmelstein, D. S. *et al.* Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife* **6**, e26726 (2017).
13. Morris, J. H. *et al.* The scalable precision medicine open knowledge engine (SPOKE): a massive knowledge graph of biomedical information. *Bioinformatics* **39**, (2023).
14. Putman, T. E. *et al.* The Monarch Initiative in 2024: an analytic platform integrating phenotypes, genes and diseases across species. *Nucleic acids research* **52**, D938–D949 (2024).
15. Chandak, P., Huang, K. & Zitnik, M. Building a knowledge graph to enable precision medicine. *Sci. Data* **10**, 67 (2023).
16. McLaughlin, J., Lagrimas, J., Iqbal, H., Parkinson, H. & Harmse, H. OLS4: a new Ontology Lookup Service for a growing interdisciplinary knowledge ecosystem. *Bioinformatics* **41**, btaf279 (2025).
17. Ong, E. *et al.* Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration. *Nucleic Acids Res.* **45**, D347–D352 (2017).
18. Balhoff, J. *et al.* Ubergraph: Integrating OBO Ontologies into a Unified Semantic Graph. *ICBO* 1–9 (2022).
19. Jacobsen, J. *et al. Geneontology/Obographs: Immutables and Splitting out OWL API*. (Zenodo, 2021). doi:10.5281/ZENODO.5070191.
20. Mungall, C. *Semantic-Sql: SQL and SQLite Builds of OWL Ontologies*. (Github, 2026).
21. Caufield, J. H., Reese, J. & Mungall, C. KG-OBO: Open bio-ontologies in knowledge graph form. (2022) doi:10.7490/f1000research.1119049.1.
22. Rodríguez-Revello, J., Barba-González, C., Rybinski, M. & Navas-Delgado, I. KNIT: Ontology reusability through knowledge graph exploration. *Expert Syst. Appl.* **228**, 120239 (2023).

Integrative interactome analysis for mechanistic discovery and drug repurposing in rare diseases: GPX4-related rare skeletal dysplasia

Kalyani B. Karunakaran[1], N. Balakrishnan[1] and Madhavi K. Ganapathiraju[*2,3]
[1]Supercomputer Education and Research Centre, Indian Institute of Science, Bangalore, 560012, India
[2]Department of Biomedical Informatics and Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, 15213, USA

Rare diseases pose a major challenge for mechanistic and therapeutic discovery because of limited molecular data and small patient populations. Building on our previously developed interactome-based methods, we present an integrative network analysis workflow for constructing disease-specific functional landscapes by combining protein-protein interactions (PPIs), disease-gene associations, phenotypes, tissue expression, gene perturbation data, and drug-target relationships. We applied it to Spondylometaphyseal dysplasia, Sedaghatian type (SMDS), an ultra-rare lethal skeletal dysplasia caused by GPX4 mutations. Disease-centric and gene-centric networks connected GPX4 with genes implicated in SMDS and related skeletal dysplasias, and with regulators of GPX4 expression, and were enriched for SMDS-relevant phenotypes, fetal-stage functional modules, and brain/testis-enriched genes. Seven high-confidence computationally predicted GPX4 interactors (APBA3, EGR4, FUT5, GAMT, GTF2F1, MATK, and ZNF197) were supported by convergent network, phenotype, and expression evidence. Integrative transcriptomic and network analyses prioritized eleven candidate repurposable drugs, with resveratrol additionally highlighted by network proximity. Nearly 75% of rare diseases are associated with single genes, many understudied with limited experimental data. This study demonstrates an approach to studying a rare disease in the context of its protein interactome to generate mechanistic and therapeutic hypotheses.

Keywords: rare disease; protein interactome; protein-protein interaction prediction; GPX4; Sedaghatian-type spondylometaphyseal dysplasia; drug repurposing

1. Introduction

There are over 10,000 rare diseases which, though rare individually, affect 10% of the population collectively; nearly half of the patients are children [1]. 75% of the rare diseases are associated with a single gene each though they may present multiple phenotypes (source: our analysis of Orphanet data https://www.orpha.net accessed July 2026 [2]). For example, many rare diseases are pediatric neurodevelopmental disorders, such as Malan syndrome, Tatton-Brown-Rahman syndrome, Sotos syndrome, Weaver syndrome, etc., and each associated with a different gene, NFIX, DNMT3A, NSD1, and EZH2, respectively. Unfortunately, hundreds of genes associated with rare diseases are also understudied biologically, presenting an impediment to inferring molecular mechanisms for a given rare disease, such as pathways or biological processes associated. Characterizing these genes

* Corresponding author: madhavi@pitt.edu

© 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.

in terms of their protein function, structure, and interactions can therefore provide a foundation for data-driven hypothesis-generation in rare disease research.

One critical avenue for discovering mechanisms of diseases and identifying therapeutic targets lies in elucidating protein-protein interactions (PPIs) involving disease-associated genes [3-5]. PPIs drive cellular processes by facilitating signal transduction, formation of cellular structures, and enzymatic reactions. Disease-associated variants are enriched in protein cores and interaction interfaces, and often alter binding affinity or establish novel interactions [6, 7]. As a result, the genetic perturbations can propagate through PPI network (the 'interactome') influencing multiple biological processes, contributing to disease pathogenesis, and presenting multiple phenotypes despite their origin in a single gene [8]. Several studies, led by our group and others, have demonstrated that shared genetic architecture and overlapping phenotypes across distinct disorders can be traced to convergent perturbations in PPI networks [9-14]. Interactomes provide a unifying data structure for integrating heterogeneous molecular, phenotypic, and functional annotations in a disease-centric manner.

Here, we present an interactome-based study of Spondylometaphyseal dysplasia, Sedaghatian type (SMDS), an ultra-rare developmental disorder caused by pathogenic variants in a single gene, GPX4. SMDS is a lethal autosomal-recessive skeletal dysplasia with metaphyseal, vertebral, cardiac, respiratory, and neurodevelopmental manifestations [15-22]. Cardiac-respiratory distress is the primary cause of death in infancy and living patients with severe hypotonia and intractable seizures have been reported [21, 23]. Studies on SMDS have described its clinical presentation [15-20, 22], and linked it to GPX4 mutations [21, 24]. This includes a detailed analysis of the homozygous GPX4 R152H mutation that disrupts enzyme stability and function, and is partially rescued by selenium supplementation or selective antioxidant therapies [25]. No approved therapies exist for SMDS; management has been limited to antioxidant supplementation in a few patients, with anecdotal benefit [24]. Nevertheless, the etiology of this severe lethal dysplasia remains largely unexplored. Hence, novel approaches are needed to understand the fundamental biological mechanisms underlying this disorder. Motivated by the efforts of affected families (www.curegpx4.org) and growing calls from the rare disease community to accelerate mechanistic understanding, we applied our previously established interactome-analysis approach to construct a GPX4-centered interactome and systematically integrated functional annotations to generate testable biological hypotheses. This study illustrates how PPI networks and their functional annotation can support mechanism discovery in single-gene rare diseases for which conventional omics-based analyses are not feasible.

2. Methods

Detailed Methods, including data sources, parameter settings and statistical tests for each analysis, are provided in the Supplementary Information (Section 5); a summary is given here. Disease associated gene sets were compiled from the DisGeNET database [26], and from Warman et al. [27]. Genes whose knockout, knockdown, mutation, or overexpression changed GPX4 expression were collected from Knockdown Atlas [28]. PPIs involving these genes which have been reported in the literature ('known PPIs') were obtained from HPRD (Human Protein Reference Database,

version 9 [29]) and BioGRID (Biological General Repository for Interaction Datasets, version 3.4.159 [30]). To systematically expand the molecular context of GPX4 and SMDS-associated genes, high-confidence computationally predicted PPIs were incorporated using the HiPPIP model (a random forest classifier), which we previously developed and validated for high precision through computational evaluations and experimental validations [11, 31-33]. Following the original HiPPIP publication [11], in which all pairs scoring above 0.5 were deemed high-confidence predictions without distinction between their individual scores, we likewise regard all pairs above this threshold as equally high-confidence predictions; accordingly, individual scores are not used to rank interactors and are not tabulated here. Protein pairs not previously reported but predicted as interacting by HiPPIP were treated as novel (predicted) PPIs. The PPI networks, consisting of direct interactors of input genes as well as shortest paths connecting them, were created using LENS web application [34]. While the public web-application provides only known PPIs, an in-house version of LENS under development, provides networks including HiPPIP-predicted PPIs. Shortest paths were computed using Dijkstra's algorithm, with path length measured as the number of edges and no cap on length, so the closest neighbor of a gene may be one, two, or more edges away. For each gene in the input set, all its closest neighbors in the set, and all shortest paths to them, were included in the network along with the intermediate proteins and interactions. Shortest paths were used to assemble the network neighborhood of the genes of interest and not as a statistical measure of biological relatedness: in scale-free interactomes, path length largely reflects the degree distribution rather than functional relatedness [35], so no inference is drawn from path lengths themselves. The networks were downloaded as SIF files, and visualized in Cytoscape [36]. LENS also provided pathways, diseases, drugs and genome-wide association study (GWAS) annotations significantly enriched among the network genes (P-values from Fisher's exact test).

Global and fetal-context functional modules, containing minimum eight nodes, were identified using shared k-nearest-neighbor graphs and Louvain clustering with HumanBase toolkit [37, 38]. Gene Ontology enrichment was evaluated by Fisher's exact test with Benjamini-Hochberg correction. Genes associated with twelve major SMDS-associated pathophenotypes were compiled from the MONARCH database [39]. Enrichment of these phenotypes within the interactomes was evaluated against a background of 15,451 genes with phenotype annotations, using Fisher's exact test (phenotype gene sets ranged in size from 4 to 355 genes); nominal (uncorrected) P-values and odds ratios are reported. Median expression across 53 postnatal tissues was obtained from GTEx [40]. Logarithmic transcripts per million (TPM) values were standardized by gene. PCA used SVD with iterative imputation in ClustVis [41]; hierarchical clustering used Pearson distance and average linkage in Heatmapper [42].

Repurposable drug candidates were identified by two independent signals: negative score reported in BaseSpace Correlation Engine [43] based on drug-induced and dysplasia-associated transcriptional profiles, and proximity between drug targets from DrugBank [44] and the GPX4/SMDS interactome. The BaseSpace Correlation Engine (formerly NextBio) computes a correlation score from the strength of overlap between two expression profiles, applies statistical criteria including correction for multiple hypothesis testing, ranks correlated profiles by statistical significance, and normalizes scores to 100 for the top-ranked result. Convergence between signals

was treated as stronger evidence; all candidate drugs are computational hypotheses, requiring experimental and clinical validations.

3. Results[†]

3.1. *GPX4 interactome construction and expansion*

We assembled literature-supported PPIs, and 'novel' PPIs of GPX4 computationally predicted by HiPPIP, a model whose claim for high-confidence prediction was previously validated computationally and experimentally [11, 31-33]. GPX4 is found to have 7 novel interactors; see Fig.

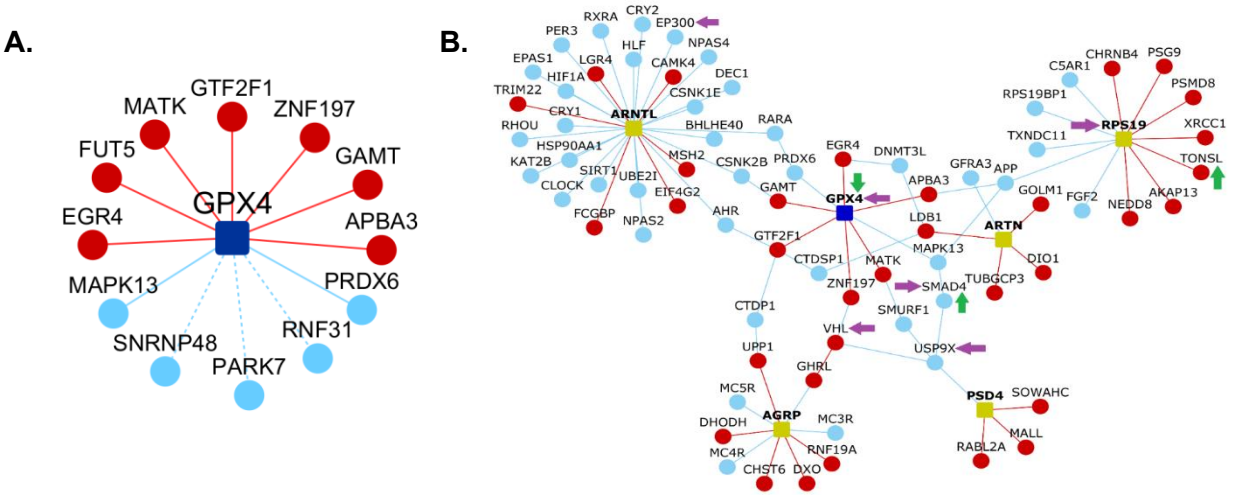


Figure 1: A. Known and Novel PPIs of GPX4. Subsequent to carrying out the analysis presented in this study, three additional PPIs have been reported in literature, which are shown in dashed lines. **B. Interactions of GPX4 with other genes speculated to be associated with Sedaghatian type spondylometaphyseal dysplasia**. This network diagram shows the shortest paths connecting the genes from the DisGeNET database (dull green colored nodes) identified through text mining to GPX4. Nodes depict proteins and edges depict PPIs. Red and light blue colored nodes denote novel and known interactors respectively. Red and light blue colored edges denote novel and known PPIs respectively. Purple colored arrows indicate genes associated with cardiac defects, whereas green colored arrows indicate those associated with skeletal defects.

1A and **Table 1**. Among the seven predicted interactors, GAMT presents phenotype-level connection: Human Phenotype Ontology annotations report both GPX4 and GAMT to be associated with hypotonia phenotype (HP:0001252). We analyzed four complementary network views of GPX4: (i) other genes putatively associated with SMDS in DisGeNET [26], (ii) genes associated with other skeletal dysplasias, (iii) genes associated with disorders phenotypically similar to SMDS and (iv) genes whose perturbation causes significant differential expression of GPX4. For each network, we identified enriched functional modules and tested enrichment for SMDS-relevant pathophenotypes.

[†] *Additional details are provided in Supplementary Information referred to in Section 5.*

Table 1: Biological relevance of novel interactors of GPX4. See Supplementary Material for additional details.

Gene	Potential relevance to GPX4/SMDS
APBA3	Connects GPX4 to skeletal dysplasia genes (IARS2), resveratrol target APP, and oxidative stress pathways.
EGR4	Brain developmental TF; co-expressed with GPX4; six shared GPX4 mouse phenotypes; implicated in cartilage/neurogenesis modules.
FUT5	Testis-enriched; co-expressed with GPX4; implicated in cartilage development and neurogenesis.
GAMT	Links GPX4 to TERF1; 4 mouse phenotypes and 1 human phenotype shared with GPX4; oxidative stress module.
GTF2F1	Connects GPX4 with PAM16 (Megarbane SMD) and resveratrol targets; participates in oxidative stress and epithelial morphogenesis modules.
MATK	Brain-enriched; participates in GPCR signaling and cartilage development modules.
ZNF197	Links GPX4 to VHL/HIF1A signaling and skeletal dysplasia genes; shared mouse phenotype; implicated in VEGFA regulation, cartilage development, ROS metabolism, and neurogenesis.

3.2. *Interactome of GPX4 and other genes putatively associated with SMDS*

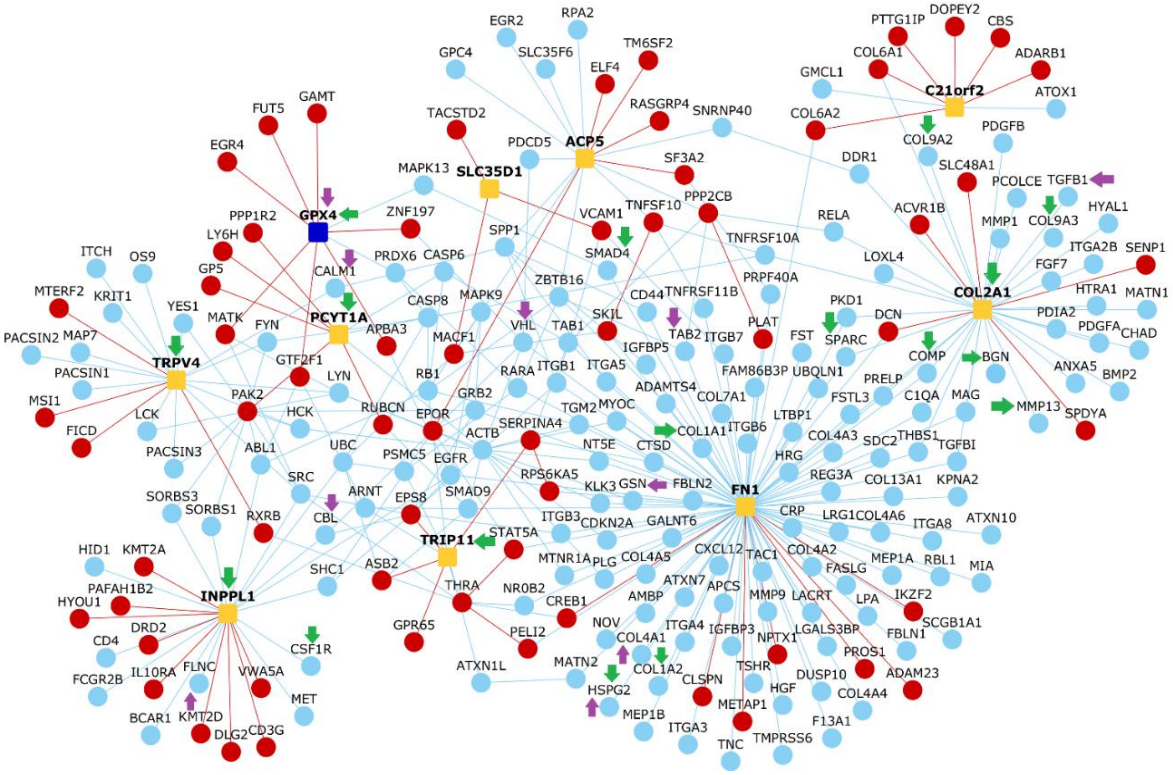


Figure 2: GPX4 interaction network linking GPX4 to genes associated with other spondylometaphyseal dysplasias and related skeletal dysplasias. Shown are the shortest paths connecting GPX4 to disease-associated genes (orange). Nodes represent proteins and edges represent PPIs; red/light blue indicate novel/known interactors and PPIs, respectively. Purple and green arrows denote genes associated with cardiac and skeletal defects.

DisGeNET lists six genes – AGRP, ARNTL, ARTN, LOH19CR1, PSD4, and RPS19 – as potentially associated with SMDS based on a prior study [45]. That study identified an SMDS-associated variant within a sperm RNA element mapping to GPX4; sperm RNA elements are short exon-sized sequences corresponding to unique coding and non-coding transcripts [45]. Their disease

association scores in DisGeNET are very low (0.01), indicating some, but minimal supporting evidence, whereas GPX4 has a higher score of 0.72. These genes were tagged as potentially linked to SMDS in DisGeNET based on co-mention of gene and disease names in literature. Given the scarcity of information on SMDS, these co-mentions may provide testable insights. We included these genes in our network analysis to assess whether they exhibited biologically meaningful connectivity to GPX4 and the broader SMDS molecular landscape.

We extracted the interactors of these six genes and the shortest paths connecting them to GPX4 in the PPI network and found that each gene was connected to GPX4 by fewer than three intermediate interactions. Notably, these paths traversed six of the seven novel interactors of GPX4 (Fig. 1B). This interactome was enriched in genes associated with myocarditis, atrial septal defect, and platyspondyly (P = 5.33E-04, 0.016 and 0.018; odds ratio = 55.8, 3.37 and 5.38, respectively). From this, we identified fetus-specific and tissue-naive (or 'global') modules. Fetus-specific modules may provide insights into the congenital underpinnings of SMDS. Cellular oxidant detoxification pathway was identified in both global (M4: *Q*-value = 2.15E-03) and fetus-specific (M3: *Q*-value = 2.77E-03) contexts. G-protein coupled receptor signaling pathway was detected both as a global (M3: *Q*-value = 4.25E-04) and fetus-specific (M2: *Q*-value = 3.4E-04) module. *Cell growth* (M1: *Q*-value = 2.38E-04) and *response to redox state* (M2: *Q*-value = 2.38E-04) were detected only as global modules.

3.3. *Interactome of GPX4 and genes associated with other skeletal dysplasias*

Nine genes representing twelve related skeletal dysplasias were curated from Warman et al. [27]. Their network consisting of direct interactors and shortest paths to GPX4 contained 93 intermediate proteins and 26 novel interactions (Fig. 2). This interactome was enriched in genes associated with arrhythmia (P-value = 0.041, odds ratio = 1.88), cardiorespiratory arrest (P-value = 9.25E-03, odds ratio = 13.38), myocarditis (*P*-value = 4.64E-03, odds ratio = 19.11), metaphyseal cupping (P-value = 1.9E-07, odds ratio = 22.5), platyspondyly (*P*-value = 4.98E-13, odds ratio = 10.43) and abnormal ribs (*P*-value = 8.9E-03, odds ratio = 3.93). Paired global/fetal modules implicated extracellular-matrix and cartilage development, vascular development/cell migration, and apoptosis. Additional fetal modules involved cytoskeletal organization, cellular maintenance, and protein catabolism, indicating convergence of GPX4-associated networks with clinically relevant skeletal and cardiovascular phenotypes.

3.4. *Interactome of GPX4 and genes associated with disorders phenotypically like SMDS*

We used Phenogrid from MONARCH to identify the diseases that were phenotypically most similar to SMDS. Seven skeletal dysplasias shared 70% phenotypic similarity with SMDS: X-linked spondyloepimetaphyseal dysplasia, acrocapitofemoral dysplasia, A4 type SMD, Dyggve-Melchior-Clausen disease, autosomal recessive Megarbane type SMD, metaphyseal acroscyphodysplasia and X-linked Dyggve-Melchior-Clausen disease. Four of them had genetic associations reported in DisGeNET: X-linked spondyloepimetaphyseal dysplasia (BGN and IARS2), acrocapitofemoral dysplasia (IHH), Dyggve-Melchior-Clausen disease (DYM), and autosomal recessive Megarbane type SMD (PAM16). GPX4 was connected to these five genes through 13 intermediate interactors

including three novel interactors of GPX4 (APBA3, GTF2F1 and ZNF197) (Fig. 3). This interactome was enriched in genes associated with arrhythmia, cardiorespiratory arrest, cerebellar hypoplasia, agenesis of corpus callosum, metaphyseal cupping, platyspondyly and abnormal ribs (P = 0.042, 5.9E-04, 8.63E-03, 1.88E-03, 3.73E-05, 5.9E-09 and 3.8E-03; odds ratio = 3.05, 54.2, 4.95, 5.65, 45.18, 19.9 and 9.57, respectively, in that order), mirroring key SMDS clinical features. Five modules were detected in both global and fetus-specific contexts: skeletal system development/embryo development/gastrulation, lipid storage/reactive oxygen species biosynthetic process, humoral response, cellular response to toxic substance and membrane organization. A fetus-specific module enriched for neurogenesis was also identified, consistent with neurodevelopmental manifestations of SMDS.

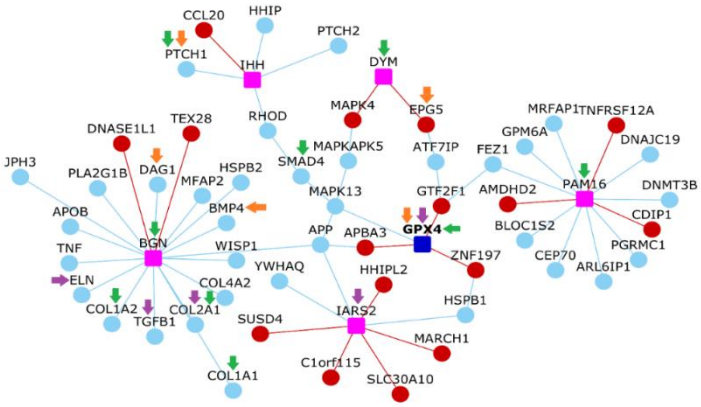


Figure 3: Interactions of GPX4 with genes associated with bone dysplasias sharing phenotypic similarity with Sedaghatian type spondylometaphyseal dysplasia. This network diagram shows the shortest paths connecting GPX4 with the genes associated with X-linked spondyloepimetaphyseal dysplasia, acrocapitofemoral dysplasia, Dyggve-Melchior-Clausen disease and autosomal recessive Megarbane type SMD (pink colored nodes). Nodes depict proteins and edges depict PPIs. Red and light blue colored nodes denote novel and known interactors respectively. Red and light blue colored edges denote novel and known PPIs respectively. Purple, green and orange arrows indicate genes associated with cardiac, skeletal and brain defects respectively.

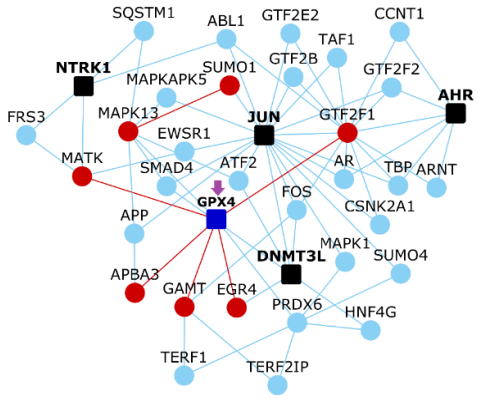


Figure 4: Four genes whose perturbation affects GPX4 expression, which are closely connected to GPX4 in the network. Nodes represent proteins and edges represent PPIs; red/light blue denote novel/known interactors and PPIs, respectively. Black nodes are the 4 genes mentioned and purple arrows indicate genes associated with heart defects.

3.5. *Interactome of GPX4 and genes that affect GPX4 expression when perturbed*

From the Knockdown Atlas (BaseSpace Correlation Engine [28]), we compiled 136 genes whose knockout, knockdown, mutation or overexpression significantly affects GPX4 expression. The connecting network included all the seven novel interactors of GPX4. Among the 136, four were closely connected to GPX4, namely, AHR, DNMT3L, JUN and NTRK1 (Fig. 4). Mutations in JUN lead to GPX4 underexpression, whereas DNMT3L knockdown, AHR knockout, and NTRK1 overexpression lead to GPX4 overexpression. This interactome was enriched in genes associated

with cardiorespiratory arrest and myocarditis (P = 3.2E-03 and 1.52E-03; odds ratio = 23.06 and 32.9, respectively), suggesting regulatory network effects relevant to SMDS-associated cardiac phenotypes.

Telomere maintenance/organization and positive regulation of neurogenesis were detected both as global and fetus-specific modules. Six additional global modules were identified: cellular response to retinoic acid, detoxification, peptidyl-threonine phosphorylation, regulation of cellular protein localization, DNA replication and protein autophosphorylation. Three additional fetal modules were detected: negative regulation of myeloid cell differentiation, TGF-β receptor signaling pathway and positive regulation of endopeptidase activity.

3.6. *Tissue-elevated gene clusters in the integrated GPX4 landscape.*

We merged the interactomes described in the previous sections to construct an integrated GPX4 functional landscape containing 342 genes and 461 PPIs. We then clustered its genes hierarchically by GTEx tissue expression [40]. GPX4 and 59 genes clustering with it showed high expression in

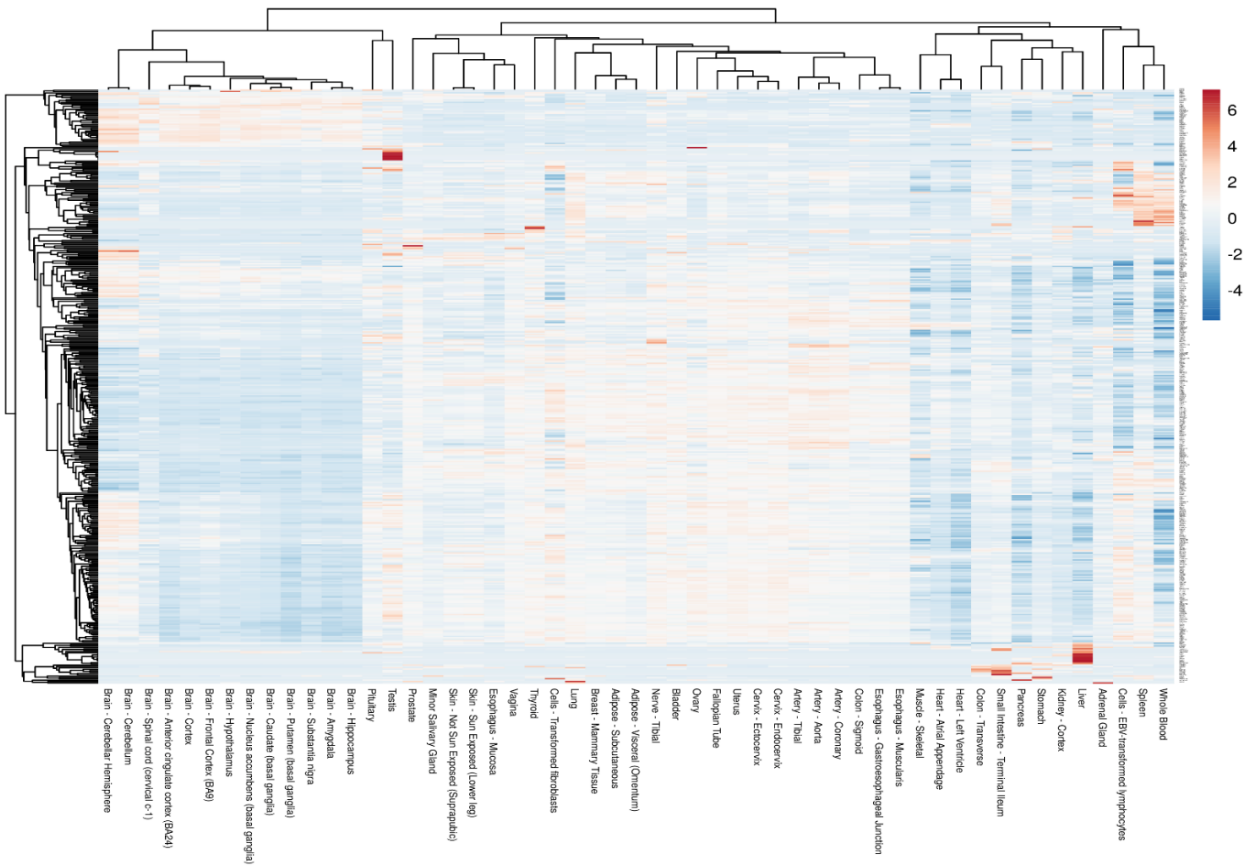


Figure 5: Clustering analysis of genes in the integrated GPX4 interactome. Tissue expression patterns of genes in the integrated **GPX4** interactome across 53 human tissues. The interactome comprises GPX4 and the shortest paths connecting it to genes associated with SMDS, other spondylometaphyseal dysplasias, and phenotypically similar skeletal dysplasias. Heatmap values represent normalized Z-scores of log-transformed TPM expression levels. Hierarchical clustering identified four major gene clusters with elevated expression in (i) brain/testis, (ii) whole blood/spleen, (iii) liver, and (iv) small intestine.

brain tissues or testis, with or without elevated expression in EBV-transformed lymphocytes and spleen (Fig. 5). This cluster included novel interactors of GPX4, namely, EGR4, MATK, and FUT5.

Brain-elevated genes were enriched for axon development, regulation of trans-synaptic signaling, regulation of ion transmembrane transport, protein localization to mitochondrion and neural nucleus development; testis-elevated genes for DNA recombination, aging, regulation of cyclin-dependent protein kinase activity, regulation of DNA metabolic process, mitotic cell cycle phase transition and extrinsic apoptotic signaling pathway; and small intestine-elevated genes for cell fate commitment, in utero embryonic development, epithelial cell proliferation, mRNA transcription, toxin transport and mesenchymal cell proliferation. None of these remained significant after multiple-testing correction, whereas the enrichments of liver-elevated genes (platelet degranulation and regulation of response to wounding) and of lymphocyte-, whole blood- and spleen-elevated genes (including leukocyte differentiation, leukocyte migration and adaptive immune response) did. We separately examined the interactome containing the intermediate interactors connecting the 4 genes (AHR, DNMT3L, JUN and NTRK1) whose perturbation affects GPX4 expression to GPX4, together with the known and novel interactors of these genes. We identified a group of nine testis-elevated genes that clustered with GPX4: FUT5, MX1, PDE6G, NCR1, MAGEC1, SUMO4, NTRK1, MSX2 and FKBP4.

3.7. *Interconnections of GPX4 and genes associated with Joubert syndrome*

Joubert syndrome with Jeune asphyxiating thoracic dystrophy (JATD) exhibited 64% phenotypic similarity with SMDS, sharing neurodevelopmental, skeletal, cardiac, and respiratory features. To investigate potential molecular links, we examined shortest paths connecting GPX4 to Joubert syndrome-associated genes (IFT140, KIAA0586, and CSPP1). The resulting network contained multiple novel GPX4 interactors and perturbation-associated genes, suggesting shared molecular pathways between these disorders (Fig. 6a). Extending this analysis to ciliary proteins identified TERF1 as the closest ciliary protein connected to GPX4, which may point to a potential role for

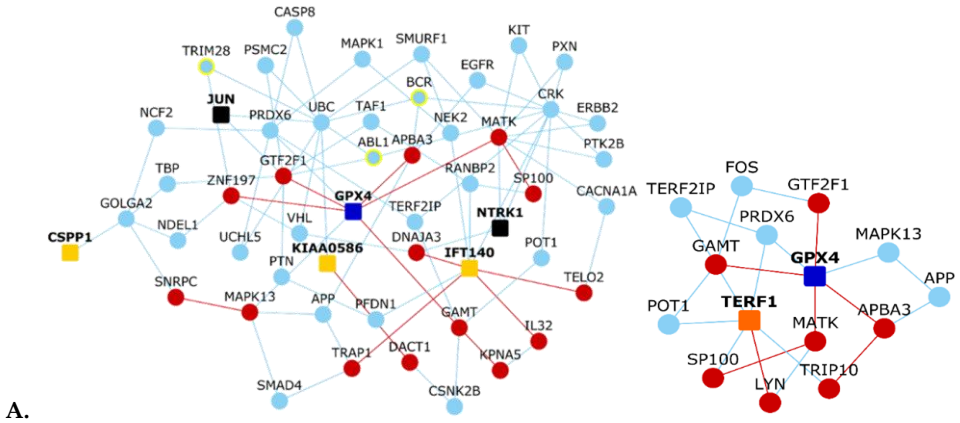


Figure 6: GPX4 interactions with ciliary proteins and Joubert syndrome-associated genes. (A) Shortest paths connecting GPX4 to three Joubert syndrome/Jeune asphyxiating thoracic dystrophy genes, together with genes whose perturbation affects GPX4, IFT140, or KIAA0586 expression. (B) Network of shortest paths connecting the ciliary protein TERF1 to GPX4. Nodes represent proteins and edges represent PPIs; red/light blue indicate novel/known interactors and PPIs.

ciliary biology in SMDS (Fig. 6b). Functional module analysis highlighted detoxification as both a global and fetus-specific module, together with enrichment of developmental, cell cycle, JAK–STAT signaling, proteostasis, and organelle-related processes.

3.8. *Repurposable drugs for SMDS*

We adopted two complementary strategies to identify repurposable drugs for SMDS using the integrated interactome of SMDS-associated genes. First, we applied a transcriptome-based drug repurposing approach that identifies compounds whose gene expression signatures are inversely correlated with disease-associated expression profiles, using the BaseSpace Correlation Engine [28, 43, 46]. An SMDS drug-protein interactome was then constructed, representing drugs that target one or more proteins within the SMDS interactome. Thirty-six drugs targeted 16 proteins in this network, including 5 novel interactors (PSMD8, XRCC1, DHODH, GHRL and VHL) (Fig. 7a). Four dysplasia-related gene expression datasets were selected, comprising two datasets corresponding to chondrodysplasia and two corresponding to immuno-osseous dysplasia (tibial growth plate hypertrophic zone of Cog and Schmid mice versus wild-type littermates, GSE30628 [47]; skin fibroblasts of Schimke immuno-osseous dysplasia cell lines SD60 and SD8 versus healthy controls, GSE35551 [48]). No SMDS expression datasets are available, so these dysplasias serve as proxies and may differ etiologically from SMDS. We then compiled a list of chemical compounds whose differential gene expression profiles (drug versus no drug) were negatively correlated with at least one of the four dysplasia-associated expression datasets (disease versus control). Using this approach, eleven candidate drugs were identified that may be evaluated for therapeutic value in SMDS (Fig. 7a): seven with expression profiles negatively correlated with chondrodysplasia and

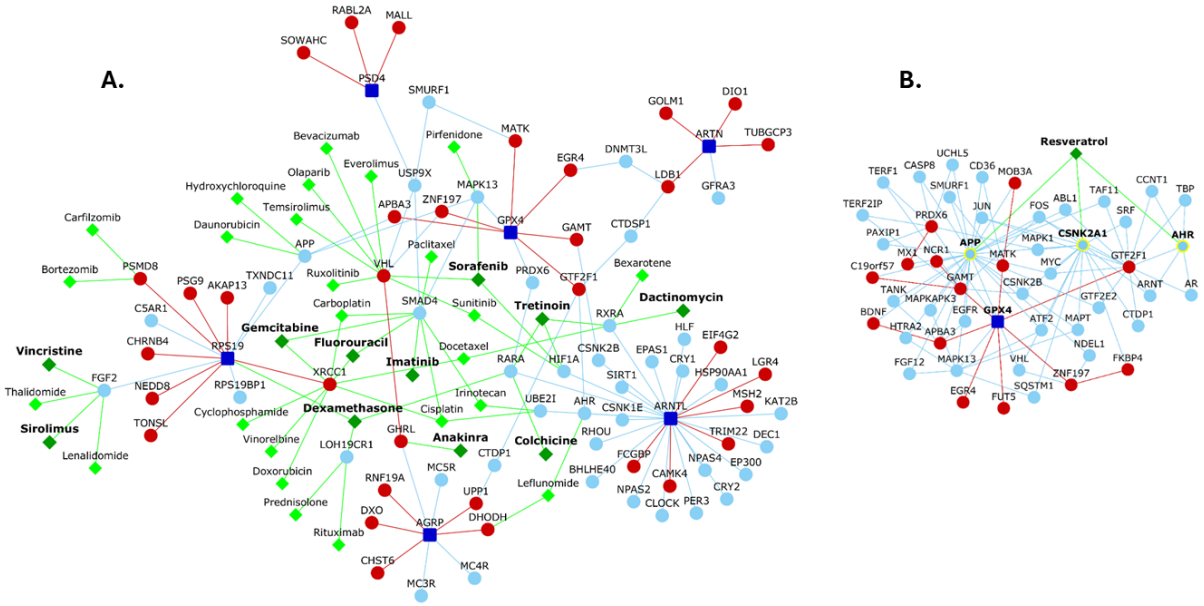


Figure 7: SMDS associated drug–protein interactome. (A) Drug–protein interactome linking drugs (green nodes) to proteins in the SMDS interactome; green edges denote drug–protein interactions. Repurposing candidates identified by comparative transcriptomic analysis are shown as dark green nodes with bold labels. (B) Network connecting GPX4 to the three targets of resveratrol (yellow-bordered nodes). Red and light blue nodes denote novel and known interactors, respectively; red and light blue edges denote novel and known PPIs, respectively.

ten with profiles negatively correlated with immuno-osseous dysplasia. Although subsets of genes were differentially expressed in the same direction for both the drug and disorder, the overall transcriptome-level signatures were anticorrelated. These drugs, namely, anakinra, colchicine, dactinomycin, dexamethasone, fluorouracil, gemcitabine, imatinib, sirolimus, sorafenib, tretinoin, and vincristine each target one to three proteins in the interactome.

In our second approach, we compiled a list of drugs that are currently approved for, or are undergoing phase I/II/III clinical trials in, different forms of skeletal dysplasias [49], and identified their protein targets using Drug Bank [44]. This resulted in eight drugs with 56 protein targets in the interactome; the drugs are etidronic acid, risedronic acid, denosumab (osteogenesis imperfecta), resveratrol (pseudoachondroplasia), carbamazepine (Schmid type metaphyseal dysplasia), asfotase alfa (hypophosphatasia), burosumab (X-linked hypophosphatemia), and N-acetylcysteine (diastrophic dysplasia). We examined the shortest paths connecting each of these protein targets to GPX4 and identified those exhibiting the closest network proximity. Three resveratrol targets – AHR, CSNK2A1 and APP – were connected to GPX4 via single intermediate interactors (Fig. 7b). Specifically, both AHR and CSNK2A1 were found to be connected to GPX4 via GTF2F1, a novel interactor of GPX4, whereas APP was connected via known (MAPK13) and novel (APBA3) interactors. Further supporting this network-based observation, resveratrol gene-expression profile was negatively correlated with the expression profile of immuno-osseous dysplasia.

4. Discussion

SMDS is an under-characterized skeletal dysplasia caused by GPX4 mutations. To expand the functional landscape of this disorder, we constructed both disease-centric and gene-centric neighborhood networks of GPX4, augmented with novel (hitherto unknown) interactors predicted by the HiPPIP algorithm. These complementary networks linked GPX4 to putative SMDS genes, related skeletal dysplasias, phenotypically similar disorders, and genes whose perturbation alters GPX4 expression. Together, they revealed consistent enrichment for SMDS-associated pathophenotypes, tissue-naïve and fetus-specific functional modules, brain/testis-enriched genes, and identified 11 candidate repurposable drugs.

Several novel interactors detected in this study have been linked to developmental deficits. EGR4 regulates hindbrain development [50], XRCC1 deficiency causes cerebellar abnormalities in mice [51], and LDB1 is required for embryonic patterning, cardiac development, and regulation of the skeletal dysplasia gene LMX1B [52, 53]. These independent observations strengthen the biological plausibility of the predicted GPX4 interaction network. Across all analyses, GPX4 emerged as a central node linking skeletal, neurodevelopmental, cardiovascular, and systemic processes. Enrichment of fetus-specific functional modules suggests developmental roles for GPX4 in cellular maintenance, cytoskeletal organization, apoptosis, and vascular development, consistent with the congenital manifestations of SMDS. The network also connected GPX4 to Joubert syndrome-associated genes and the ciliary protein TERF1, suggesting ciliary biology as a possible contributor to disease mechanisms. Enrichment of detoxification pathways in both global and fetal contexts further reinforces the established role of GPX4 in oxidative stress protection, while links

to telomere maintenance and TGF-β signaling suggest additional mechanisms warranting investigation.

The tissue-specific functional landscape of SMDS reveals strong expression of a gene cluster including GPX4 in both the brain and testis. Guo et al. have reported that the human brain and testis exhibit the highest similarity in gene expression patterns among 17 examined tissues [54]. This has been attributed to shared exocytotic processes and receptors, and to the involvement of both tissues in human speciation [55]. Functionally, the brain-elevated genes identified in this cluster were only nominally enriched (not significant after multiple-testing correction) for axon development, synaptic signalling, and mitochondrial protein localization – processes central to neural connectivity and plasticity – and would, if confirmed, be consistent with known roles of GPX4 in protecting neurons from oxidative damage. In contrast, testis-elevated genes were likewise only nominally associated with DNA recombination, cell cycle regulation, and apoptosis, which would be consistent with a role for GPX4 in germ cell maintenance and reproductive fitness. Of note, GTEx does not include bone or cartilage samples, so expression of these genes in skeletal tissue could not be assessed here. The skeletal relevance of the GPX4 interactome therefore rests on the enrichment for skeletal dysplasia phenotypes, the cartilage- and extracellular-matrix-related modules, and the links to known skeletal dysplasia genes described above, rather than on tissue expression; the relevance of the brain and testis signal to the skeletal phenotype of SMDS remains to be established.

We shortlisted 11 drugs for potential prioritization in SMDS. The prominence of anticancer drugs among the candidates likely reflects the involvement of essential cellular pathways common to both oncogenesis and skeletal dysplasia. Identification of fluorouracil, gemcitabine, and imatinib points to shared signaling pathways, including SMAD4 and TGF-β signaling [56], between dysplasia and oncogenic mechanisms. Seven of the eleven candidates (dactinomycin, fluorouracil, gemcitabine, imatinib, sirolimus, sorafenib and vincristine) are cytotoxic or broadly immunosuppressive agents, and SMDS presents in neonates, in whom such drugs carry substantial pediatric and developmental safety risks; the realistic therapeutic window for these compounds is therefore narrow, and we regard them as hypotheses about the pathways implicated here rather than as clinical candidates. The identification of anakinra and dexamethasone – both modulators of immune responses [57] – supports a role for immune dysregulation in SMDS pathophysiology. Network-based analysis further highlighted resveratrol, in our view the most defensible candidate because, unlike the cytotoxic agents above, it has reached a phase II trial in a related skeletal dysplasia (pseudoachondroplasia; NCT03866200; terminated due to recruitment challenges), implicating oxidative and inflammatory pathways [58] as potential intervention points. Finally, the observed network connections between SMDS and other skeletal dysplasias are consistent with the fact that the drugs used for osteogenesis imperfecta, hypophosphatasia, and pseudoachondroplasia target proteins linked to GPX4. Collectively, these findings nominate repurposable therapeutic candidates that warrant experimental validation. Because patient-derived material is not available, a realistic next step is to test the predicted GPX4 interactions and the highest-priority compounds (resveratrol first) in cellular or model-organism systems of GPX4 deficiency, carrying forward only candidates with mechanistic support and an acceptable pediatric safety profile. Since SMDS is typically lethal early in life, the immediate value of findings in this work is likely to lie in clarifying disease mechanism and in identifying pathways that would need to be targeted very early in

development; the same pipeline may be more directly actionable for rare genetic disorders with later onset and longer survival, where small patient populations similarly preclude conventional discovery.

This study has several limitations. First, all networks were assembled from shortest paths through the human interactome, which are biased toward highly connected, well-studied genes; enrichment for common phenotypes or for brain- and testis-elevated expression may therefore partly reflect properties of the interactome and of study bias rather than GPX4-specific biology. We did not perform a degree-preserving randomization (null model) to quantify this, and we did not compare our results against a baseline that uses only the direct interactors of GPX4 without shortest-path expansion. For the same reason, observations of network proximity (for example, of the ciliary protein TERF1 or of the resveratrol targets to GPX4) are reported descriptively and were not tested against degree-matched expectations. Second, the tissue-level functional enrichments did not remain significant after correction for multiple testing, and phenotype-enrichment P-values are nominal (uncorrected); both are presented as hypothesis-generating. Third, because no SMDS expression data exist, the transcriptome-based drug analysis uses chondrodysplasia and immuno-osseous dysplasia datasets as proxies, which assumes that the transcriptional signature of SMDS resembles theirs; this assumption is untested. Fourth, the seven novel GPX4 interactors are computational predictions that (albeit believed to be of high confidence) were not experimentally tested here, and the six DisGeNET-derived genes rest on literature co-mention with minimal supporting evidence.

The methods and tools used here have been applied previously by our group to other disorders, so the contribution of this work is their application to SMDS rather than a new method, and we did not benchmark the approach against alternative rare-disease gene-prioritization methods (such as network propagation or phenotype-driven tools), because no ground truth beyond GPX4 itself is available for SMDS. The novel interactors predicted by HiPPIP substantially expand the molecular landscape of GPX4 and provide experimentally testable hypotheses for elucidating the molecular mechanisms underlying SMDS. Together, our findings position GPX4 as a focal node at the intersection of skeletal dysplasia, neurodevelopmental disorders, and systemic homeostasis, providing a structured framework for prioritizing genes, pathways, and therapeutic hypotheses for this rare disease.

5. Supplementary Material

Additional details and vector graphics figures, which can be magnified without loss of resolution and searched by gene name, are available on Zenodo (https://doi.org/10.5281/zenodo.23091097) and at https://severus.dbmi.pitt.edu/gpx4/. All analyses were performed using web-based tools, so no scripts were used.

6. Author Contributions

MKG formulated the project idea to put the interactome-based analysis together motivated by the call “Let’s Find a Cure Together” (https://www.curegpx4.org/). KBK carried out the analysis of the interactomes, presenting literature-based evidence, supervised by NB and MKG. Manuscript has been drafted by KBK and edited by MKG.

References

1. Dumbuya JS, Zeng C, Deng L, Li Y, Chen X, Ahmad B, Lu J: **The impact of rare diseases on the quality of life in paediatric patients: current status**. *Frontiers in Public Health* 2025, **13**:1531583.
2. Weinreich SS, Mangon R, Sikkens JJ, Teeuw ME, Cornel MC: **[Orphanet: a European database for rare diseases]**. *Ned Tijdschr Geneeskd* 2008, **152**(9):518-519.
3. Ganapathiraju MK, Bhatia T, Deshpande S, Wesesky M, Wood J, Nimgaonkar VL: **Schizophrenia Interactome-Derived Repurposable Drugs and Randomized Controlled Trials of Two Candidates**. *Biol Psychiatry* 2024, **96**(8):651-658.
4. Kuppuswamy U, Ananthasubramanian S, Wang Y, Balakrishnan N, Ganapathiraju MK: **Predicting gene ontology annotations of orphan GWAS genes using protein-protein interactions**. *Algorithms Mol Biol* 2014, **9**(1):10.
5. Kuzmanov U, Emili A: **Protein-protein interaction networks: probing disease mechanisms using model systems**. *Genome Med* 2013, **5**(4):37.
6. Laddach A, Ng JC-F, Chung SS, Fraternali F: **Genetic variants and protein–protein interactions: a multidimensional network-centric view**. *Current opinion in structural biology* 2018, **50**:82-90.
7. Sahni N, Yi S, Taipale M, Bass JIF, Coulombe-Huntington J, Yang F, Peng J, Weile J, Karras GI, Wang Y: **Widespread macromolecular interaction perturbations in human genetic disorders**. *Cell* 2015, **161**(3):647-660.
8. Barabási A-L, Gulbahce N, Loscalzo J: **Network medicine: a network-based approach to human disease**. *Nature reviews genetics* 2011, **12**(1):56.
9. Sakai Y, Shaw CA, Dawson BC, Dugas DV, Al-Mohtaseb Z, Hill DE, Zoghbi HY: **Protein interactome reveals converging molecular pathways among autism disorders**. *Science translational medicine* 2011, **3**(86):86ra49-86ra49.
10. Lim J, Hao T, Shaw C, Patel AJ, Szabó G, Rual J-F, Fisk CJ, Li N, Smolyar A, Hill DE: **A protein–protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration**. *Cell* 2006, **125**(4):801-814.
11. Ganapathiraju MK, Thahir M, Handen A, Sarkar SN, Sweet RA, Nimgaonkar VL, Loscher CE, Bauer EM, Chaparala S: **Schizophrenia interactome with 504 novel protein–protein interactions**. *Schizophrenia* 2016, **2**:16012.
12. Karunakaran KB, Jain S, Brahmachari SK, Balakrishnan N, Ganapathiraju MK: **Parkinson's disease and schizophrenia interactomes contain temporally distinct gene clusters underlying comorbid mechanisms and unique disease processes**. *Schizophrenia* 2024, **10**(1):26.
13. Karunakaran KB, Chaparala S, Lo CW, Ganapathiraju MK: **Cilia interactome with predicted protein-protein interactions reveals connections to Alzheimer's disease, aging and other neuropsychiatric processes**. *Sci Rep* 2020, **10**(1):15629.
14. Malavia TA, Chaparala S, Wood J, Chowdari K, Prasad KM, McClain L, Jegga AG, Ganapathiraju MK, Nimgaonkar VL: **Generating testable hypotheses for schizophrenia and rheumatoid arthritis pathogenesis by integrating epidemiological, genomic, and protein interaction data**. *NPJ Schizophr* 2017, **3**:11.
15. Sedaghatian M, Opitz JM: **Congenital lethal metaphyseal chondrodysplasia: a newly recognized complex autosomal recessive disorder**. *American journal of medical genetics* 1980, **6**(4):269-274.
16. Opitz JM, Spranger JW, Stöss HR, Pesch HJ, Azadeh B, Reynolds JF: **Sedaghatian congenital lethal metaphyseal chondrodysplasia—observations in a second Iranian family and histopathological studies**. *American journal of medical genetics* 1987, **26**(3):583-590.
17. Peeden Jr J, Rimoin D, Lachman R, Dyer M, Gerard D, Gruber H: **Spondylometaphyseal dysplasia, Sedaghatian type**. *American journal of medical genetics* 1992, **44**(5):651-656.
18. Elçioglu N, Hall CM: **Spondylometaphyseal dysplasia–sedaghatian type**. *American journal of medical genetics* 1998, **76**(5):410-414.
19. Koutouby A, Habibullah J, Moinuddin FA: **Spondylometaphyseal dysplasia: Sedaghatian type**. *American journal of medical genetics* 2000, **90**(3):199-202.
20. English SJ, Gayatri N, Arthur R, Crow YJ: **Sedaghatian spondylometaphyseal dysplasia with pachygyria and absence of the corpus callosum**. *American Journal of Medical Genetics Part A* 2006, **140**(17):1854-1858.

21. Smith AC, Mears AJ, Bunker R, Ahmed A, MacKenzie M, Schwartzentruber JA, Beaulieu CL, Ferretti E, Majewski J, Bulman DE: **Mutations in the enzyme glutathione peroxidase 4 cause Sedaghatian-type spondylometaphyseal dysplasia**. *Journal of medical genetics* 2014, **51**(7):470-474.
22. Aygun C, Celik FC, Nural MS, Azak E, Kucukoduk Ş, Ogur G, Incesu L: **Simplified gyral pattern with cerebellar hypoplasia in Sedaghatian type spondylometaphyseal dysplasia: a clinical report and review of the literature**. *American Journal of Medical Genetics Part A* 2012, **158**(6):1400-1405.
23. Peshimam N, Farah H, Caswell R, Ellard S, Jan W, Calder AD, Cobben J, Kariholu U, Leitch HG: **Sedaghatian spondylometaphyseal dysplasia in two siblings**. *Eur J Med Genet* 2022, **65**(8):104541.
24. Cheff DM, Muotri AR, Stockwell BR, Schmidt EE, Ran Q, Kartha RV, Johnson SC, Mittal P, Arner ESJ, Wigby KM, Hall MD, Ramesh SK: **Development of therapies for rare genetic disorders of GPX4: roadmap and opportunities**. *Orphanet J Rare Dis* 2021, **16**(1):446.
25. Liu H, Forouhar F, Seibt T, Saneto R, Wigby K, Friedman J, Xia X, Shchepinov MS, Ramesh SK, Conrad M: **Characterization of a patient-derived variant of GPX4 for precision therapy**. *Nature chemical biology* 2022, **18**(1):91-100.
26. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, García-García J, Sanz F, Furlong LI: **DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants**. *Nucleic acids research* 2016, **45**(D1):D833–D839.
27. Warman ML, Cormier-Daire V, Hall C, Krakow D, Lachman R, LeMerrer M, Mortier G, Mundlos S, Nishimura G, Rimoin DL: **Nosology and classification of genetic skeletal disorders: 2010 revision**. *American journal of medical genetics Part A* 2011, **155**(5):943-968.
28. Kupershmidt I, Su QJ, Grewal A, Sundaresh S, Halperin I, Flynn J, Shekar M, Wang H, Park J, Cui W: **Ontology-based meta-analysis of global collections of high-throughput public data**. *PloS one* 2010, **5**(9):e13066.
29. Keshava Prasad T, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A: **Human protein reference database—2009 update**. *Nucleic acids research* 2008, **37**(suppl_1):D767-D772.
30. Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets**. *Nucleic acids research* 2006, **34**(suppl_1):D535-D539.
31. Dunham B, Ganapathiraju MK: **Benchmark Evaluation of Protein–Protein Interaction Prediction Algorithms**. *Molecules* 2022, **27**(1):41.
32. Zhu J, Zhang Y, Ghosh A, Cuevas RA, Forero A, Dhar J, Ibsen MS, Schmid-Burgk JL, Schmidt T, Ganapathiraju MK: **Antiviral activity of human OASL protein is mediated by enhancing signaling of the RIG-I RNA sensor**. *Immunity* 2014, **40**(6):936-948.
33. Karunakaran KB, Yanamala N, Boyce G, Becich MJ, Ganapathiraju MK: **Malignant pleural mesothelioma interactome with 364 novel protein-protein interactions**. *Cancers* 2021, **13**(7):1660.
34. Handen A, Ganapathiraju MK: **LENS: web-based lens for enrichment and network studies of human proteins**. *BMC Med Genomics* 2015, **8 Suppl 4**(Suppl 4):S2.
35. Embar V, Handen A, Ganapathiraju MK: **Is the average shortest path length of gene set a reflection of their biological relatedness?** *J Bioinform Comput Biol* 2016, **14**(6):1660002.
36. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome research* 2003, **13**(11):2498-2504.
37. Krishnan A, Zhang R, Yao V, Theesfeld CL, Wong AK, Tadych A, Volfovsky N, Packer A, Lash A, Troyanskaya OG: **Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder**. *Nature neuroscience* 2016, **19**(11):1454-1462.
38. Sealfon RSG, Theesfeld CL, Funk J, Sauerwald N, Wong AK, Troyanskaya OG: **HumanBase: an interactive AI platform for human biology**. *Nature Methods* 2026, **23**(3):483-484.
39. Shefchek KA, Harris NL, Gargano M, Matentzoglu N, Unni D, Brush M, Keith D, Conlin T, Vasilevsky N, Zhang XA, Balhoff JP, Babb L, Bello SM, Blau H, Bradford Y, Carbon S, Carmody L, Chan LE, Cipriani V, Cuzick A, Della Rocca M, Dunn N, Essaid S, Fey P, Grove C, Gourdine JP, Hamosh A, Harris M, Helbig I, Hoatlin M, Joachimiak M, Jupp S, Lett KB, Lewis SE, McNamara C, Pendlington ZM, Pilgrim C, Putman T, Ravanmehr V, Reese J, Riggs E, Robb S, Roncaglia P, Seager J, Segerdell E, Similuk M, Storm AL, Thaxon C, Thessen A, Jacobsen JOB, McMurry JA, Groza T, Kohler S, Smedley D, Robinson PN, Mungall CJ, Haendel MA, Munoz-Torres MC, Osumi-Sutherland D: **The Monarch Initiative in 2019: an integrative**

data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res* 2020, **48**(D1):D704-D715.
40. GTEx-Consortium: **The GTEx Consortium atlas of genetic regulatory effects across human tissues**. *Science* 2020, **369**(6509):1318-1330.
41. Metsalu T, Vilo J: **ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap**. *Nucleic acids research* 2015, **43**(W1):W566-W570.
42. Babicki S, Arndt D, Marcu A, Liang Y, Grant JR, Maciejewski A, Wishart DS: **Heatmapper: web-enabled heat mapping for all**. *Nucleic acids research* 2016, **44**(W1):W147-W153.
43. Chattopadhyay A, Ganapathiraju MK: **Demonstration Study: A Protocol to Combine Online Tools and Databases for Identifying Potentially Repurposable Drugs**. *Data* 2017, **2**(2):15.
44. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J: **DrugBank: a comprehensive resource for in silico drug discovery and exploration**. *Nucleic acids research* 2006, **34**(suppl_1):D668-D672.
45. Burl RB, Clough S, Sendler E, Estill M, Krawetz SA: **Sperm RNA elements as markers of health**. *Systems biology in reproductive medicine* 2018, **64**(1):25-38.
46. Sirota M, Dudley JT, Kim J, Chiang AP, Morgan AA, Sweet-Cordero A, Sage J, Butte AJ: **Discovery and preclinical validation of drug indications using compendia of public gene expression data**. *Science translational medicine* 2011, **3**(96):96ra77-96ra77.
47. Cameron TL, Bell KM, Tatarczuch L, Mackie EJ, Rajpar MH, McDermott BT, Boot-Handford RP, Bateman JF: **Transcriptional profiling of chondrodysplasia growth plate cartilage reveals adaptive ER-stress networks that allow survival but disrupt hypertrophy**. *PLoS One* 2011, **6**(9):e24600.
48. Baradaran-Heravi A, Cho KS, Tolhuis B, Sanyal M, Morozova O, Morimoto M, Elizondo LI, Bridgewater D, Lubieniecka J, Beirnes K, Myung C, Leung D, Fam HK, Choi K, Huang Y, Dionis KY, Zonana J, Keller K, Stenzel P, Mayfield C, Lucke T, Bokenkamp A, Marra MA, van Lohuizen M, Lewis DB, Shaw C, Boerkoel CF: **Penetrance of biallelic SMARCAL1 mutations is associated with environmental and genetic disturbances of gene expression**. *Hum Mol Genet* 2012, **21**(11):2572-2587.
49. Marzin P, Cormier-Daire V: **New perspectives on the treatment of skeletal dysplasia**. *Therapeutic advances in endocrinology and metabolism* 2020, **11**:2042018820904016.
50. Bae C-J, Jeong J, Saint-Jeannet J-P: **A novel function for Egr4 in posterior hindbrain development**. *Scientific reports* 2015, **5**(1):1-10.
51. Hoch NC, Hanzlikova H, Rulten SL, Tétreault M, Komulainen E, Ju L, Hornyak P, Zeng Z, Gittens W, Rey SA: **XRCC1 mutation is associated with PARP1 hyperactivation and cerebellar ataxia**. *Nature* 2017, **541**(7635):87-91.
52. Mukhopadhyay M, Teufel A, Yamashita T, Agulnick AD, Chen L, Downs KM, Schindler A, Grinberg A, Huang S-P, Dorward D: **Functional ablation of the mouse Ldb1 gene results in severe patterning defects during gastrulation**. *Development* 2003, **130**(3):495-505.
53. Dreyer SD, Morello R, German MS, Zabel B, Winterpacht A, Lunstrum GP, Horton WA, Oberg KC, Lee B: **LMX1B transactivation and expression in nail–patella syndrome**. *Human molecular genetics* 2000, **9**(7):1067-1074.
54. Guo J, Zhu P, Wu C, Yu L, Zhao S, Gu X: **In silico analysis indicates a similar gene expression pattern between human brain and testis**. *Cytogenetic and genome research* 2003, **103**(1-2):58-62.
55. Matos B, Publicover SJ, Castro LFC, Esteves PJ, Fardilha M: **Brain and testis: more alike than previously thought?** *Open Biology* 2021, **11**(6):200322.
56. Wang Q, Xiong F, Wu G, Wang D, Liu W, Chen J, Qi Y, Wang B, Chen Y: **SMAD proteins in TGF-β signalling pathway in cancer: regulatory mechanisms and clinical applications**. *Diagnostics* 2023, **13**(17):2769.
57. Mehta P, Cron RQ, Hartwell J, Manson JJ, Tattersall RS: **Silencing the cytokine storm: the use of intravenous anakinra in haemophagocytic lymphohistiocytosis or macrophage activation syndrome**. *The Lancet Rheumatology* 2020, **2**(6):e358-e367.
58. Meng T, Xiao D, Muhammed A, Deng J, Chen L, He J: **Anti-inflammatory action and mechanisms of resveratrol**. *Molecules* 2021, **26**(1):229.

Leveraging host omic-response signatures to identify microbiome metabolizers of drugs in the vaginal and gut microbiome

Dali Nemecio† and Douglas K. Brubaker
Department of Biomedical Engineering, School of Medicine,
Case Western Reserve University, 10900 Euclid Avenue,
Cleveland, OH 44106, United States of America
Email: dxn254@case.edu, dkb50@case.edu

The human microbiome interacts with drugs and contributes to side effects and loss of efficacy. Methods to identify strains with potential to biotransform drugs are inefficient, often fail to scale, have species-level specificity, and rarely account for host factors. We present an *in silico* method to predict the microbial metabolism of drugs using a Microbe-Drug-Host (MDH) signature. The signature integrates non-treated human omics, microbiome composition, and drug-treated *in vitro* transcriptomics to capture features that predict metabolic potential of a given microbe-drug pair. We built support vector machines (SVM) using MDH signatures to predict drug metabolism by vaginal and gut microbes. We established the framework with a single-drug vaginal test case and predicted metabolism of tenofovir with models with high performance metrics (AUC, precision, accuracy, recall, specificity equal to and above 0.84). Generalizing to multidrug gut models, 27 drugs across 130 bacteria were used to generate metabolism probabilities with overall high performance metrics with AUC, precision, accuracy, specificity above or equal to 0.81 and recall at 0.75. The gut microbes most implicated to metabolize drugs include *Bacteroides cellulosilyticus, Bacteroides thetaiotaomicron, and Blautia hansenii*, consistent with existing literature. Our method using the MDH signature addresses deficiencies of existing methods by providing a disease-agnostic, tissue-specific method to identify microbe-drug metabolism interactions.
Keywords: machine learning, microbiome, drug metabolism, host-microbiome interactions

1. Background

The microbiome dynamically contributes to host heath outcomes in beneficial and deleterious manners. The gut microbiome aids in digestion of dietary fibers, synthesizes micronutrients, and assists with bile metabolism and creates a barrier of protection against pathogens.[1–14] In the vaginal microbiome, bacterial strains provide protection against pathogens by maintaining a low pH environment through lactic acid production and releasing anti-microbial compounds such as hydrogen peroxide and bacteriocins.[15–17] Microbiomes can biotransform exogenous drugs through reduction or oxidation reactions, resulting in side effects on the host such as toxicity or reduced treatment efficacy.[18,19] These interactions have been best characterized in the gut microbiome.[20–23] For instance, sulfasalazine requires an azo bond cleavage facilitated by gut microbiota-secreted azoreductases to become activated. [24,25] Digoxin has a well-established inactivating metabolic relationship with *E. lenta* strains that possess a two-gene cytochrome-encoding operon.[26–29] In the

† Work supported by Startup funding granted by Case Western Reserve University and University Hospitals Cleveland Medical Center
© 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.

vaginal microbiome, the efficacy of the HIV pre-exposure prophylaxis drug tenofovir (TV) is compromised through metabolism by microbes such as *Gardnerella vaginalis*, *Prevotella* (*bivia* and *amnii*), *Mobiluncus mulieris* and *Escherichia coli. These microbes* are implicated in increasing HIV incidence rate to 61%, compared to 18% in *Lactobacillus*-dominant microbiomes.[30 31] In both the gut and vagina many drugs experience biotransformation by the whole microbiome, but few explicit bacterial strains have been identified. This remains a blind-spot for pharmacology towards comprehensively understanding the impact the microbiome has on drug therapies.[19,32,33]

High-throughput methods for identifying microbe-drug interactions are extremely limited. Experimental methods use *in vitro* and *ex vivo* whole microbiome and single-species culturing with drugs.[34 35] Computational methods for inference scale more effectively, but existing tools primarily predict *human* enzymatic interactions with drugs, and rarely predict *bacterial* metabolic capacity.[36–38 39,40] Models that predict drug-microbiome interactions and resulting metabolic products primarily explore whole-microbiome meta-metabolome output, or use functional annotations or metagenomic screens to identify bacteria carrying pre-characterized enzymes.[41–45 46] *In silico* models specific to microbial metabolic products primarily use physiologically based pharmacokinetic (PBPK) models and constraint-based reconstruction and analysis (COBRA).[47 48] Vaginal models that predict bacterial metabolic capacity primarily use the metabolome to explore host-microbiome interactions using genome-scale metabolic models and flux balance analysis methods.[49,50 51] Gut models are more commonly found with wider applications and input data types. Some models explore prediction of drug metabolic product but only in the context of therapeutic effect in disease states.[47,52] Models using ML methods require pre-existing molecular feature identification of the drugs (such as SMILE notations).[53,54] Models like AGORA and AGORA2 identify species-level drug biotransformations but require manual curation of tissue-specific bacterial species, rather than utilizing tissue-derived microbiome composition.[44,55] Other models instead focus on biotransformation due to host diet or microbe population cross-feeding.[56–61] Overall, most existing vaginal and gut models identify whole-microbiome biotransformations and few are species-specific. Our method does not require manual curation of bacterial species, nor pre-identified drug molecular features, and is disease-agnostic. There is no existing *in silico* methods that integrate *both* host multi-omics and *in vitro* data, making our method exceptionally unique in its approach.

Here, we develop a novel computational framework using host, microbiome, and *in vitro* drug screening data to predict species-specific metabolism of drugs across multiple microbiomes using a “microbe-drug-host (MDH)” signature. Each microbe-drug pair contains a unique MDH signature that integrates the human, microbe, and reference drug data into a set of class-separating features that best reflects the metabolic relationship. We establish the framework with a case study of vaginal microbiome metabolism of TV and generalize the pipeline to a multidrug classifier for species-specific drug metabolism in the gut microbiome. This method takes advantage of the non-disease state of human *in vivo* omics combined with high-throughput *in vitro* drug transcriptomics to create a scalable framework to support large-scale screening and prediction for clinically meaningful microbe-drug metabolic relationships.

2. Materials and Methods

2.1. Pipeline

Fig. 1 depicts overall data preprocessing, feature selection, testing and training, and performance.

2.1.2. Data pre-processing

We generated a matrix that contains the integrated host omics, microbiome composition, and *in vitro* drug-treated transcriptomics information. Human transcriptomics or proteomics and paired microbiome composition undergo a Spearman correlation to produce a microbe-host ρ matrix. These ρ values are multiplied by each drug's Library of Network-based Cellular Signatures characteristic direction (LINCS CD) to produce a product matrix. The CD is a mean gene expression over many cell types and treatment conditions. The point-wise product combines drug-treated *in vitro* data with *in vivo* microbe-host ρ. The relationship of a microbe-drug pair is denoted in the model with a binary class, +1 indicates a metabolic and -1 a non-metabolic relationship.

Microbe-drug pairs without literature-established relationships, denoted class of NA, are considered unknown and excluded from the training and testing data. Any microbe-drug pairs with unknown metabolic class are sequestered to the auxiliary predictions dataset which will be used for microbe-drug pair predictions after the model is generated.

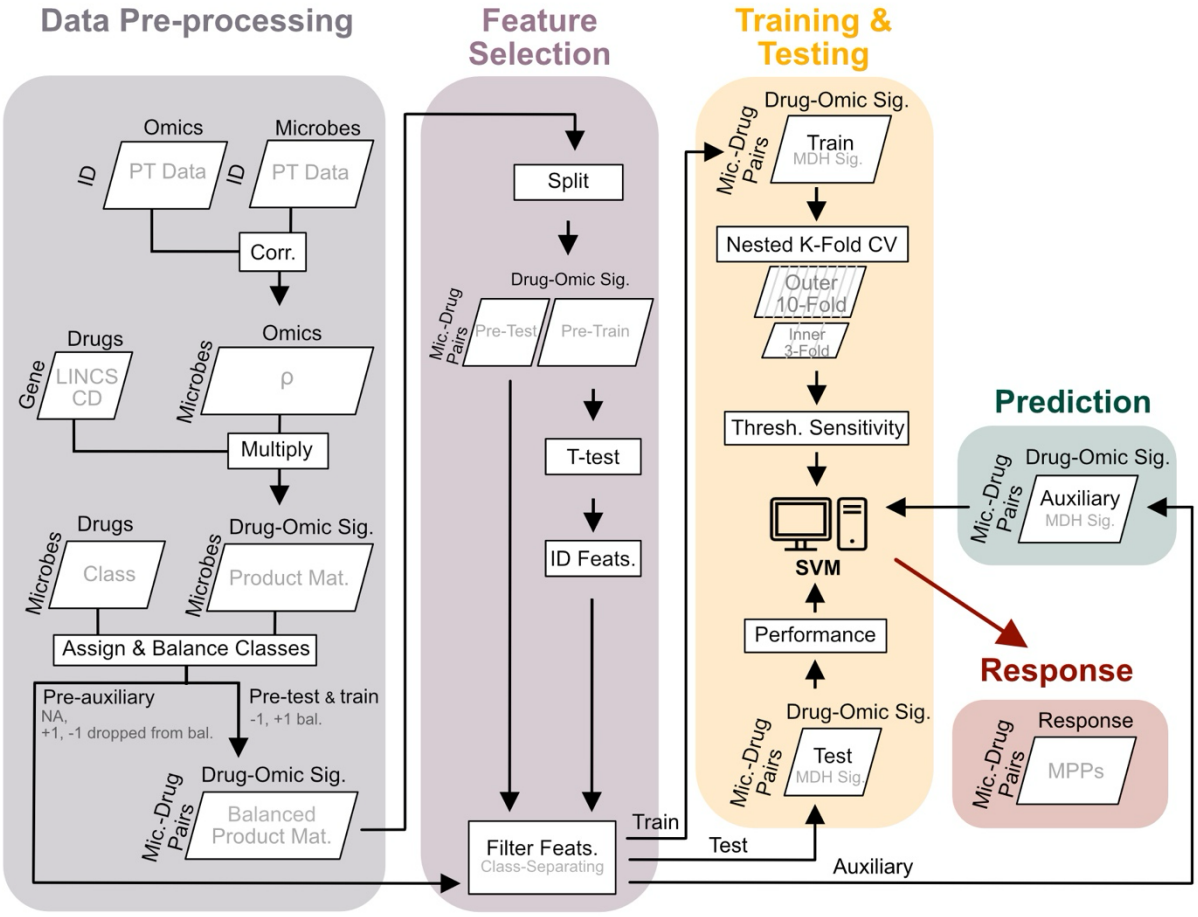


Fig.1: Generic pipeline to generate the MDH signature and SVM model to predict the metabolic relationship of the microbe-drug pairs. The Tv-VagGene, TV-VagProt and MultiDrugV1 and V2 models are adapted from this pipeline.

2.1.3. Feature selection of the MDH signatures

We split the product matrix into testing and training data (split 20/80% respectively). The training data undergoes feature selection against the classes associated with the microbe-drug pairs via T-test with Benjamini-Hochberg false discovery rate (FDR) correction. Features with the lowest adjusted p-values define and compose the MDH signatures for model training. SVMs lack built-in feature selection, and by reducing data dimensionality we mitigate potential model overfitting.

2.1.4. Training and testing

A SVM model was selected due to its ability to handle high dimensional data well and its inherent property of identifying the largest margin of separation between classes of interest. We used 10-fold cross validation (CV) with nested 3-fold hold out for hyperparameter optimization and determined performance metrics which include precision, accuracy, recall, and specificity for both test and training data. Overall performance calculated from the aggregated CV data uses a prediction threshold of probability 0.5, where >=0.5 is metabolic and < 0.5 probability is non-metabolic. We performed sensitivity analysis to determine the optimal threshold for each metabolic class at 0.05 increments. Optimal prediction thresholds were selected based on best F1 score for metabolic prediction responses and negative predictive value for non-metabolic prediction responses, with recall and specificity used to resolve ties respectively. If ties are not resolved, the lowest threshold for the metabolic class is selected and the highest threshold for the non-metabolic class is selected.

2.1.5. Predictions and response

Novel predictions are made with the optimized model using microbe-drug pairs from the auxiliary data that were excluded from the testing and training data. We assessed the prediction responses of the model independently on the testing and training data in addition to the auxiliary predictions. Predictions contain metabolic prediction probabilities (MPPs) for each microbe-drug pair based on the respective MDH signature. Prediction responses are finalized based on the optimal prediction thresholds determined from the sensitivity analysis. MPPs and the corresponding microbe-drug pairs that fall between the metabolic and non-metabolic classification thresholds are considered indeterminate.

2.2 Vaginal Data Participant Cohort

The vaginal models used HIV negative vaginal swab samples from the Partner's PrEP study.[30,62,63] The TV-VagGene cohort contains 20 woman (25-47 years old) with 27 microbes represented and 9451 genes in vaginal tissue transcriptomics. (Fig. S1A) The TV-VagProt cohort contains 192 woman (18-49 years old) with 150 microbes represented and 434 identified proteins from host mucosa. (Fig. S1C)

2.3 Gut Data Participant Cohort

The gut models use 10 control, non-IBD adults from the Inflammatory Bowel Disease Multiomics Database (IBDMDB) as part of the Integrative Human Microbiome Project.[64] In total, 19 samples of ileum and rectum tissue transcriptomics with corresponding microbiome abundance was used. In the microbiome composition, 27 of the microbes were also found in the *in vitro* data, which will assign the microbe-drug "true" relationship for the testing and training data. (Fig. S1B) This is further described below in the 2.4.2 *Preparing Zimmerman Data* section.

2.4 Adapted Pipelines

The pipelines to generate the four models follow the same steps of data preprocessing, feature selection, testing and training, prediction and response. The vaginal models add bootstrapping and a modified feature selection, with testing and training data partitioned earlier in the pipeline. The gut model uses Zimmerman et al 2019 *in vitro* data from a microbe-drug screen to define which microbe-drug pairs to include in the testing and training data.[65] Adapted pipeline details are visualized in Fig. S2.

2.4.1 Vaginal Models Pipeline

Pre-processing with Bootstrapping

The vaginal models, TV-VagGene and TV-VagProt, undergo modified data preprocessing steps that include early testing and training partitioning, and bootstrapping as depicted in Fig. S2A. For Tv-VagGene we generated class-balanced training data from 15 participants and test data from 5 participants via independent bootstrapping. For Tv-VagProt we generated a class-balanced training data from 144 participants and test data from 48 participants via independent bootstrapping. The participant omics data and microbe composition ρ were bootstrapped for 100 iterations with 3 or 10 participants selected for each iteration to produce a single microbe-host ρ matrix for the training and test data respectively. The microbe-host correlations were multiplied by the LINCS CD to generate the product matrix. These steps are repeated to generate a set of 10 data frames. Metabolic relationships of TV were based on *in vitro* validated metabolic capabilities defined by Klatt et al 2017.[30] Microbes with metabolic capacity for TV include *G. vaginalis, M. mulieris, P. amnii, P. Bivia,* and *E. coli*. Microbes with non-metabolic capacity for TV include *L. crispatus* and *L. iners.* These classes are balanced in the training data and then filtered to exclude microbes with unknown metabolic capacity.

Feature Selection

Each of the 10 data frames in the training data underwent feature selection, where features with lowest 500 adjusted p-values were selected. The same features found across all 10 data frames were chosen as the MDH signature. This allows for identification of *consistent* features that distinguish the metabolic relationships *independent* of the noise introduced with bootstrapping. The first data frame is arbitrarily chosen to represent the training partition to build the model. TV-VagProt training and testing data both represented 7 microbes with the respective class composition, (200 +1, 200 -1) and (200 +1, 497 -1). TV-VagGene training data represented 4/3 microbes with the respective class composition (169 +1, 170 -1) and (195 +1, 100 -1).

2.4.2 Gut model's in vitro-defined relationships

The literature-established metabolic relationships needed for building the gut models is defined by *in vitro* data published by Zimmerman et al 2019.[65] This group assessed drug-metabolic activity of 76 human gut bacteria against 271 oral drugs, and reported percentage of drug consumed for each microbe-drug pair. The preprocessing is outlined in Fig. S2B. For our application, this *in vitro* data must be filtered to contain the same drugs and microbes found in the LINCS CD and IBDMDB cohorts respectively. The resulting filtered *in vitro* Zimmerman matrix contains 27 microbes shared with the IBDMDB cohort, and 130 drugs shared with LINCS CD.

Microbe Proportion and Percent Consumed Thresholds

We apply to the prepared *in vitro* Zimmerman matrix the *microbe proportion threshold* and *percent consumed threshold.* These thresholds impose cut-offs on the percent consumed values to in order to assign metabolic classes to the microbe-drug pairs. The first threshold is the *microbe proportion threshold,* which specifies for each drug the percent of microbes with a nonzero value of drug consumed. For example, a threshold of 0.25 means 25% of the 27 microbes tested with a drug contained *some* drug consumed. Any drugs and their respective microbes that do not meet this threshold are assigned a metabolic class of NA, and are dropped from the testing and training data.

The *percent consumed threshold* is applied to the remaining drugs and microbes, which specifies the percent of drug consumed to be considered metabolized by the respective microbe. For example, a threshold of 0.50 means 50% or more of the drug must be consumed to be considered a metabolic relationship and assigned a "true" class of +1, otherwise it is denoted as non-metabolic and assigned a class of -1. Each microbe-drug pair in the filtered *in vitro* Zimmerman matrix will have a metabolic class assigned based on the *percent consumed threshold*. The metabolic relationships are either metabolic (+1 class), non-metabolic (-1 class), or unknown (NA class). Microbe-drug pairs with 1) unknown metabolic relationships and 2) pairs dropped due to class balancing, compose the auxiliary data and are used for prediction at the end of the pipeline. Multi-GutV1 model contains *microbe proportion threshold* of 0.50 and *percent consumed threshold of 0.50*. After sensitivity analysis was conducted on these thresholds, Multi-GutV2 contains *microbe proportion threshold* of 0.25 and *percent consumed threshold of 0.50*. For each gut model the testing, training, and auxiliary data together account for 3510 unique combinations of microbe-drug pairs derived from 27 taxa and 130 drugs. Multi-GutV1 training and test data respectively contain representation of 27/25 microbes, 18/17 drugs, with a class composition of (120 +1, 111 -1) and (31 +1, 28 -1). Multi-GutV2 training and test data respectively contain representation of 27/27 microbes, 49/48 drugs, with a class composition of (310 +1, 326 -1) and (78 +1, 82 -1).

3. Results

3.1 Vaginal Models' Performance

Feature selection identified 455 and 91 features for the TV-VagGene and TV-VagProt model respectively. The hyperparameter optimization determined the TV-VagGene model to contain the following parameters: kernel = polynomial, degree = 1, cost=12.1, and gamma=1.41. The hyperparameter optimization determined the TV-VagProt model to contain the following parameters: kernel = polynomial, degree = 3, cost = 8.55, and gamma = 0.507. The performance metrics determined from the testing and aggregated CV training folds for both models are comparable and reflect useful classification from both methods. Notably, precision, accuracy, recall and specificity are all above 0.84 from the testing data and above 0.99 for the aggregated CV training data (Fig. 2C, Fig. S3B) Both models reflect an area under the curve (AUC) above 0.99 in their respective testing and aggregated CV fold performances. The recall performance indicates that both models can identify metabolic microbes within the dataset (0.84 TV-VagGene, 0.99 TV-VagProt). Similarly, the high specificity indicates reliable identification of non-metabolic prediction responses (1.00 TV-VagGene, 0.87 TV-VagProt). The patient pool bootstrapped to produce the training data cannot introduce biological variation beyond what is represented in the original cohort. The small patient pool results in limited heterogeneity and potentially reduces the generalizability of the model to unseen data. However, the learning curve for both models reflects

negligible overfitting, as the generalization gap between the accuracy and mean misclassification error are less than 0.01 for both models. (Fig. S5)

3.2 Vaginal Model Sensitivity Analysis

Sensitivity analysis on the classification threshold in both TV-VagGene and TV-VagProt reflect minimal impact on the overall performance. (Fig. 2A/B) The TV-VagGene model optimal prediction threshold for response is 0.50, with an associated F1 value of 0.91 and recall of 0.84. The non-metabolic prediction threshold is also 0.50 with a determined negative predictive value of 0.92 and specificity of 1.00. The TV-VagProt model optimal metabolic prediction threshold is 0.80 with an associated F1 value of 0.98 and recall of 0.97. The nonmetabolic prediction threshold is 0.25 with an associated negative predictive value of 0.99 and specificity of 0.79.

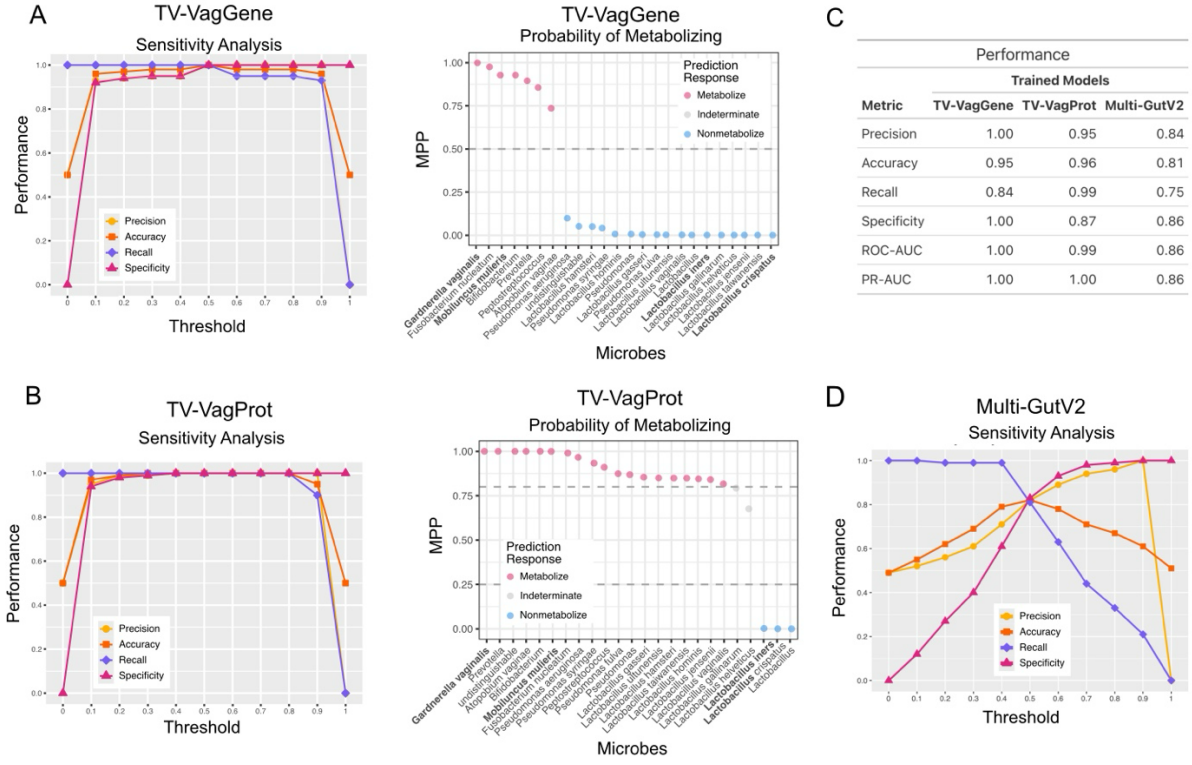


Fig. 2: Vaginal model results and performance metrics from TV-VagGene, TV- VagProt, and MultiDrug-V2. A/B) The sensitivity analysis on the classification threshold are determined from the aggregated CV folds. The MPPs (A, B right) for all direct participant MDH signatures are depicted for each model. Microbes found in both transcriptomics and proteomics data are displayed in the plots. Bolded microbes were used in testing and training. Complete predictions are found in Supplementary Tables 1 and 2. Optimized prediction thresholds determined from the sensitivity analysis are denoted as horizontal dashed lines. C) Performance metrics from three models are depicted from test data on threshold of 0.5. (D) The sensitivity analysis on Multi-GutV2 classification threshold for aggregated CV folds reflects optimal performance at 0.45.

3.3 Vaginal Model Predictions

TV-VagGene and TV-VagProt model prediction responses on various microbes reflect a range of metabolic and non-metabolic responses. (Fig. 2 A/B) Both MPP plots reflect predictions in shared microbes from the TV-VagGene and TV-VagProt cohorts, all microbe MPPs are in Table 1 and 2 respectively. As expected, microbes used for training are predicted correctly with MPPs that support the response. Microbes predicted to metabolize TV includes *G. vaginalis* (1.00 MPP, 1.00 MPP), and *M. Mulieris* (0.93 MPP, 1.00 MPP) for TV-VagGene and TV-VagProt respectively. TV-VagProt also predicted metabolic response for *P. amnii, P. Bivia,* (both 1.00 MPP) and *E. coli* (0.86 MPP). Prediction response of nonmetabolic class is as expected for *L. Crispatus* (5.6e-4 MPP,

3.5e-4 MPP) and *L. iners* (1.2e-3 MPP, 2.4e-3 MPP) for TV-VagGene and TV-VagProt respectively. Comparatively, TV-VagGene returns more conservative predictions across all microbes, despite this, there are consistent predictions between both models.

Novel metabolic predictions shared by both models include *A. vaginae* (0.74 MPP, 1.00 MPP*)*, Bifidobacterium species (0.93 MPP, 1.00 MPP), Prevotella species (0.90 MPP, 1.00 MPP) and Peptostreptococcus species (0.86 MPP, 0.91 MPP) from TV-VagGene and TV-VagProt respectively. Notably, Bifidobacterium and Atopobium genus are highly predicted to be metabolizers of TV. Due to the lack of availability in the cohort data, the TV-VagGene model was not trained on *P. amanii* and *P. bivia*; However, the model correctly identifies the metabolic potential of Prevotella genus. This supports the integrity of the method as the identification of this genus is correct, despite its absence from the training data.

The two models agreed on non-metabolic prediction of Lactobacillus genus. However, there are 11 microbes with conflicting predictions, primarily where TV-VagProt response predicts metabolic relationship (MPP range 0.82-1.00) and TV-VagGene does not (MPP range 1.0e-3 to 0.10). These include Pseudomonas species*, P. aeruginosa, P. syringae, P. fulva, L. gasseri, L. ultunensis, L. hamsteri, L. taiwanensis, L. hominis, L. jensenii,* and *L. vaginalis.* A summary of the predictions between the TV-VagGene and TV-VagProt models is depicted in Fig. S3C.

3.4 Multi-GutV1

The results from Multi-GutV1 model with the *microbe proportion* threshold of 0.5 and *percent consumed* threshold of 0.5 depicts decent performance. (Fig. S6) The model's parameters are: kernel = polynomial, degree = 2, cost = 19.2, gamma = 0.491 with 500 features. The testing fold performance depicts a range of 0.75 - 0.77 for precision, accuracy, recall and specificity. (Fig. S6A) The AUC of the sensitivity and specificity curve and precision-recall curve are 0.83 and 0.87 respectively. These performance values and AUC support moderate performance by the model, and its ability to distinguish the metabolic class. (Fig. S6C) The sensitivity analysis on the prediction threshold reflects a 0.5 threshold is optimal for moderate performance across metrics. (Fig. S6B) The model's predictions grouped by drug reflect higher mean MPPs compared to grouping by microbe. (Fig. S6D/E) This reflects that MPPS are more influenced by the drug type rather than microbe species.

3.5 Multi-GutV2 Performance

During Multi-Gut model development we interrogated *microbe proportion threshold* and *percent consumed threshold* to identify a model with improved performance. Denoted as Multi-GutV2, this model contains a *microbe proportion threshold* of 0.25 and a *percent consumed threshold* of 0.50. The model's parameters are: kernel = polynomial, degree = 2, cost = 15.3, gamma = 0.83 and contains 500 features. This improved Multi-GutV2 model contains performance metrics of precision, accuracy, recall, and specificity between 0.75-0.86. (Fig. 2C, Fig. S7A) The learning curves reflect some minor potential overfitting with a generalization gap of 0.21, likely due to the high model complexity. (Fig. S5) The model built on a balanced dataset can differentiate between the classes well (accuracy 0.81, specificity 0.86) and most positive predictions are correct (precision 0.84). This provides reassurance the prediction responses may provide meaningful MPPs worth interpreting. Sensitivity analysis on the prediction threshold reflects optimal performance at prediction threshold of 0.45 (F1 0.79, recall 0.83). (Fig. 2D) The optimal performance for nonmetabolic class is prediction threshold of 0.25 (negative predictive value 0.9,

specificity 0.31). Performance metrics and confusion matrices for the aggregated CV folds are depicted in Fig. S3A/B.

3.6 Multi-GutV2 Trends in Responses

The Multi-GutV2 responses from the testing and training data show metabolic trends when MPPs are grouped by drugs rather than microbes. When grouped by microbes, the two leading microbes with the highest mean MPPs are *B. cellulosilyticus* and *B thetaiotaomicron* (mean MPP 0.58, mean MPP 0.55). (Fig. S7B, Table 3) Interestingly, when the same MPPs are grouped by drug identity, certain drugs are favored for metabolic relationships by multiple species of microbes. (Fig. 3A, Table 3) The leading drugs with highest mean MPPs are artemisinin and tenatoprazole (mean MPP 0.88, mean MPP 0.84). 37% (18/49) of drugs contain mean MPPs that are above the metabolic prediction threshold of 0.45. There are no structural or pharmaceutical classes overrepresented in these 18 drugs. However, some of these drugs share metabolic mechanisms, specifically involving hydrolysis and reduction reactions. These 18 include prodrugs that require activation and contain the respective mean MPPs which include: bisacodyl (0.76), sulfasalazine (0.61), famciclovir (0.59), racecadotril (0.56), deflazacort (0.51). These also include known drug substrates for reduction and cleavage with the following mean MPPs: sulfasalazine (0.61), phenazopyridine (0.62), and tinidazole (0.75).

The predictions on microbe-drug pairs from the testing data reflects consistent classification of the model's response and the corresponding true class. For microbe-drug pairs assigned a class "truth" of metabolic, the median MPP is 0.60 and median percent consumed is 0.96. For the "true" non-metabolic microbe-drug pairs, the median MPP is 0.13 and median percent consumed is 0.0. A Spearman correlation between the MPPs for testing microbe-drug pairs and the corresponding percent consumed *in vitro* value returns a ρ of 0.53 with a p-value of 7.3e-12. The positive direction, meaningful magnitude and notable p-value reflect the *in vitro* percent consumed values associated with each microbe-drug pair are highly correlated to the MPP produced by the model. This correlation reflects the MDH signature can capture wide metabolic range, and corresponding MPP's are reflective of measured activity in the *in vitro* values used to assign the classes. Together, these values support the useful classification abilities of the model.

Compared to the testing and training MPPs, the predictions on the auxiliary microbe-drug pairings assigned a truth of NA reflect a smaller range of MPPs. This is evident when grouping the MPPs by microbe and by drug. The two leading microbes identified to metabolize the drugs include *B. hansenii* (mean MPP 0.50) and *B. thetaiotaomicron* (mean MPP 0.50). (Fig. S7C, Table 3) The two leading drugs predicted to be metabolized by the microbes are metaxalone (mean MPP 0.64) and fenofibrate (mean MPP 0.54). (Fig. 3B) 44% (36/81) of the drugs contain mean MPPs above the metabolic prediction threshold of 0.45. The auxiliary microbe-drug pairs were selected based on failure to meet the *microbe proportion threshold* and *percent consumed threshold*, and therefore it is reasonable that we see more MPPs trending towards non-metabolic responses. Similar to the testing and training predictions, the auxiliary microbe-drug pair responses are more distinguishable based on the drug rather than the microbes. Experimentally these auxiliary microbe-drug pairs had very little metabolism representation across the microbes. The MPPs are less certain compared to those produced by the testing and training data, but are still above our classification threshold for metabolic predictions.

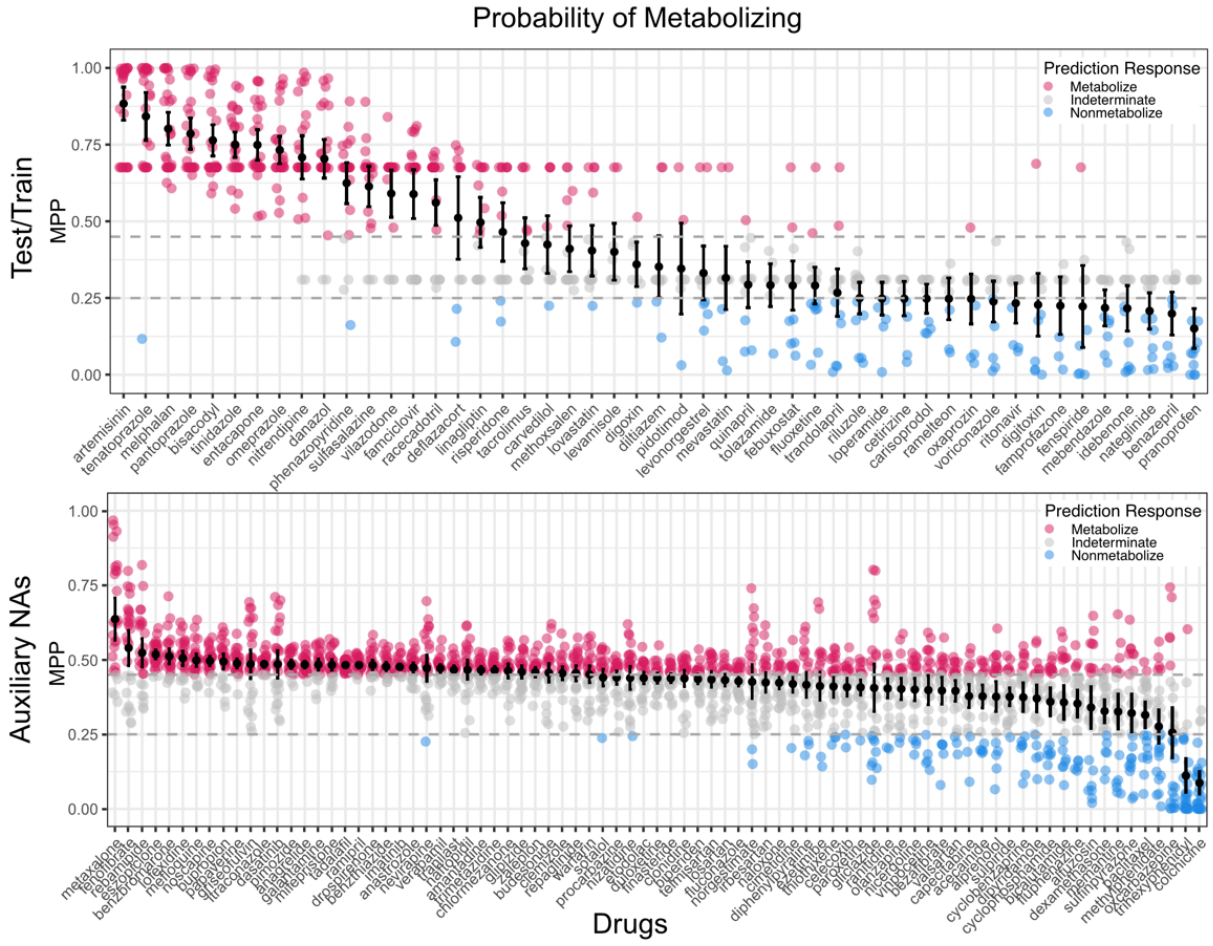


Fig. 3: Multi-GutV2 prediction responses grouped by drugs from the training and testing data (top) and auxiliary NA data (bottom). Optimized prediction thresholds determined from the sensitivity analysis are denoted as horizontal dashed lines.

3.7 Literature Exploration of Multi-GutV2 Responses

We asses Multi-GutV2 responses with literature-established gut microbiome metabolic relationships to explore our results. These include microbe-drug pairs from across the testing, training and auxiliary data. (Table 3) We explore known relationships between sulfasalazine, omeprazole, digoxin and levamisole metabolism evidence. The sulfasalazine literature-established relationship with the gut microbiome is reflected in the MPPs. Microbe-sulfasalazine pairs predicted as metabolic returned a relatively high median MPP of 0.68. For sulfasalazine, 78% (21/27) of microbes are predicted as metabolic and 22% (6/27) are indeterminate. Interestingly, for these six the true assigned class is denoted as non-metabolic, but since the MPPs are not low enough meet the non-metabolic prediction threshold of 0.25, they fall within the indeterminate class. This suggests the associated MDH signatures for these six sulfasalazine-microbe pairs still contain enough "metabolic" signature so that the model is unable to confidently predict them to be non-metabolic. Our results reflect *P. distasonis* and *B. vulgatus* contain high MPPs (0.89, 0.68 respectively). Interestingly, *ex vivo* human fecal cultures treated with sulfasalazine displayed a significantly increased species abundance of these same two bacterial species.[66] Additionally, of the Bacteroides genera, *B. caccae* and *B. Fragillis* (MPP 0.77, MPP 0.75) are most likely metabolic towards sulfasalazine. Within literature, *B. caccae* has been identified as a biomarker for children with ulcerative colitis.[67] At the same time, *B. caccae* is also reported as a predictor for successful

treatment outcomes with anti-cytokine therapies.[20] There are no findings associating successful treatment outcomes with sulfasalazine in children with ulcerative colitis and *B. caccae* abundance. However, the Bacteroides genera has been identified to contain azoreductases implicated in activating sulfasalazine through reduction.[68] Indeed our predictions reflect this trend, where 83% (10/12) of the Bacteroides species are predicted to be metabolic of sulfasalazine with a median MPP of 0.68. There are two Bacteroides species that are denoted as undetermined, *B. stercoris* and *B. finegoldii*, however the *in vitro* assigned truth is non-metabolic. The lack of consistency between the truth and response may be due to metabolic capacity being captured in the MDH signature. Reductive metabolism of omeprazole has been reported in *in vitro* conditions by gut bacteria to produce sulfide metabolites however there are few species explicitly identified.[26,69] In our model, 96% (26/27) of microbe-omeprazole pairs were assigned a truth of metabolic and 100% (27/27) had a predicted response of metabolic. Digoxin has an interesting literature established relationship with *E. lenta*, where only strains that contain a two-gene cytochrome-encoding operon metabolize the drug. Our results (MPP 0.31, indeterminate) reflect this range of *E. Lenta* response, since it is not a metabolic relationship held by the entire species and thusly the MPP is low enough to be placed in the indeterminate response. Due to the MDH signature capturing some of this metabolic capacity the MPP is not below the non-metabolic threshold. Microbe-levamisole pairs had responses that deviated from the "truth" assigned. There are two microbes *A. colihominis* and *E. coli* out of the 27 levamisole-microbe pairs assigned a truth of metabolic. Interestingly, the model prediction responses identified seven microbes to be metabolic (median MPP 0.53). Evidence suggests Levamisole may be metabolized by Bacteroides species, and indeed three of these levamisole-microbe pairs included *B. uniformis* (MPP 0.72). *B. cellulosilyticus* (MPP 0.54), and *B. thetaiotaomicron* (MPP 0.50).[70] Our results with sulfasalazine, omeprazole, digoxin and levamisole are explored due to their existing literature evidence to have metabolic properties with the gut microbiome. These results highlight the model's capability to return predictions that align with literature-based findings. Drug-metabolizing capacity can vary among strains therefore taxonomic trends do not necessarily indicate conserved metabolic capacity across all members. Discrepancies between the response and truth may indicate the MDH signature's ability to discern between the *in vivo* metabolic capacity *independent* of the assigned "true" class based on the *in vitro* data.

Within the auxiliary prediction data, we explore the prediction responses for microbe-drug pairs with an unknown true class. (Fig. 3, Fig. S7C, Table 3) Some of the 18 drugs identified to be metabolized include capecitabine, sulfinpyrazone, ranitidine and nizatidine. Our results reflect varying degrees of model responses that align with literature-established evidence. Capecitabine is reduced by dihydropyrimidine dehydrogenase, an enzyme previously identified in 85 species of bacteria through comparative genomics.[44] In our results, capecitabine returned 22% (6/27) microbes as having a metabolic relationship (median MPP 0.49). Sulfinpyrazone can experience s-oxide reduction by microbes found in human stool.[71,72] For sulfinpyrazone, 30% (8/27) microbes were predicted to have metabolic relationship (median MPP 0.52). Ranitidine and nizatidine, when incubated with human stool experience a nitro reduction via cleavage of an N-oxide bond.[73,74] Ranitidine-microbe pairs had 33% (9/27) predicted to be metabolic (median MPP 0.47). Nizatidine-microbe pairs had 52% (14/27) microbes predicted to metabolic (median MPP 0.51). Despite these microbe-drug pairs being sequestered to the auxiliary data during preprocessing class assignment due to their inability to meet the *percent consumed* and *microbe proportion* threshold, we still see predictions for microbes with metabolic capacity. The median MPPs are lower compared to those from the testing/training. This is somewhat reasonable, since the auxiliary data

was not used in training, and therefore the MDH signature is not tuned to detect any metabolic capacity within these drugs. Despite this, the metabolic predictions demonstrate the generalizable, MDH signature *can* distinguish metabolic capacities on drugs that did not contribute to the training of the model.

Metabolic relationships identified from *in vitro* screening with drugs can contribute to our understanding of drug-microbe interactions. However, it does not capture dynamic microbiome composition which often impacts bacterial metabolic need and activity. While we can state what the assigned "truth" is based off the *in vitro* data, the endogenous behavior of the bacteria species in the host microbiome context may be different. Due to this misalignment, it is possible that behavior observed *in vitro* is not behavior observed in the endogenous host microbiome. These results reflect the MDH signature derived from *in vivo* patient data can capture metabolic relationships, which are difficult to replicable *in vitro*. The MDH signature improves on the existing *in vitro* and *in silico* methods as it integrates the *context* of the complete endogenous microbiome interaction with the host. The Multi-GutV2 model has identified species-specific microbes with MDH signatures consistent with metabolic activity.

4. Discussion

The MDH signature can predict species-specific metabolic relationships. Our model is distinct from existing methods since it is derived from human omics data and microbiome composition from disease-free subjects and integrates the drug-treatment variable through *in vitro* derived data. Our findings highlight the novelty and utility of using a MDH signature to identify drug-microbial metabolic relationships in the vaginal and gut microbiome. Using this MDH signature we build vaginal microbiome models, TV-VagGene and TV-VagProt, that predict metabolism of TV. These models contain high performance metrics on test data (precision, accuracy, recall, specificity equal to and above 0.84) and an AUC equal to and above 0.99. The two models agree on the novel predictions of TV metabolism by the vaginal bacteria: *A. vaginae*, *F. nucleatum*, Peptostreptococcus genus, and Bifidobacterium genus.

We further demonstrate the greater capacity of the MDH signature to capture multiple microbe-drug metabolic relationships in the gut microbiome with the Multi-GutV1 and Multi-GutV2 models. The Multi-GutV2 model was identified through sensitivity training of the V1 model and contains useable performance metrics from the testing data (precision, accuracy, specificity above or equal to 0.81 and recall at 0.75), with an AUC of 0.87. For the Multi-GutV2 model the leading microbes most implicated in metabolism of the assessed drugs are *B. cellulosilyticus*, *B thetaiotaomicron*, and *B. hansenii*. There are 54 total drugs with mean MPPs across the respective microbe-drug pairs assed that reflect metabolic potential with multiple microbe species. The responses also reflect trends in drug literature-established metabolic relationships. The Multi-GutV2 model produced MPPs in the testing data which are highly correlated (ρ of 0.55, p-value 1.3e-53) to the percent consumed values from the *in vitro* data. This reflects the MDH signature's ability to capture and discern the range of metabolic capacity of each unique microbe-drug pair.

The method to produce the MDH signatures is highly customizable as it requires 1) non-disease paired omics and 2) a few microbe-drugs with defined metabolic relationships. It is advantageous the human omics and microbiome composition is *non-disease* as this expands the range of applicable available datasets. The human paired omics and microbiome composition are tissue-specific and disease-agnostic. The training data requires defined metabolic relationships for microbe-drug pairs to establish the model. These can be taken from literature (as shown in TV-

VagGene and TV-VagProt models) or custom *in vitro* experiments with a threshold to delineate defined metabolic relationships (as done in the Multi-GutV1 and V2 models).

Future applications of this model could adapt the existing pipeline and shift the specificity of the model to suit the research interests. The utility of this model is constrained to the microbiome diversity of patient samples. This can impact the performance of the model though limited samples for testing and training or limit the amount of novel microbe-drug pairs that are predicted on. Future applications of this method could generate models unique to specific drug classes or hypothesized enzymatic mechanisms. Drugs in similar classes or with similar functional groups have been shown to share metabolic properties consistent across different microbes with the same enzymes.[19,32,33,65,75] Future applications can incorporate *in vitro* or *ex vivo* validation of the microbe-drug prediction, which is a limitation in our interpretation. Another model constraint is the feature selection T-test, which may omit interaction effects leveraged by the SVM. Future implementations could consider multivariate approaches with built-in feature selection such as Elastic Net, LASSO, or PCA.

Another improvement could be applying tissue-specific *in vitro* drug-treated cell line data rather than the LINCS CD values. The LINCS CD readings include a vast diversity of cell types across various tissue locations. The models may have better performance if the *in vitro* values were more tissue-specific and therefore would have more contextually accurate drug-treatment values to contribute to the MDH signatures. Future applications of this model could also expand using a larger applied drug pool. Here, we limited our microbe-drug pairs to drugs that were found in both our *in vitro* drug-treated microbes from the Zimmerman et al 2019 publication to simplify interpretation of our results and to ensure we had *in vitro* values to compare our generated MPP per microbe-drug pair. Future applications could use the whole LINCS catalogue to generate microbe-drug pair predictions for all LINCS drugs to generate a reference database. This generated library of predictions could be cited to preemptively identify any potential impact of the gut microbiome on the drug of interest. We could also make individual SVMs per drug for each microbiome, similar to the individual models we made for TV metabolism. These single-drug models may contain higher precision since the class-separating features in the MDH signatures capture each specific drug's metabolic properties. The weakness of a multidrug model such as the Multi-GutV1 and V2 is the class-distinguishing features are *not* specific to the drug class nor the structures, which can contribute to loss of performance. This is reasonable, since introducing additional drugs may obscure the features that best distinguish the classification abilities. However, despite this, in a multidrug model the features are more generalizable *because* the features are not tied to one drug. We have shown reliable performance is still viable in a multidrug model.

The utility of these models is most evident within the clinical setting. We can preemptively determine how a patient's microbiome may impact the therapeutic effect of the drug. With this knowledge we can either pivot drug treatment to select a therapeutic that will not interact with the microbiome or provide probiotics to shift the microbiome populations to be more receptive to the drug of choice. This is a low-cost, high throughput method to assess the microbe metabolism of drugs in an accessible manner. Clinicians can reference this tool to incorporate microbiome state into patient treatment plans going forward, thereby improving personalized medicine and treatment success.

5. Supplementary Information

All supplementary figures, tables, code and high-resolution main figures can be found here: https://github.com/Brubaker-Lab/microbe-host-omic_nemecio

6. References

1. Jiménez-Escrig, A. & Sánchez-Muniz, F. J. Dietary fibre from edible seaweeds: Chemical structure, physicochemical properties and effects on cholesterol metabolism. *Nutr. Res.* **20**, 585–598 (2000).

2. Holscher, H. D. Dietary fiber and prebiotics and the gastrointestinal microbiota. *Gut Microbes* **8**, 172–184 (2017).

3. Williams, B. A., Mikkelsen, D., Flanagan, B. M. & Gidley, M. J. 'Dietary fibre': moving beyond the 'soluble/insoluble' classification for monogastric nutrition, with an emphasis on humans and pigs. *J. Anim. Sci. Biotechnol.* **10**, 45 (2019).

4. Kumar, J., Rani, K. & Datt, C. Molecular link between dietary fibre, gut microbiota and health. *Mol. Biol. Rep.* **47**, 6229–6237 (2020).

5. Biesalski, H. K. Nutrition meets the microbiome: micronutrients and the microbiota. *Ann. N. Y. Acad. Sci.* **1372**, 53–64 (2016).

6. Barone, M., D'Amico, F., Brigidi, P. & Turroni, S. Gut microbiome–micronutrient interaction: The key to controlling the bioavailability of minerals and vitamins? *BioFactors* **48**, 307–314 (2022).

7. Rowland, I. *et al.* Gut microbiota functions: metabolism of nutrients and other food components. *Eur. J. Nutr.* **57**, 1–24 (2018).

8. Soto-Martin, E. C. *et al.* Vitamin Biosynthesis by Human Gut Butyrate-Producing Bacteria and Cross-Feeding in Synthetic Microbial Communities. *mBio* **11**, e00886-20 (2020).

9. Hofmann, A. F. The continuing importance of bile acids in liver and intestinal disease. *Arch. Intern. Med.* **159**, 2647–2658 (1999).

10. Ridlon, J. M., Kang, D. J., Hylemon, P. B. & Bajaj, J. S. Bile acids and the gut microbiome. *Curr. Opin. Gastroenterol.* **30**, 332–338 (2014).

11. Xu, Z. & Knight, R. Dietary effects on human gut microbiome diversity. *Br. J. Nutr.* **113**, S1–S5 (2015).

12. Shreiner, A. B., Kao, J. Y. & Young, V. B. The gut microbiome in health and in disease. *Curr. Opin. Gastroenterol.* **31**, 69–75 (2015).

13. Jones, M. L., Martoni, C. J., Ganopolsky, J. G., Labbé, A. & Prakash, S. The human microbiome and bile acid metabolism: dysbiosis, dysmetabolism, disease and intervention. *Expert Opin. Biol. Ther.* **14**, 467–482 (2014).

14. Karlsson, F. H., Ussery, D. W., Nielsen, J. & Nookaew, I. A closer look at bacteroides: phylogenetic relationship and genomic implications of a life in the human gut. *Microb. Ecol.* **61**, 473–485 (2011).

15. Witkin, S. & Linhares, I. Why do lactobacilli dominate the human vaginal microbiota? *BJOG Int. J. Obstet. Gynaecol.* **124**, 606–611 (2017).

16. Witkin, S. S. *et al.* Influence of vaginal bacteria and D- and L-lactic acid isomers on vaginal extracellular matrix metalloproteinase inducer: implications for protection against upper genital tract infections. *mBio* **4**, e00460-13 (2013).

17. V. Sgibnev, A. & A. Kremleva, E. Vaginal Protection by H2O2-Producing Lactobacilli. *Jundishapur J. Microbiol.* **8**, e22913 (2015).

18. Wilson, I. D. & Nicholson, J. K. Gut Microbiome Interactions with Drug Metabolism, Efficacy and Toxicity. *Transl. Res. J. Lab. Clin. Med.* **179**, 204 (2016).

19. Enright, E. F., Gahan, C. G. M., Joyce, S. A. & Griffin, B. T. The Impact of the Gut Microbiota on Drug Metabolism and Clinical Outcome. *Yale J. Biol. Med.* **89**, 375–382 (2016).

20. Lee, J. W. J. *et al.* Multi-omics reveal microbial determinants impacting responses to biologic therapies in inflammatory bowel disease. *Cell Host Microbe* **29**, 1294-1304.e4 (2021).

21. Ananthakrishnan, A. N. *et al.* Gut Microbiome Function Predicts Response to Anti-integrin Biologic Therapy in Inflammatory Bowel Diseases. *Cell Host Microbe* **21**, 603-610.e3 (2017).

22. Kostrzewska, M. *et al.* The effect of omeprazole treatment on the gut microflora and neutrophil function. *Clin. Res. Hepatol. Gastroenterol.* **41**, 575–584 (2017).

23. Koo, S. H. *et al.* Effects of proton pump inhibitor on the human gut microbiome profile in multi-ethnic groups in Singapore. *Singapore Med. J.* **60**, 512–521 (2019).

24. Sousa, T. *et al.* On the Colonic Bacterial Metabolism of Azo-Bonded Prodrugsof 5-Aminosalicylic Acid. *J. Pharm. Sci.* **103**, 3171–3175 (2014).

25. Seradj, D.-S. *et al.* Influence of Different Diets on the Degradation of Sulfasalazine by Colon Bacteria Determined Using MimiCol3. *Pharmaceuticals* **16**, 1128 (2023).

26. Guo, J. *et al.* Gut microbiota and host Cyp450s co-contribute to pharmacokinetic variability in mice with non-alcoholic steatohepatitis: Effects vary from drug to drug. *J. Adv. Res.* **39**, 319–332 (2022).

27. Saha, J. R., Butler, V. P., Neu, H. C. & Lindenbaum, J. Digoxin-inactivating bacteria: identification in human gut flora. *Science* **220**, 325–327 (1983).

28. Haiser, H. J. *et al.* Predicting and Manipulating Cardiac Drug Inactivation by the Human Gut Bacterium Eggerthella lenta. *Science* **341**, 295–298 (2013).

29. Lindenbaum, J., Rund, D. G., Butler, V. P., Tse-Eng, D. & Saha, J. R. Inactivation of Digoxin by the Gut Flora: Reversal by Antibiotic Therapy. *N. Engl. J. Med.* **305**, 789–794 (1981).

30. Klatt, N. R. *et al.* Vaginal bacteria modify HIV tenofovir microbicide efficacy in African women. *Science* **356**, 938–945 (2017).

31. Cheu, R. K. *et al.* Impact of vaginal microbiome communities on HIV antiretroviral-based pre-exposure prophylaxis (PrEP) drug metabolism. *PLOS Pathog.* **16**, e1009024 (2020).

32. Sun, C., Chen, L. & Shen, Z. Mechanisms of gastrointestinal microflora on drug metabolism in clinical practice. *Saudi Pharm. J.* **27**, 1146–1156 (2019).

33. Martinelli, F. & Thiele, I. Microbial metabolism marvels: a comprehensive review of microbial drug transformation capabilities. *Gut Microbes* **16**, 2387400.

34. Wu, Q. *et al.* Activity of GPCR-targeted drugs influenced by human gut microbiota metabolism. *Nat. Chem.* **17**, 808–821 (2025).

35. Verdegaal, A. A. *et al.* A drug–microbiome–drug interaction impacts co-prescribed medications for Parkinson's disease. *Nat. Microbiol.* **11**, 1387–1409 (2026).

36. Lettieri, M., Rodda, M. & Carlucci, V. Machine Learning in Early Prediction of Metabolism of Drugs. in *Computational Toxicology: Methods and Protocols* (ed. Nicolotti, O.) 275–291 (Springer US, New York, NY, 2025). doi:10.1007/978-1-0716-4003-6_13.

37. Kazmi, S. R., Jun, R., Yu, M.-S., Jung, C. & Na, D. *In silico* approaches and tools for the prediction of drug metabolism and fate: A review. *Comput. Biol. Med.* **106**, 54–64 (2019).

38. Dudas, B. & Miteva, M. A. Computational and artificial intelligence-based approaches for drug metabolism and transport prediction. *Trends Pharmacol. Sci.* **45**, 39–55 (2024).

39. Litsa, E. E., Das, P. & Kavraki, L. E. Machine learning models in the prediction of drug metabolism: challenges and future perspectives. *Expert Opin. Drug Metab. Toxicol.* **17**, 1245–1247 (2021).

40. Kazmi, S. R., Jun, R., Yu, M.-S., Jung, C. & Na, D. In silico approaches and tools for the prediction of drug metabolism and fate: A review. *Comput. Biol. Med.* **106**, 54–64 (2019).

41. Guthrie, L., Wolfson, S. & Kelly, L. The human gut chemical landscape predicts microbe-mediated biotransformation of foods and drugs. *eLife* **8**, e42866 (2019).

42. Javdan, B. *et al.* Personalized Mapping of Drug Metabolism by the Human Gut Microbiome. *Cell* **181**, 1661-1679.e22 (2020).

43. Wensel, C. R., Pluznick, J. L., Salzberg, S. L. & Sears, C. L. Next-generation sequencing: insights to advance clinical investigations of the microbiome. *J. Clin. Invest.* **132**, e154944.

44. Heinken, A. *et al.* Genome-scale metabolic reconstruction of 7,302 human microorganisms for personalized medicine. *Nat. Biotechnol.* **41**, 1320–1331 (2023).

45. Malla, M. A. *et al.* Exploring the Human Microbiome: The Potential Future Role of Next-Generation Sequencing in Disease Diagnosis and Treatment. *Front. Immunol.* **9**, (2019).

46. Bustion, A. E., Nayak, R. R., Agrawal, A., Turnbaugh, P. J. & Pollard, K. S. SIMMER employs similarity algorithms to accurately identify human gut microbiome species and enzymes capable of known chemical transformations. *eLife* **12**, e82401 (2023).

47. Bhalla, P., Rengaswamy, R., Karunagaran, D., Suraishkumar, G. K. & Sahoo, S. Metabolic modeling of host–microbe interactions for therapeutics in colorectal cancer. *Npj Syst. Biol. Appl.* **8**, 1 (2022).

48. Stevanoska, M., Folz, J., Beekmann, K. & Aichinger, G. Physiologically based kinetic (PBK) modeling as a new approach methodology (NAM) for predicting systemic levels of gut microbial metabolites. *Toxicol. Lett.* **396**, 94–102 (2024).

49. Hu, T. *et al.* Genome-scale metabolic modeling reveals specific vaginal Lactobacillus strains and their metabolites as key inhibitors of Candida albicans. *Microbiol. Spectr.* **13**, e02984-24 (2025).

50. Glass, E. M., Kolling, G. L. & Papin, J. A. Genome-scale metabolic modelling identifies vaginal microbiome members as potential probiotics. *Nat. Microbiol.* **11**, 2034–2046 (2026).

51. Lee, C. Y., Dillard, L. R., Papin, J. A. & Arnold, K. B. New perspectives into the vaginal microbiome with systems biology. *Trends Microbiol.* **31**, 356–368 (2023).

52. Zimmermann, M., Zimmermann-Kogadeeva, M., Wegmann, R. & Goodman, A. L. Separating host and microbiome contributions to drug pharmacokinetics and toxicity. *Science* **363**, eaat9931 (2019).

53. McCoubrey, L. E. *et al.* Machine Learning Predicts Drug Metabolism and Bioaccumulation by Intestinal Microbiota. *Pharmaceutics* **13**, 2001 (2021).

54. Malwe, A. S., Srivastava, G. N. & Sharma, V. K. GutBug: A Tool for Prediction of Human Gut Bacteria Mediated Biotransformation of Biotic and Xenobiotic Molecules Using Machine Learning. *J. Mol. Biol.* **435**, 168056 (2023).

55. Magnúsdóttir, S. *et al.* Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat. Biotechnol.* **35**, 81–89 (2017).

56. Chen, X., Alonso, A. P., Allen, D. K., Reed, J. L. & Shachar-Hill, Y. Synergy between (13)C-metabolic flux analysis and flux balance analysis for understanding metabolic adaptation to anaerobiosis in E. coli. *Metab. Eng.* **13**, 38–48 (2011).

57. Vijayakumar, S., Rahman, P. K. S. M. & Angione, C. A Hybrid Flux Balance Analysis and Machine Learning Pipeline Elucidates Metabolic Adaptation in Cyanobacteria. *iScience* **23**, 101818 (2020).

58. Burgard, A. P., Pharkya, P. & Maranas, C. D. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* **84**, 647–657 (2003).

59. Oskotsky, T. T. *et al.* A transcriptomics-based computational drug repurposing pipeline identifies simvastatin and primaquine as therapeutics for endometriosis. *iScience* **29**, (2026).

60. Heinken, A. & Thiele, I. Systematic prediction of health-relevant human-microbial co-metabolism through a computational framework. *Gut Microbes* **6**, 120–130 (2015).

61. Bauer, E., Zimmermann, J., Baldini, F., Thiele, I. & Kaleta, C. BacArena: Individual-based metabolic modeling of heterogeneous microbes in complex communities. *PLOS Comput. Biol.* **13**, e1005544 (2017).

62. Heffron, R. *et al.* Efficacy of oral pre-exposure prophylaxis (PrEP) for HIV among women with abnormal vaginal microbiota: a post-hoc analysis of the randomised, placebo-controlled Partners PrEP Study. *Lancet HIV* **4**, e449–e456 (2017).

63. Baeten, J. M. *et al.* Antiretroviral prophylaxis for HIV prevention in heterosexual men and women. *N. Engl. J. Med.* **367**, 399–410 (2012).

64. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).

65. Zimmermann, M., Zimmermann-Kogadeeva, M., Wegmann, R. & Goodman, A. L. Mapping human microbiome drug metabolism by gut bacteria and their genes. *Nature* **570**, 462 (2019).

66. Li, L. *et al.* RapidAIM: a culture- and metaproteomics-based Rapid Assay of Individual Microbiome responses to drugs. *Microbiome* **8**, 33 (2020).

67. Mouzan, M. E., Sarkhy, A. A. & Assiri, A. Gut microbiota predicts the diagnosis of ulcerative colitis in Saudi children. *World J. Clin. Pediatr.* **13**, (2024).

68. Braccia, D. J. *et al.* Gut Microbiome–Wide Search for Bacterial Azoreductases Reveals Potentially Uncharacterized Azoreductases Encoded in the Human Gut Microbiome. *Drug Metab. Dispos.* **51**, 142–153 (2023).

69. Watanabe, K., Yamashita, S., Furuno, K., Kawasaki, H. & Gomita, Y. Metabolism of Omeprazole by Gut Flora in Rats. *J. Pharm. Sci.* **84**, 516–517 (1995).

70. Shu, Y.-Z., Kingston, D. G. I., Van Tassell, R. L. & Wilkins, T. D. Metabolism of levamisole, an anti-colon cancer drug, by human intestinal bacteria. *Xenobiotica* **21**, 737–750 (1991).

71. Renwick, A. G., Evans, S. P., Sweatman, T. W., Cumberland, J. & George, C. F. The role of the gut flora in the reduction of sulphinpyrazone in the rat. *Biochem. Pharmacol.* **31**, 2649–2656 (1982).

72. Strong, H. A., Renwick, A. G., George, C. F., Liu, Y. F. & Hill, M. J. The reduction of sulphinpyrazone and sulindac by intestinal bacteria. *Xenobiotica* **17**, 685–696 (1987).

73. Basit, A. W., Newton, J. M. & Lacey, L. F. Susceptibility of the H2-receptor antagonists cimetidine, famotidine and nizatidine, to metabolism by the gastrointestinal microflora. *Int. J. Pharm.* **237**, 23–33 (2002).

74. Basit, A. W. & Lacey, L. F. Colonic metabolism of ranitidine: implications for its delivery and absorption. *Int. J. Pharm.* **227**, 157–165 (2001).

75. Jourova, L., Anzenbacher, P. & Anzenbacherova, E. Human gut microbiota plays a role in the metabolism of drugs. *Biomed. Pap.* **160**, 317–326 (2016).

Combination Overlap in Drug Synergy Benchmarks: An Audit of Cell-Line Holdout Protocols

Kusal Debnath and Preetam Ghosh†

Department of Computer Science, Virginia Commonwealth University, Richmond, Virginia 23284, USA

†*E-mail: pghosh@vcu.edu*

Drug-combination synergy models are generally benchmarked on combinatorial screens, in which a fixed panel of drug pairs is assayed across a fixed panel of cell lines. Random splits over measurements are permissive in nature, as the same drug pair may show up in both the train and test set, albeit alongside different cell lines, allowing the model to interpolate rather than generalize. As a possible mitigation, the field has adopted cold-start protocols, namely, leave-drug-out, leave-cell-line-out, leave-tissue-out, and leave-combination-out. Nevertheless, these protocols differ in what they actually withhold. In this work, we show that holding out a cell line does not hold out a combination. Whenever a screen assays its pair matrix across its entire cell panel, leave-cell-line-out retains every test combination in training by construction. This holds exactly on the fully-crossed ALMANAC, FRIEDMAN, and O'Neil screens, and reaches 92% on the sparser AstraZeneca screen. Additionally, a harmonised DrugComb corpus still retains 72.8%, suggesting that pooling across screens dilutes the effect instead of removing it. To further quantify the consequence, we compare a null model using no molecular information, comprised of a per-cell-line mean along with a per-pair mean, against a symmetric neural baseline and a larger pathway-aware model under matched splits. Under the leave-cell-line-out setup, the null model outperforms both, and collapses when combinations are held out instead. Between leave-combination-out and leave-cell-line-out, the sign of structure's contribution flips, while leave-drug-out yields no resolvable difference and leave-tissue-out is too variable across folds to adjudicate. Moreover, for this one pathway model, ablation locates the advantage in the drug-target branch rather than in pathway topology, although the branch's contribution cannot be separated from the capacity and auxiliary supervision it also supplies. Our work thus emphasizes that predicting a known combination in an uncharacterized context is legitimate and clinically relevant, but distinct from combination generalization; retained overlap is a property of a protocol to be reported alongside the claim it supports, rather than a defect to be corrected.

Keywords: Drug combination synergy; evaluation methodology; cold-start evaluation; train–test overlap; benchmark design; combinatorial drug screens

1. Introduction

In cancer treatment, combination therapy has emerged as the standard protocol. Nevertheless, exhaustive screening of compounds is tedious due to the large candidate space. In silico adoption for screening, pioneered by Preuer et al.,[1] aids in this scenario.

© 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.

Evolution of research in this field has been majorly measured against benchmarks derived from combinatorial screens, namely, O'Neil,[2] NCI-ALMANAC,[3] and their aggregation in DrugComb.[4,5] These screens share a design property that is a virtue experimentally but a hazard statistically: a fixed panel of drug pairs is assayed across a fixed panel of cell lines. As a result, random shuffling and splitting causes the same cell lines to appear in both the train and test sets, just with different pairs. During testing, the model encounters drug pairs and cell lines it has already seen, but with a new intersection. Therefore, the credibility of such test results is a matter of debate due to the inherent bias in the predictions.

Evaluating on novel combinations instead of random measurements, Preuer et al. reported that every method they compared degraded when extrapolating to unexplored drugs or cell lines. Later, Dewulf et al. formalized the cold-start subtasks for drug-drug interaction prediction and argued that matched validation schemes are critical to correct assessment.[6] Recent methods routinely reported leave-drug-out, leave-combination-out and leave-cell-line-out alongside headline numbers.[7–9]

As an instance of a general problem, Kapoor et al. surveyed seventeen scientific fields and observed leakage affecting 294 papers, often producing extremely overoptimistic conclusions. In their reproducibility study of civil-war prediction, correcting leakage eliminated the entire advantage of complex models over vanilla logistic regression.[10,11] Bernett et al. emphasized that leakage is especially hard to detect in biological data due to the complex dependencies between samples,[12] while Rosenblatt et al. showed that repeated subjects drastically inflate performance in connectome models.[13]

In this work, we seek an answer to a question that, to our knowledge, has not been addressed: within each cold-start protocol, how much combination overlap actually survives? Holding out a cell line removes a context, not the combination, as the same pair is measured in the cell lines that remain. The importance of such a scenario is an empirical question related to the redundancy of the screens.

We treat these protocols as definitions of distinct tasks rather than as defective instruments. Leave-cell-line-out defines a well-posed and clinically meaningful task, predicting a profiled combination in an uncharacterized context, and retaining that combination in training is precisely what the task licenses. What we audit is the distance between that definition and the generalization claims commonly attached to the numbers it produces.

In summary, this work is best placed as a method to characterize what each evaluation protocol withholds and what follows for the interpretation of reported performance. We quantify how much combination overlap each cold-start protocol actually leaves, protocol by protocol and screen by screen. Specifically, our work thus contributes the following:

- An audit across six protocols to measure the fraction of test combinations also present in training,
- A null model quantifying the consequence, using a per-cell plus per-pair mean,
- A demonstration that the protocol determines the verdict on held-out combinations, and,
- A component attribution to biological priors that might partially reflect database completion.

2. Background and Related Work

2.1. *Synergy Quantification*

Synergy is defined relative to a reference model of non-interaction. Standard references are Loewe additivity, Bliss independence, Highest Single Agent(HSA), and Zero Interaction Potency(ZIP). These references depict different assumptions and disagree in practice,[14–16] to the point that the choice of dose-response fitting software can flip a combination between synergistic and antagonistic.[17] In this work, we use Loewe as the reference model for quantifying synergy.

2.2. *Cold-Start Protocols*

As one of the earliest approaches, Dewulf et al. identified the distinct subtasks surfaced when drugs, pairs or contexts are novel, and proposed matched validation schemes.[6] In their review, Abbasi et al. concluded with a classic cold-start issue by stating that the methods they implemented performed well for known drugs and cell lines but fall short for novel ones.[18]Yousefabadi et al. showcased that permissive cross-validation allowing drug pairs across folds inflates performance, and thus adopted pair-held-out validation for antibacterial synergy.[19] In a more recent work, Agboola et al. reported a random forest outperforming their deep model under cold-drug validation.[20]

2.3. *Structure-informed Synergy Models*

Incorporating pathway or network structure is an active direction, spanning network-proximity approaches,[21] pathway-activation scores,[22] interpretable subsystem models,[23] graph and transformer architectures,[9,24,25] multi-task formulations,[26,27] multi-omics ensembles[28,29] and proteomics-based models.[8]

2.4. *Leakage in ML-based Science*

Non-independence between train and test samples is one of eight leakage types discussed by Kapoor et al. in their taxonomy,[10,11] addressed by reporting checklists[12,30] and quantified in neuroimaging[13] and other applied fields.[31,32] Cold-start protocols have been already adopted in the subfield as remedy, however, the shortcomings of such a remedy can only be seen by scrutinizing pairs.

Combination overlap is a measurable property of a partition: the fraction of test combinations that also occur in training, computable without fitting a model. Task mismatch is a gap between the entity a protocol withholds and the entity a claim generalizes over, as when leave-cell-line-out, which withholds biological contexts, is used to support a claim about unseen combinations. Leakage is the case in which that gap invalidates the performance estimate for the claim being made. Overlap alone is therefore neither necessary nor sufficient for leakage: a protocol that retains most of its test combinations is performing exactly as intended when the claim concerns a known combination in a new context, and the same retention becomes leakage only under a claim of combination generalization.

In a recent work, Çandır et al.[33] replaced the drug and cell-line representations of five published synergy models with one-hot identifiers and obtained comparable or slightly better performance, concluding that these models exploit identity-linked covariation in the synergy labels rather than the chemical and molecular features they are supplied.

3. Methods

3.1. *Corpus*

We harmonized DrugComb[4] to 634,631 measurements after removing self-pairs and rows lacking a Loewe score, resulting in 73,108 unique unordered pairs across 288 cell lines. Eliminating instances of incomplete drug and cell features drastically shrinks the measurements to 180,746 over 1,945 drugs, 159 cell lines and 15 tissues, drawn from ALMANAC, FRIEDMAN, O'NEIL and ASTRAZENECA. Every protocol, split, overlap figure and performance metrics reported in this study is computed on this feature-complete subset of 180,746 measurements.

3.2. *Protocols*

Here, six protocols are implemented over the common corpus: random, leave-combo-out (unseen combination, both drugs seen), leave-drug-out (one/both variants), leave-cell-out, leave-tissue-out, leave-study-out. Each of the protocols persisted as row indices with a corpus fingerprint so a split replays exactly without redistributing data.

The drug degree distribution is heavy-tailed: a small number of drugs participate in a large share of the pairs and therefore carry a disproportionate share of the rows. Partitioning entities uniformly is consequently unstable, as a single high-degree entity landing on the held-out side moves the realized test fraction far from its nominal value. Instead of uniform partitioning, we therefore assigned entities by mass-aware greedy packing. The packer orders groups by row mass, assigns each to the bin with the largest remaining deficit, and asserts the achieved test fraction before a split is accepted. Every split reported in this paper is produced by this packer on the 180,746-row feature-complete corpus.

The assertion applies to the mass-packed protocols only. The leave-tissue-out and leave-study-out partition on an explicit categorical assignment. As a result, their test fraction is essentially what the held-out block weighs, hence the check is waived consciously. The leave-tissue-out is restricted to the eight tissues with at least 3,500 covered rows, making the distribution severely skewed (skin 71,638 rows, 39.6% of the 180,746-row feature-complete corpus; pancreas 6 rows).

3.3. *Baselines*

We denote the null model as B1. For a test row carrying unordered pair p in cell line c, B1 predicts $y(p, c) = m_c + d_p$, where m_c is the mean Loewe score over the training rows of the cell line c and d_p is the mean of the residual $y_i - m_{c(i)}$ over the training rows i and pair p, both estimated on training rows only. Next, the unseen entities back off: m_c is replaced by the global training mean when c contributes no training row, and d_p is set to zero when p contributes none, so a pair unseen in a cell line unseen receives the global training mean as a constant prediction.

B1 therefore has no fitted weights, no regularization and no hyperparameters, and its score under a given protocol is a property of the split rather than of a training procedure. It uses no structure, targets, pathways or omics since the objective here is diagnosing a benchmark on which B1 is competitive. Signal from such a benchmark might be largely recoverable from pair and context identity alone, without chemistry.

Two additional baselines are a gradient boosting on PCA-reduced embeddings (B2) and a symmetric MLP over drug embeddings and cell features with no ontology, graph or target projection (B3). Both consume the drug embeddings and the cell context specified in Section 3.4, encoded symmetrically as the sum and the absolute difference of the two drug vectors concatenated with the cell vector, so that drug order cannot affect a prediction. B2 reduces the drug embeddings with a 64-component PCA fitted on the stacked training embeddings of both drugs and fits a histogram gradient-boosting regressor with 300 boosting iterations and otherwise default settings. B3 applies the same encoding to the unreduced embeddings and passes it through a 512-unit layer with layer normalization, GELU activation and dropout 0.1, a 256-unit GELU layer and a scalar head, trained for 30 epochs with AdamW at learning rate 0.001, weight decay 0.01 and batch size 512, minimizing mean squared error on winsorized and standardized Loewe scores with no early stopping. B2 is a learned but non-neural comparator that sits between B1 and the neural models, while B3 acts as the reference for "does structure help".

3.4. *Features and Model Configuration*

Cell context is a 160-dimension entity, comprising of 14 PROGENy activities,[34] 64 principal components of protein-coding expression, 64 of CRISPR gene effect over the ontology's gene universe, and 18 genotype flags, all from the DepMap portal.[35] Drug embeddings are generated using MolFormer-XL.[36] The drug-target profiles come from ChEMBL[37] over a 72-gene panel, where only 24.7% of trainable rows have both drugs annotated. Hence, a target-imputation head predicts profiles from structure.

Pathway vocabulary, gene membership, graph topology and mechanism taxonomy are a versioned specification with no assumption about relevant biology is embedded in the architecture. Here, we use MSigDB Hallmark (50 gene sets)[38] and aggregate a pathway-pathway graph from STRING v12 physical interactions at combined score at least 700[39] with no hand-specified edges. The panel genes are absent from every Hallmark set acting as the druggable kinome, while Hallmark sets are expression-response signatures and are placed by STRING guilt-by-association. Finally, the drug target profiles project onto pathways, compose across the pair over the graph, and read out through a synergy head.

3.5. *Training and Model Selection*

Every split is persisted as disjoint training, validation and test bins. The mass-packed protocols divide the row mass in proportions of 0.8, 0.1 and 0.1, while leave-tissue-out and leave-study-out validate on a second tissue and on FRIEDMAN, and the tissue cross-validation serves evaluation only. Because the protocol is the object of comparison, each model keeps one configuration, fixed in advance and shared across protocols, folds and seeds, and no hyper-

parameter search is performed. The pathway model is a two-layer, two-head GAT with node dimension 16, hidden width 256, encoder width 512, dropout 0.1 and max projection, auxiliary target, node and sparsity losses weighted 0.1, 0.05 and 0.01, and an imputation head frozen for the first ten epochs. It is trained with AdamW at learning rate 3×10^{-4}, weight decay 0.01, batch size 512 and gradient clipping at 1.0. B1 has nothing to select, and the learned models select only through stopping rules that never see the test bin. The pathway model keeps its best validation-Spearman checkpoint within 60 epochs (patience 8), B2 applies the scikit-learn default early stopping to a tenth of the training rows, and B3 trains for a fixed 30 epochs, which can only favor the pathway model in Table 5. All preprocessing is fitted on training rows, each test bin is scored once, and the untuned scores are therefore conservative.

4. Results

4.1. *Combination overlap survives cold-start protocols*

As shown in Table 1, leave-study-out retains only 9.8% of test combinations, whereas leave-cell-line-out and leave-tissue-out retain 72.8% and 85.4%, respectively. Replicating this audit within individual contributing screens (Table 2) reveals that pooling the corpus actually dilutes the true overlap. For fully crossed screens, such as ALMANAC, FRIEDMAN, and O'Neil, holding out a cell line removes absolutely no combinations, resulting in 100% overlap, while the partially crossed AstraZeneca screen yields 92% retention. When evaluating leave-one-tissue-out across all eight eligible tissues (Table 3), the retained overlap varies drastically depending on the held-out tissue, ranging from 56.4% for skin to 100.0% for lung, ovary, and kidney. Across these tissue folds, the performance of the context-plus-pair null model correlates almost perfectly with the degree of combination overlap ($r = 0.977$). This association is not an artifact if fold size, as the retained overlap is uncorrelated with the number of test rows in a fold ($r = 0.081$ against log test rows, $p = 0.85$). In addition, the overlap - B1 association is undiminished when fold size is partialled out (partial $r = 0.984$, $p < 0.001$).

Retention under biological partitioning follows from the design of fully crossed screens, which assay every pair across every available cell line, so withholding a context masks the context-specific measurements without removing the pair from training.The correlation between baseline performance and retained overlap is consistent with simple models recovering the mean effect of a specific pair wherever retention is high. We report it as an association rather than a causal estimate: folds differ in tissue biology and in size as well as in overlap, and it is the analytic argument of Section 4.2, not this correlation, that isolates the contribution of pair retention.

4.2. *The overlap is large enough to invert model rankings*

Across the evaluated protocols (Table 4, Fig. 1), the performance of the null model broadly tracks the degree of combination overlap ($r = 0.97$). Under leave-cell-line-out (72.8% overlap), the null model outperforms both the symmetric multilayer perceptron and the pathway-informed model, by 0.055 and 0.121 Pearson respectively; paired over the six seeds, both margins are significant ($t = 7.71$, $p = 0.0006$ and $t = 8.81$, $p = 0.0003$). The same ordering holds under leave-tissue-out, where the null model exceeds the multilayer perceptron in all

Table 1. Fraction of test combinations overlap with training.

Protocol	Test Rows	Test Pairs Present in Train Set	Held Out Combinations
Random	18,076	94.0% ± 0.3%	no
Leave-tissue-out	4,372–71,638	85.4% ± 18.6%	no
Leave-cell-out	18,076	72.8% ± 0.1%	no
Leave-study-out	15,279	9.8% (fixed)	mostly
Leave-drug-out	31,728	0.0%	yes
Leave-combo-out	18,074	0.0%	yes

Table 2. Per-study Replication Statistics for leave-cell-out and leave-combo-out protocols.

Study	Rows	Pairs	Rows Per Pair	Random	Leave Cell Out	Leave Combo Out
ALMANAC	109,897	2,992	36.7	100% (0.677)	100% (0.653)	0% (0.057)
FRIEDMAN	50,107	2,926	17.1	100% (0.380)	100% (0.400)	0% (0.100)
ONEIL	9,369	315	29.7	100% (0.682)	100% (0.600)	0% (0.135)
ASTRAZENECA	5,910	407	14.5	100% (0.421)	92% (0.298)	0% (0.323)

Table 3. Statistics for leave-one-tissue-out on 8-fold cross-validation.

Held Out Tissue	Test Rows	Overlap	B1	B3	Pathway Model
Lung	25,612	100.0%	0.668	0.574	0.795
Ovary	16,912	100.0%	0.677	0.578	0.631
Kidney	11,845	100.0%	0.628	0.582	0.829
Large intestine	16,736	99.9%	0.642	0.544	0.561
Breast	23,644	93.1%	0.609	0.489	0.508
Brain	4,372	67.9%	0.488	0.471	0.795
Haemato./Lymphoid	4,693	66.3%	0.418	0.349	0.313
Skin	71,638	56.4%	0.325	0.210	0.192
Mean over folds		85.4%	0.557	0.475	0.578
Weighted by test rows		79.6%	0.505	0.403	0.461

eight cross-validation folds of Table 3 by a mean of 0.082 ($t = 6.52$, $p = 0.0003$); against the pathway model, however, the fold-to-fold spread is too wide to resolve a difference (-0.021, $p = 0.73$), which is the same limitation recorded for that protocol in Table 5. An off-the-shelf gradient-boosting regressor exceeds both neural baselines under leave-cell-line-out, by 0.005 and 0.071. The inversion is specific to this regime rather than to overlap in general: under random partitioning, where overlap is highest at 94.0%, the null model does not invert the ranking but instead recovers 79% of the multilayer perceptron's correlation (0.588 against 0.745) without using any molecular information. Leave-cell-line-out is the more consequential case precisely because it is the protocol adopted to control for the permissiveness of random splitting, and it is there that a model with no chemistry wins outright. Conversely, under protocols that strictly enforce combination novelty, such as leave-study-out (9.8% overlap) and leave-combo-out (0.0% overlap), the performance of all models decreases precipitously, and the null model performs poorly.

Table 4. Summary of Pearson Correlation on Loewe Reference Model.

Protocol	Overlap	B1	B2	B3	Pathway Model
Random	94.0%	0.588	0.630	0.745	0.744
Leave-tissue-out	85.4%	0.557	0.499	0.475	0.578
Leave-cell-out	72.8%	0.425	0.375	0.370	0.304
Leave-study-out	9.8%	0.111	0.146	0.139	0.123
Leave-drug-out	0.0%	0.125	0.216	0.245	0.224
Leave-combo-out	0.0%	0.200	0.378	0.594	0.618

This inversion of model rankings is consistent with high combination overlap permitting baseline models to exploit pair-specific memorization rather than learning transferable representations. Complex neural models, constrained by their architecture and regularization, may fail to memorize these pair-specific effects as efficiently as simpler models or explicit baselines, leading to lower apparent performance on the overlapping rows. It is a mismatch rather than a leakage when the resulting number is read as evidence of generalization to unseen combinations. When combination novelty is enforced instead, the memorization advantage vanishes, and all models are forced to rely on genuine generalization, resulting in uniformly lower performance that reflects the actual complexity of predicting unseen drug combinations.

Under leave-cell-line-out every test cell line is unseen, so the per-cell-line term of B1 backs off to the global mean on every test row (Section 3.3) and B1 reduces to the global mean plus the per-pair deviation. On test rows whose pair is absent from training, the pair term backs off as well, leaving an identical constant prediction on every such row and no predicted variance to correlate against. By construction, the leave-cell-line-out correlation of 0.425 is therefore carried entirely by the 72.8% of test pairs retained in training. This is an isolation of pair overlap at a fixed split, as it holds the protocol, the training data and the novelty of the held-out cell line constant, and varies only whether a test row's pair was seen. It applies to B1 analytically and not to the learned models, for which the corresponding control, holding the test rows fixed and deleting from training every observation that involves a test pair, requires retraining on a purpose-built partition; we did not run it and record it in Section 6 as the experiment that would complete this argument. Across protocols, the comparison is correspondingly associative, since the held-out entity, the training and test sizes and the intrinsic difficulty of the task all move together with overlap.

4.3. *The protocol determines the verdict*

When comparing the pathway-informed model to a standard neural baseline across different protocols (Table 5), the inclusion of biological structure yields a statistically significant performance improvement of 0.025 only when combinations are strictly held out ($p < 0.0001$). Under random partitioning and strict drug hold-out, the structural prior provides no significant benefit, while under cell-line hold-out, the inclusion of the pathway structure actually results in a statistically significant performance degradation of 0.067 ($p = 0.0018$). Leave-tissue-out carries the largest nominal effect of any protocol, +0.103, yet it is the one protocol under

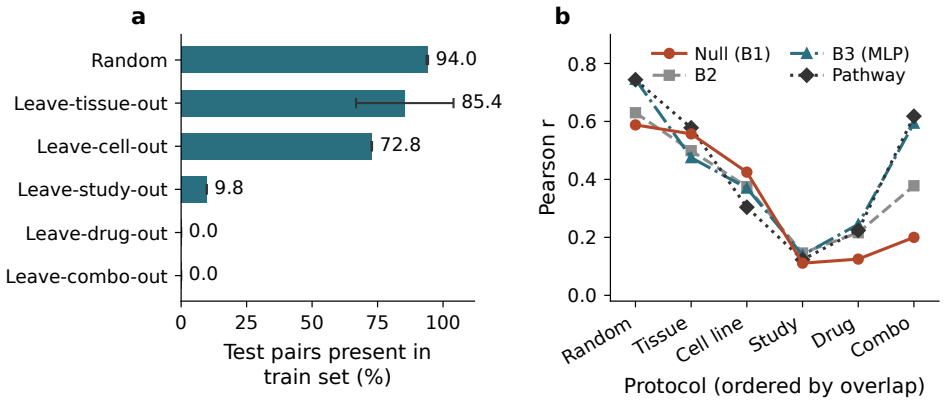


Fig. 1. (a) Combination overlap retained by each protocol. (b) Pearson correlation of each method under the same protocols, ordered by overlap. The null model B1 tracks overlap across the six protocols ($r = 0.97$) and across the eight leave-one-tissue-out folds of Table 3, in which only the held-out block varies ($r = 0.977$); protocols differ in more than overlap, so these are associations.

Table 5. Statistics for Pathway model minus MLP.

Protocol	Pearson Delta	T	P	Conclusion Supported
random	-0.001 ± 0.005	-0.58	0.59	structure is irrelevant
leave-combo-out	+0.025 ± 0.002	29.7	< 0.0001	structure helps
leave-drug-out	-0.021 ± 0.036	-1.44	0.21	structure is irrelevant
leave-cell-out	-0.067 ± 0.027	-6.04	0.0018	structure hurts
leave-tissue-out	+0.103 ± 0.139	2.11	0.073	not resolvable

which no verdict can be reached: the fold-to-fold spread is 0.139, wider than the effect itself, giving $p = 0.073$ over the eight folds. The obstacle is heterogeneity between tissues rather than a small effect, and it is consistent with the fold-level variation documented in Table 3, where retained overlap ranges from 56.4% to 100.0%. A protocol that partitions on a coarse categorical block therefore cannot adjudicate architecture at this sample size, which is itself a reason to prefer protocols whose held-out unit matches the entity under test. The magnitude of these architecture-driven effects is remarkably small, particularly when contrasted with the massive performance fluctuations induced simply by changing the evaluation protocol, where the pathway model's performance ranges from 0.123 to 0.744.

One reading is that where overlap is high the memorized pair effect is available and a structural prior is largely redundant, whereas enforcing combination novelty removes that shortcut and leaves the prior something to contribute. Whatever the mechanism, a single-protocol evaluation cannot separate a genuine benefit from biological structure from an interaction with the pair coverage that a given split leaves in place.

4.4. *What the pathway model actually relies on*

Component-wise ablation locates this model's performance in the drug-target branch rather than in the topology of the pathway graph. What the ablation does not establish is why that

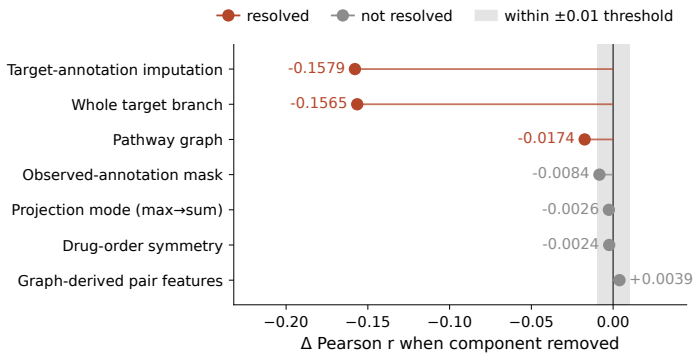


Fig. 2. Ablation deltas on leave-combo-out. Only the pathway graph and the target-annotation imputation exceed the 0.01 resolution threshold, and they differ by an order of magnitude. Removing the imputation head and removing the whole target branch are indistinguishable, which is why the branch's contribution cannot be attributed to annotation completion alone.

branch matters.

Ablation studies conducted under strict combination hold-out (Fig. 2, full model $r = 0.6246$) demonstrate that removing the pathway graph structure results in a negligible decrease in predictive performance of 0.017. In stark contrast, removing the imputation of missing drug-target annotations, or removing the entire target-processing branch, causes a massive drop in performance of approximately 0.158. Furthermore, stratifying model performance by the annotation status of the drug pairs reveals that the model performs better on pairs where neither drug has measured target annotations (0.682) compared to pairs where both drugs are fully annotated (0.582).

Database completion is one reading, under which imputation densifies a sparse annotation matrix and hands the predictor a less noisy profile than the measured annotations alone. It is not the only one, and the ablation cannot select among them: removing the imputation head (-0.1579) and removing the entire target branch (-0.1565) cost the same, so the intervention deletes the branch's learned chemical representation, auxiliary supervision, parameters and optimization path together with the imputed annotations. A capacity-matched control, a head of equal size trained without target supervision, would apportion the drop, and we did not run it (Section 6). The stratified comparison is likewise suggestive rather than decisive, as the advantage on unannotated pairs (0.682 against 0.582) is confounded with the study bias of public databases. What the ablation does support is the contrast that motivates it: the pathway graph is worth -0.0174 here, an order of magnitude less than the target branch, so a reported benefit from biological structure should be resolved into topology and annotation handling before it is attributed to either. Resting on a single architecture, this does not extend to structure-informed models in general.

5. Discussion

Cold-cell-line prediction is pivotal in predicting how a known combination behaves in an uncharacterized cell line. Although such a finding is legitimate and clinically relevant, it cannot be relied upon as a test of combination generalization. A model might perform well on the

former while learning essentially nothing about the underlying chemistry. Hence, a single performance metric cannot adjudicate both capabilities. These are two tasks, not one task and one flaw, and neither is simply the harder version of the other. The symmetric neural baseline scores higher under leave-combo-out than under leave-cell-line-out (0.594 against 0.370), while the null model shows the opposite ordering (0.200 against 0.425), so the protocols differ in the kind of information they leave available rather than along a single axis of difficulty. The corrective we propose is therefore reporting rather than replacement: pair a protocol with the claim it can carry, and publish the combination overlap that tells a reader which claim that is.

Combinatorial screens are dense in biological contexts but sparse in combination space, hence, withholding cell lines barely perturbs the coverage of drug pairs. This gap remains invisible in aggregate model performance and becomes apparent only when pair-level overlap is considered explicitly.

Attributing performance gains to biological structure likewise warrants caution. In the model examined here the advantage sits in the drug-target branch rather than in pathway topology, and studies reporting benefits from biological priors would strengthen the claim by separating knowledge injection from data completion with capacity-matched ablations.

Relative to the shortcut-learning, what the previous work doe not measure is the residue that produces the effect: how many test combinations each cold-start protocol actually leaves in training. Our work quantifies that residue protocol by protocol and screen by screen, then attribute it to full crossing of the pair-by-cell-line matrix, and use it to show that the protocol determines the verdict on structural priors and that the apparent benefit of those priors is traceable to annotation completion.

6. Limitations

DrugComb is the only corpus used in this study. The overlap audit is executed within four contributing screens. Among those, three are fully crossed while one is only partially crossed. Hence, the study addresses the effect of pooling across sources but not that of harmonization across assays.

The ablation study is anchored to a single pathway-informed model. The study does not claim that all such models will behave identically. Extending it would require the same ablations on other structure-injection mechanisms, such as subsystem models, network-proximity scores and graph transformers. Rather, we observe that the attribution of gains to biological structure remains unmeasured in prior work reporting pathway benefits.

We observe that two of the non-neural baselines surpass both neural architectures under leave-cell-line-out. This observation alone is unable to establish that no neural model could recover the per-pair mean from learned embeddings. A hybrid design, comprised of B3 with the per-cell-line and per-pair terms learned by B1 as explicit features would disentangle whether the protocol leaves too much exploitable signal.

In this study, we quantify overlap and demonstrate its consequences using our own baselines. Confirming that specific published results change under an overlap-controlled protocol requires re-training those methods, and we regard this as the most important extension.

Two controls are absent, one per central claim. For pair retention, holding the test rows fixed and deleting from training every observation involving a test pair would vary retention alone, with the held-out cell lines, the test distribution and the model constant; we offer instead the analytic reduction of B1 (Section 4.2) and the size-adjusted association of Section 4.1. For the ablation, a capacity-matched head trained without target supervision would separate annotation completion from the representation and capacity the target branch also contributes. Until both are run, the corresponding statements here are associations and component locations rather than causal attributions.

7. Conclusion

In summary, this study showcased that standard cold-start evaluation protocols for drug-combination synergy prediction retain a substantial and screen-dependent fraction of test combinations in training, and that this residue dominates both the reported numbers and the apparent utility of biological priors. The retention is consistent with the task leave-cell-line-out defines and should not be inferred as leakage, as it becomes an inferential error only when such numbers are taken as evidence of generalization to unseen combinations. Leave-cell-line-out and leave-one-tissue-out evaluations often retain the majority of test combinations within the training set depending on the underlying screen design. Under leave-cell-line-out, the protocol adopted specifically to correct for permissive random splitting, a simple baseline predictor lacking molecular information outperforms both conventional neural architectures outright, and it recovers most of their correlation under random splitting as well. As the choice of evaluation protocol can invert the apparent utility of the structured biological information, we recommend that future studies routinely report combination overlap alongside a context-plus-pair null model. Both are computationally inexpensive and would make the combination coverage a split leaves in place visible at the point where performance is reported.

Acknowledgements

This work was partially supported by the Commonwealth Health Research Board (CHRB-2360623, PI: Das), the National Science Foundation (NSF-2316003, PI: Ghosh), the National Institutes of Health (1S10OD038214-01, PI: Ghosh; 1T32TR006000-01, PI: Cifu; 1R01AR086251-01A1, PI: Boyan), the VCU Breakthrough award (PI: Ghosh), and VCU CCTR Endowment funds (PI: Thangarajh) awarded to P.G.

Code Availability

Scripts to reproduce the results presented in this work can be accessed at https://github.com/debnathk/drugsyn-audit

Use of Large Language Models

A large language model was used as a coding and drafting assistant: for parts of the analysis and plotting code. All experiments, numerical results, and scientific claims were produced and verified by the authors, who take full responsibility for the content.

References

1. K. Preuer, R. P. I. Lewis, S. Hochreiter, A. Bender, K. C. Bulusu, and G. Klambauer. DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. *Bioinformatics*, 34:1538–1546, 2018.
2. J. O'Neil, Y. Benita, I. Feldman, M. Chenard, B. Roberts, Y. Liu, et al. An Unbiased Oncology Compound Screen to Identify Novel Combination Strategies. *Mol. Cancer Ther.*, 15:1155–1162, 2016.
3. S. L. Holbeck, R. Camalier, J. A. Crowell, J. P. Govindharajulu, M. Hollingshead, L. W. Anderson, et al. The National Cancer Institute ALMANAC: A Comprehensive Screening Resource for the Detection of Anticancer Drug Pairs with Enhanced Therapeutic Activity. *Cancer Res.*, 77:3564–3576, 2017.
4. B. Zagidullin, J. Aldahdooh, S. Zheng, W. Wang, Y. Wang, J. Saad, et al. DrugComb: an integrative cancer drug combination data portal. *Nucleic Acids Res.*, 47:W43–W51, 2019.
5. S. Zheng, J. Aldahdooh, T. Shadbahr, Y. Wang, D. Aldahdooh, J. Bao, et al. DrugComb update: a more comprehensive drug sensitivity data repository and analysis portal. *Nucleic Acids Res.*, 49:W174–W184, 2021.
6. P. Dewulf, M. Stock, and B. De Baets. Cold-Start Problems in Data-Driven Prediction of Drug-Drug Interaction Effects. *Pharmaceuticals*, 14:429, 2021.
7. S. Monem et al. A multi-view feature representation for predicting drugs combination synergy based on ensemble and multi-task attention models. *J. Cheminform.*, 2024.
8. B. Boonyarit et al. SynProtX: a large-scale proteomics-based deep learning model for predicting synergistic anticancer drug combinations. *GigaScience*, 2025.
9. W. Alam et al. Unlocking the therapeutic potential of drug combinations through synergy prediction using graph transformer networks. *Comput. Biol. Med.*, 2024.
10. S. Kapoor and A. Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4:100804, 2023.
11. S. Kapoor and A. Narayanan. Leakage and the Reproducibility Crisis in ML-based Science. *arXiv* preprint, 2022.
12. J. Bernett et al. Guiding questions to avoid data leakage in biological machine learning applications. *Nat. Methods*, 2024.
13. M. Rosenblatt et al. Data leakage inflates prediction performance in connectome-based machine learning models. *Nat. Commun.*, 2024.
14. A. H. C. Vlot, N. Aniceto, M. P. Menden, G. Ulrich-Merzenich, and A. Bender. Applying drug synergy metrics to oncology combination screening data: agreements, disagreements and pitfalls. *Drug Discov. Today*, 2019.
15. D. Duarte and N. Vale. Evaluation of synergism in drug combinations and reference models for future orientations in oncology. *Curr. Res. Pharmacol. Drug Discov.*, 2022.
16. K. R. Roell, D. M. Reif, and A. A. Motsinger-Reif. An Introduction to Terminology and Methodology of Chemical Synergy — Perspectives from Across Disciplines. *Front. Pharmacol.*, 2017.
17. O. Fuentealba-Manosalva et al. Mind the Curve: Dose-Response Fitting Biases the Synergy Scores across Software Used for Chemotherapy Combination Studies. *Int. J. Mol. Sci.*, 2023.
18. F. Abbasi and J. Rousu. New methods for drug synergy prediction: a mini-review. *Curr. Opin. Struct. Biol.*, 2024.
19. H. Yousefabadi et al. Bioactivity-Driven Prediction of Antibacterial Synergy Using Machine Learning Models. *bioRxiv* preprint, 2025.
20. O. E. Agboola et al. End-to-end multimodal attention fusion for drug-target interaction prediction with cold-scaffold validation on the Davis kinase benchmark. *J. Mol. Graph. Model.*, 2026.

21. F. Cheng, I. A. Kovacs, and A. L. Barabasi. Network-based prediction of drug combinations. *Nat. Commun.*, 10:1197, 2019.
22. Q. T. Trac et al. Pathway activation model for personalized prediction of drug synergy. *eLife*, 2025.
23. B. M. Kuenzi et al. Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. *Cancer Cell*, 38:672–684.e6, 2020.
24. H. I. Kuru, O. Tastan, and A. E. Cicek. MatchMaker: A Deep Learning Framework for Drug Synergy Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 19:2334–2344, 2022.
25. W. B. Sherine et al. Graph Enhanced Transformers for Predicting Drug using Drug Synergies and Antagonistic Interactions. In *ICESC*, 2025.
26. D. Chen et al. Predicting anticancer synergistic drug combinations based on multi-task learning. *BMC Bioinformatics*, 2023.
27. Z. Wang et al. DEML: Drug Synergy and Interaction Prediction Using Ensemble-Based Multi-Task Learning. *Molecules*, 2023.
28. A. J. Preto et al. SYNPRED: prediction of drug combination effects in cancer using different synergy metrics and ensemble learning. *GigaScience*, 2022.
29. J. Lin et al. Pisces: A multi-modal data augmentation approach for drug combination synergy prediction. *Cell Genomics*, 2023.
30. S. Kapoor et al. REFORMS: Consensus-based Recommendations for Machine-learning-based Science. *Sci. Adv.*, 2024.
31. J. Zhu et al. Machine Learning in Environmental Research: Common Pitfalls and Best Practices. *Environ. Sci. Technol.*, 2023.
32. H. Semmelrock et al. Reproducibility in Machine Learning-based Research: Overview, Barriers and Drivers. *AI Magazine*, 2024.
33. Emine Beyza Çandır, Halil İbrahim Kuru, Magnus Rattray, A Ercüment Çiçek, and Oznur Tastan. One-hot news: drug synergy models shortcut molecular features. *Bioinformatics*, 42(3):btag040, 2026.
34. M. Schubert, B. Klinger, M. Klunemann, A. Sieber, F. Uhlitz, S. Sauer, et al. Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nat. Commun.*, 9:20, 2018.
35. A. Tsherniak, F. Vazquez, P. G. Montgomery, B. A. Weir, G. Kryukov, G. S. Cowley, et al. Defining a Cancer Dependency Map. *Cell*, 170:564–576.e16, 2017.
36. J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh, and P. Das. Large-scale chemical language representations capture molecular structure and properties. *Nat. Mach. Intell.*, 4:1256–1264, 2022.
37. B. Zdrazil, E. Felix, F. Hunter, E. J. Manners, J. Blackshaw, S. Corbett, et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res.*, 52:D1180–D1192, 2024.
38. A. Liberzon, C. Birger, H. Thorvaldsdottir, M. Ghandi, J. P. Mesirov, and P. Tamayo. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.*, 1:417–425, 2015.
39. D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.*, 51:D638–D646, 2023.