THE PLATONIC GENE: CONVERGENT REPRESENTATIONS ACROSS BIOLOGICAL MODELS LEARN CELL PLAUSIBILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein coding genes can be represented in myriad ways, including their nucleic acid or protein sequences, molecular networks, textual annotations, co-expression in cells, and perturbation phenotypes. The Platonic representation hypothesis proposes that increasingly capable models converge toward shared representations of the underlying world, even when trained on different modalities. Here, we test this hypothesis for gene biology. We find that representations from different modalities, architectures, and training objectives can agree on the relationships between genes. This agreement is stronger among models that predict biological processes better, in accordance with the Platonic representation hypothesis. Nonetheless, convergence is not perfect, raising the question whether differences between representations capture complementary biological information that can be used to obtain a more holistic representation. We show that combining six models into a shared PLANTAGENET consensus representation improves gene-property prediction. We use the geometry of genes in the PLANTAGENET consensus space to assess the biological plausibility of single-cell expression profiles. In a combined human blood and brain evaluation, cell coherence and dominant-dimensionality scores distinguish real observed cells from gene-shuffled, independently sampled gene counts, and cross-tissue chimeric expression profiles, achieving AUROCs of 1.000, 0.992, and 0.845, respectively, outperforming individual models. These results suggest that biological models capture both shared and complementary gene relationships, and that integrating these relationships provides a geometric approach to assess gene-expression coherence, with potential implications for biological discovery and cell engineering.

1 INTRODUCTION

Biological systems are both multi-scale, across different levels of biological organization – from atoms to cells, tissues, organisms and populations – and multi-modal, with multiple ways to measure both within and across different scales. Experimental assays are constantly being developed to measure and characterize biology across scales, with each method capturing one or more different biological aspects. Genes, the basic physical and functional units of heredity, can be represented across modalities (Figure 1); as nucleotide (Avsec et al., 2026) or amino acid (Lin et al., 2023) sequences, as structures of their encoded RNAs and proteins, as members of biological pathways (Szklarczyk et al., 2023), and as the outcome of their genetic perturbation at the molecular, cellular, tissue or organism level, including on cell

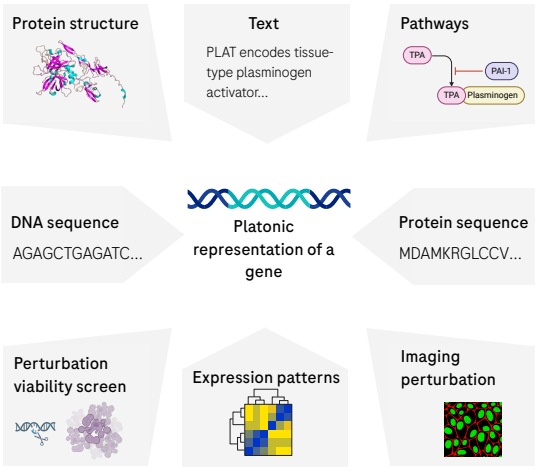


Figure 1: *Views of the same gene, PLAT, across modalities provide a single Platonic representation.*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

morphology (Funk et al., 2022) or gene-expression signatures (Dixit et al., 2016). Deep learning methods, statistical models, and text-based descriptors are widely used to capture cellular- and gene-level representations across these modalities, but have been historically challenging to combine into a fully coherent view.

The Platonic representation hypothesis (Huh et al., 2024) proposes that powerful models understand the world behind data and therefore learn similar representations of given objects, even when trained on different modalities, resembling ideal Platonic forms (Plato, c. 375 BC). Related findings in biology suggest that representations learned from one modality can capture information about another, as illustrated by protein structure prediction from sequence models (Lin et al., 2023) and the prediction of expression defined neuronal types from neuronal activity measurements (Wu et al., 2025). Recent work identified shared representational geometry between independently learned human brain embeddings (Marcos-Manchón et al., 2026), and between perturbed cell RNA and morphological profiles (Kudo et al., 2026).

These findings motivate a corresponding question for genes: do models trained on different biological modalities recover shared gene relationships, and do their differences capture complementary information? Both aspects of the question are biologically intriguing; convergence that preserves relationships among genes may reflect shared biological organization, and cross modality-differences can reveal complementary information that strengthens joint representations. Integrating these shared and modality-specific gene properties could support a more comprehensive and unified computational view of cell biology (Rood et al., 2024). Such cell-level representations could enable biological discovery through improvement of cell engineering by helping distinguish biologically plausible cell states from those that may only appear realistic (Bunne et al., 2024).

In this work, we test the Platonic representation hypothesis for genes and introduce PLANTAGENET[1] consensus, a shared embedding space that integrates complementary views of gene biology learned from different modalities. Our contributions are:

- We compare gene representations across modalities, architectures, and training objectives, showing that cross-modal agreement correlates with models' biological understanding.
- We construct PLANTAGENET consensus embeddings, showing that cross-modal integration improves performance on a benchmark of gene-property prediction tasks.
- We use the PLANTAGENET consensus space geometry to define cell coherence and dominant-dimensionality scores, which distinguish true observed biological cells from biologically implausible profiles.

2 PRIOR WORK

The analogy with Plato's ideal forms (Plato, c. 375 BC) has also informed accounts of biological organization (Denton et al., 2002; Levin, 2026). Denton et al. (2002) argue that protein folds can be understood as lawful natural forms constrained by physics, not just contingent products of evolution. Such proposals concern the emergence of biological forms; they are distinct from the empirical question of whether learned representations converge.

In representation learning, Huh et al. (2024) introduced the Platonic representation hypothesis, reporting increasing agreement among representations learned by more capable vision and language models. Gröger et al. (2026) subsequently showed that standard similarity measures can be inflated by model width and depth; their calibrated analysis preserves evidence for shared local neighborhoods, while challenging global geometric convergence. Moving from similarity to translation across modalities, Jha et al. (2025) studied a stronger formulation, showing that independently trained text embedding spaces can be translated across without paired data.

Related biological studies recovered shared information across measurements; protein language models capture information that supports atomic-level structure prediction (Lin et al., 2023), and Wu et al. (2025) learn representations from neuronal activity that can predict expression-based cell types. These findings demonstrate translation, but do not establish convergence between cross-modal representations. More directly, Marcos-Manchón et al. (2026) investigate the strong Platonic

[1]The name is a hybrid of "platonic" and "gene" and stands for the House of Plantagenet, which brought diverse territories under common rule in the twelfth-century Angevin Empire.

representation hypothesis in the human visual cortex. They independently learn subject-specific embeddings from repeated fMRI responses to natural images and align these spaces using orthogonal transformations fitted without paired cross-subject samples. Combining the pairwise alignments into a common latent space improves cross-subject retrieval, supporting an approximately shared representational geometry across individuals.

To the best of our knowledge, there is no prior work discussing the Platonic representations of gene modeling. A blog post by Kalashnikov (2026) explores a related question, an agreement of drug embeddings obtained from cellular morphology and gene-expression measurements. At the gene level, Kan-Tor et al. (2024) introduce a modality-agnostic benchmark that evaluates gene embeddings through the prediction of curated biological properties. This benchmark assesses how accurately these properties can be predicted from embeddings, rather than whether different models learn similar representations. We use linear probes on the benchmark's classification tasks to assess the biological information captured by individual and PLANTAGENET consensus embeddings.

3 METHODS

3.1 GENE EMBEDDING MODELS

We evaluated 33 gene representations that span complementary views of gene biology (Figure 1, Table S1). The representations capture: textual descriptions, protein and regulatory DNA sequences, curated pathways and protein-interaction networks, atlas-scale single-cell and spatial gene-expression profiles, and perturbation phenotypes measured through viability or microscopy. These representations are produced by language and sequence-to-function models, graph-embedding methods, matrix-factorization approaches, single-cell foundation models, and already integrated embeddings, allowing us to compare representations that differ simultaneously in data modality, training objective, and architecture. We also include random (isotropic Gaussian) vectors of dimensions 128, 512, and 5120 as negative controls. We define $\mathcal{G}(m_1, m_2)$ to be the set of genes shared across models m_1 and m_2, $\mathcal{G}$ to be the set of genes shared between all models $m_i \in \mathcal{M}$, and $X_{m_i} \in \mathbb{R}^{|\mathcal{G}| \times d_{m_i}}$ to describe the native gene embeddings produced by model m_i.

3.2 TESTING REPRESENTATION CONVERGENCE

Following Huh et al. (2024), we use the mutual nearest neighbor (mNN) metric as a measure of representation convergence. Let $d_{\text{knn}}(g, m_i)$ be the set of k nearest neighbors of gene g (excluding g itself) in the gene embedding space of the model m_i. Then:

$$mNN(m_1, m_2) = \frac{1}{k|\mathcal{G}(m_1, m_2)|} \sum_{g \in \mathcal{G}(m_1, m_2)} |d_{\text{knn}}(g, m_1) \cap d_{\text{knn}}(g, m_2)|\,, \tag{1}$$

using cosine distances between gene embeddings and setting $k = 10$. The raw overlap $mNN(m_1, m_2)$ is not comparable across model pairs; the chance level of neighbor agreement depends on the number of shared genes $|\mathcal{G}|(m_1, m_2)$: pairs with less shared genes are biased towards higher overlap. To correct for this, we normalize each pair against its own null permutation, mNN_{null}; given by averaging P occurrences of σ_p, a uniformly random permutation of the shared gene set. We set $P = 127$ and report the fold over random, defined as $mNN_{excess} := mNN/mNN_{\text{null}}$.

3.3 GENE-PROPERTY PREDICTION BENCHMARK

We evaluate gene embeddings using classification tasks from the gene-property prediction benchmark introduced by Kan-Tor et al. (2024). We exclude regression tasks, treat each binary or multiclass task as a single property-prediction problem, and evaluate each label of a multi-label task separately. A property is used if at least ten genes in the set of genes under study have it. The tasks cover five categories: biological processes (9 tasks, 320 properties), protein properties (3 tasks, 99 properties), localization (*i.e.* tissue specificity) (19 tasks, 166 properties), regulatory functions (4 tasks, 4 properties), and genomic properties (5 tasks, 58 properties). For each property, we train a logistic regression probe on the gene embeddings and evaluate performance using five-fold cross-validation. Each tasks' AUROC is the median across its properties' AUROC. The 647 properties

listed above are those retained when the set of genes is the intersection of the eight views used in PLANTAGENET VIII consensus (Table 1, and Figure S1) and for the individual models, the linear probe was applied to the 128 PCs embeddings reduced. In Figure 2, Figure S3 and Figure S4, the set of genes is the union of the 33 views, yielding 663 properties, and the linear probe was applied to the native embeddings.

3.4 CONSTRUCTION OF PLANTAGENET CONSENSUS-SPACES

We first map embeddings of every model m_i to a normalized 128-dimensional representation (for equal contribution to the consensus):

$$Z_m := \mathrm{RowNorm}(\mathrm{PCA}_{128}(\mathrm{Center}(\mathrm{RowNorm}(X_{m_i})))) \,.$$

Using these preprocessed views, we construct the concatenation and Procrustes PLANTAGENET consensus spaces described in Algorithm 1 and Algorithm 2, respectively.

3.5 SELECTING BUILDING BLOCKS FOR THE PLANTAGENET CONSENSUS SPACE

We construct the initial PLANTAGENET consensus space from eight gene-representation views, six selected models (Figure S1) and two atlas-derived PCA representations form the original 33 models (Table S1). We first exclude models whose gene set covers less than 15,000 of the 20,020 canonical GENCODE genes (Mudge et al., 2025, Table S2). We rank eligible models using cross-family mutual 10-nearest-neighbor agreement (Figure 2, Figure S5, and Figure S6). For each model, we identify its highest agreement with any member of other model family, and average these family-specific maxima to obtain an overall cross-family agreement score used to select the highest-ranking model from each family (Figure S5, Figure S6, with an exception for the two foundation model group, see below).

Algorithm 1: Concatenation consensus

Require: $\{Z_m\}_{m\in\mathcal{M}}$
Ensure: Z^{concat}
$Z^{\mathrm{concat}} \leftarrow [Z_1 \mid \cdots \mid Z_{|\mathcal{M}|}]$
$Z^{\mathrm{concat}} \leftarrow \mathrm{RowNorm}(Z^{\mathrm{concat}})$
return Z^{concat}

Algorithm 2: Procrustes consensus

Require: $\{Z_m\}_{m\in\mathcal{M}}$, tolerance ε
Ensure: Z^{proc}
$\bar{Z} \leftarrow |\mathcal{M}|^{-1} \sum_m Z_m$
repeat
  for all $m \in \mathcal{M}$ **do**
    $U_m \Sigma_m V_m^\top \leftarrow \mathrm{SVD}(Z_m^\top \bar{Z})$
    $A_m \leftarrow Z_m U_m V_m^\top$
  end for
  $\bar{Z}' \leftarrow |\mathcal{M}|^{-1} \sum_m A_m$
  $\Delta \leftarrow \|\bar{Z}' - \bar{Z}\|_F$
  $\bar{Z} \leftarrow \bar{Z}'$
until $\Delta \le \varepsilon$
$Z^{\mathrm{proc}} \leftarrow \mathrm{RowNorm}(\bar{Z})$
return Z^{proc}

The six selected models comprise of: *GenePT-protein* which represents genes through natural-language descriptions (Chen & Zou, 2025); each gene is described by a `text-embedding-3-large` embedding of the UniProt summary of its protein product. *AlphaGenome* is a sequence-to-function model trained to predict genomic measurements from DNA (Avsec et al., 2026); each gene represented by mean-pooled encoder features from a 524,288-bp DNA sequence centered on its transcription start site. *ESM-2*, a protein language model trained on amino-acid sequences (Lin et al., 2023); each gene is represented by the mean of the residue embeddings of its protein product. *DepMap CRISPR* represents genes through their effects on cancer-cell viability across cell lines. We use the representation from PRESAGE (Littman et al., 2025), comprising gene loadings from a 128-component PCA of DepMap 24Q2 CRISPR gene-effect profiles (DepMap, Broad, 2024). The remaining two are gene-expression based foundation models, *scGPT* pretrained on human single-cell RNA-seq (Cui et al., 2024), and *CellPLM* pretrained on both single-cell RNA-seq and spatial transcriptomic data (Wen et al., 2024, we retain both to leverage the spatial information captured in the latter). For both models, we extract the static learned gene vectors from their pretrained gene-embedding tables.

We supplement these six models with two direct representations of tissue-specific gene-expression variation: *PCA (blood)*, derived from the blood cells in Tabula Sapiens (Tabula Sapiens Consortium, 2022), and *PCA (brain)*, derived from the Siletti adult-human-brain atlas (Siletti et al., 2023). PCA is fitted separately to each atlas, and each gene is represented by its loadings on the first 128 principal components.

3.6 CELL GEOMETRY SCORES

For a cell c expressing gene $g \in \mathcal{G}_c \subset \mathcal{G}$ with raw count $x_{c,g}$, we define gene weights $w_{c,g}$ as the normalized log-transformed count, described in Algorithm 3. Then, we extract two cell geometry scores: a first moment score using a vector cell representation $z_c \in \mathbb{R}^d$ (Algorithm 3) and a second moment score using a matrix cell representation $Z_c \in \mathbb{R}^{d \times |\mathcal{G}_c|}$ (Algorithm 4). The vector cell representation z_c is the weighted average of its expressed genes' embeddings and the coherence score ρ_c is defined as $||z_c||$. The matrix cell representation is defined by its column vectors being $\sqrt{w}_g z_g$. The left singular vectors $u_i \in \mathbb{R}^d$ from the SVD decomposition of Z_c can be interpreted as the cell's program directions in the gene embedding space, with associated singular values λ_i. We define the number of dominant directions D_c as $\left(\sum_i \lambda_i\right)^2 / \sum_i \lambda_i^2$, equaling 1 when a single program dominates and d if all d directions contribute equally, so $D_c \in [1, d]$. In the cell rejection experiment (Table 2), when using ρ_c, we reject cells for which ρ_c is below a certain threshold whereas when using D_c, we reject cells for which D_c is above a certain threshold. Treating real cells as positives and corrupted cells as negatives, we generate a ROC curve for each corrupted cell type, for each cell geometry score.

Algorithm 3: Coherence ρ_c

Require: gene embeddings $\{z_g\}$, counts $\{x_{c,g}\}$, $\mathcal{G}_c$
Ensure: $\rho_c \in [0, 1]$
1: $w_{c,g} \leftarrow \log(1 + x_{c,g}) / \sum_{h \in \mathcal{G}_c} \log(1 + x_{c,h})$
2: $z_c \leftarrow \sum_{g \in \mathcal{G}_c} w_{c,g}\, z_g$
3: $\rho_c \leftarrow \|z_c\|_2$
4: **return** ρ_c

Algorithm 4: Number of dominant directions D_c

Require: gene embeddings $\{z_g\}$, counts $\{x_{c,g}\}$ of cell c, $\mathcal{G}_c$
Ensure: $D_c \in [1, d]$
1: $w_{c,g} \leftarrow \log(1 + x_{c,g}) / \sum_{h \in \mathcal{G}_c} \log(1 + x_{c,h})$
2: $Z_c \leftarrow \left[\sqrt{w_{c,g}}\, z_g\right]_{g \in \mathcal{G}_c} \in \mathbb{R}^{d \times |\mathcal{G}_c|}$
3: $USV^\top \leftarrow \mathrm{SVD}(Z_c); \quad \lambda_i \leftarrow S_{ii}^2$
4: $D_c \leftarrow \dfrac{\left(\sum_i \lambda_i\right)^2}{\sum_i \lambda_i^2}$
5: **return** D_c

3.7 CELL TYPE MARKER GENE SETS

We obtained 178 human cell types marker gene sets using PanglaoDB[2] (Franzén et al., 2019). We grouped the 178 PanglaoDB cell types into ten cell lineages: neuronal, glial and neural support, immune/hematopoietic, epithelial, endocrine, endothelial, muscle and mural, stromal/mesenchymal, germ/reproductive/placental, and stem/progenitor/embryonic. A gene was then assigned a cell lineage label only if every marker gene sets it belonged to was in the same lineage. The remaining genes were left unlabeled. This resulted in 3,073 labeled genes among the 18,127 genes from the PLANTAGENET VI consensus gene universe.

3.8 CORRUPTED CELLS

Our initial pool of measurements of real cells consists of scRNA-seq profiles of brain and blood cells from the Siletti adult-human-brain (Siletti et al., 2023) and Tabula Sapiens' blood cell atlases (Tabula Sapiens Consortium, 2022), respectively (Table S1). We sampled each atlas with a per-cell type cap of 2,500 and obtained 47,006 cells (24,872 blood, 22,134 brain, cell types distributions shown in Figure S7). We performed three experiments: one with blood cells only, one with brain cells only, one with blood and brain cells. From each pool of real cells, we sampled 20,000 reference real cells and constructed 20,000 *shuffle* cells, 20,000 *marginal* cells and 20,000 *chimeras*. We constructed shuffle cells by taking a real cell scRNA-Seq profile and shuffling its genes. Each marginal cell profile gets a total number of counts taken from a real cell profile. Then, each count is sampled independently from the empirical distribution of gene counts across the pool of real cell profiles. Chimeras are made by interleaving two real cell profiles at random: each gene count is picked at random from one of the two cell profiles. In the experiment with blood and brain cells, the chimeras are always made from one blood and one brain cell profile.

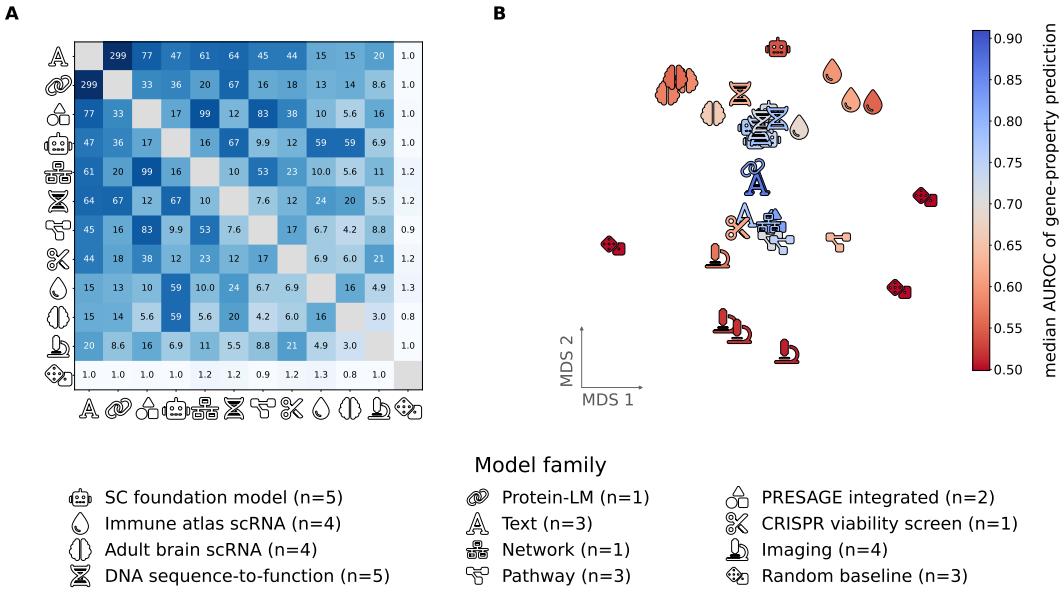


Figure 2: *Platonic representation hypothesis holds in biology for gene embedding models trained on different modalities*. **a**, excess of 10 mutual nearest neighbors over the random expectation for best gene representation per embedding family. **b**, multidimensional scaling plot based on mNN agreement of different gene embedding models. Colors reflect prediction performance on the gene property benchmark (663 properties, Section 3.3).

4 Results

4.1 Gene Embedding Biology Demonstrates Platonic Convergence Behavior

First, we applied the mutual Nearest Neighbor (mNN) framework to gene embeddings derived from different models as well as random embeddings (Section 3.1). We observed a consistently higher than random mNN overlap between all model groups (Figure 2a, Figure S5, Figure S6). GenePT embeddings of protein text descriptions and ESM-2 embeddings of protein sequences reached the highest cross-model family agreement with 299 folds mNN excess over random or 0.16 average fraction of 10 nearest neighbors. This might be expected because these models were trained on similar data. However, even models trained on completely orthogonal modalities, such as perturbation screen of cancer cells with imaging readout and single-cell RNA-Seq models of brain, showed 3 times larger mNN overlap than expected at random. Average agreement between modalities was highest for text embeddings (60x over random) and lowest for imaging (9x over random). It agrees with intuition: text integrates all available information for genes, while imaging only captures cell-morphology changes following gene knockouts in a simple cell line model.

The Platonic representation hypothesis postulates models' alignment increases with scale and performance (Huh et al., 2024). Focusing on models' performance, we evaluate whether this holds for gene embeddings,using linear probes prediction accuracy in the gene-property prediction benchmark (Section 3.3, Figure S3). We observe strong correlation between prediction performance on biological-process tasks and mNN excess (Pearson's $r = 0.75$, $p = 6.6 \times 10^{-7}$, Figure S4), and multidimensional scaling shows that models with better prediction performance cluster together, despite being trained on different modalities (Figure 2b). Weaker models show little structure beyond clustering by modality. Overall, our results demonstrate that properties of the Platonic representation hypothesis hold for cross-modal gene embeddings.

[2]data available at https://panglaodb.se/

	bio. proc. 9 tasks	prot. prop. 3 tasks	localiz. 19 tasks	reg. func. 4 tasks	genomic prop. 5 tasks	all 40 tasks	family mean
AlphaGenome	0.752	0.736	0.894	0.794	0.678	0.842	0.771
ESM-2	0.878	<u>0.951</u>	0.719	0.721	0.724	0.758	0.799
GenePT-protein	<u>0.948</u>	0.922	0.743	0.736	0.753	0.759	0.820
DepMap CRISPR	0.618	0.614	0.596	0.618	0.599	0.668	0.609
scGPT	0.759	0.752	<u>0.913</u>	0.768	0.684	0.849	0.775
CellPLM	0.722	0.705	0.900	0.689	0.656	0.799	0.734
PCA (blood)	0.646	0.667	0.743	0.657	0.613	0.753	0.665
PCA (brain)	0.634	0.655	0.717	0.651	0.602	0.712	0.652
PLANT. VIII (concat)	0.941	0.948	0.884	**0.829**	0.764	0.860	<u>0.873</u>
PLANT. VIII (Procrustes)	0.918	0.903	<u>0.913</u>	<u>0.823</u>	0.771	<u>0.882</u>	0.865
PLANT. VI (concat)	**0.953**	**0.956**	0.889	0.805	<u>0.784</u>	0.876	**0.877**
PLANT. VI (Procrustes)	0.935	0.915	**0.917**	0.805	**0.794**	**0.886**	<u>0.873</u>

Table 1: Gene-property prediction benchmark by biological task family. Each family's AUROC is the median across its tasks' AUROCs. The all tasks' AUROC is the median across all 40 tasks. The family mean is the mean AUROC across the five families. Best gene representation in **bold**, second best are <u>underlined</u>. PLANTAGENET VIII (PLANT. VIII) uses all eight individual views whereas PCA (blood) and PCA (brain) are removed in PLANTAGENET VI(PLANT. VI).

4.2 A GENE CONSENSUS SPACE TO CAPTURE GENE BIOLOGY'S PLATONICITY

Despite a significant convergence of gene representations, the mNN overlap was still limited, especially between model families (Figure 2, Figure S5, Figure S6). Thus, we asked whether the differences between the gene embeddings are biologically meaningful and capture modality-specific information. We selected a subset of 8 of the 33 models (Table S1) as building blocks for a consensus space of gene embeddings, retaining one model with the highest average cross-modality agreement score from each model family: six family representatives (AlphaGenome, ESM-2, GenePT-protein, DepMap, scGPT, CellPLM) along with the blood and brain PCA atlas embeddings. The individual models performs best on different biological task families in the gene-property benchmark: text and protein models (GenePT-protein and ESM-2) capture biological processes, protein properties and genomic properties tasks, RNA and DNA models (scGPT, CellPLM, and AlphaGenome) excel at localization tasks, and Alpha Genome drive regulatory functions tasks (Table 1).

Combining the eight views into a single PLANTAGENET consensus space (PLANTAGENET VIII), by either concatenation or Procrustes alignment (Section 3.4), yielded a representation that outperforms the individual models on the gene-property prediction benchmark (Table 1). PLANTAGENET VIII consensus improves on the single best performing individual model on the two aggregate measures, the median AUROC across all 40 tasks (0.882 vs. 0.849 for scGPT) and the family-mean AUROC (0.873 vs. 0.820 for GenePT-protein). At the family level, it beats the best individual model on two of the five task families (regulatory functions and genomic properties).

To ensure a more balanced holistic cross-modality representation, avoiding gene-expression bias and organ specific biology we removed the two PCA views, forming PLANTAGENET VI consensus. Notably, the formed PLANTAGENET VI consensus improves the representation further: it outperforms the best individual model on all five task families as well as PLANTAGENET VIII on the median over all 40 tasks (0.886 vs. 0.882) and the family-mean AUROC (0.877 vs. 0.873). These results complement the Platonic representation hypothesis by showing that differences between representations are as valuable as commonalities. By leveraging modality specific information and learning consensus representation, a true Platonic form of the underlying objects can be better approximated.

4.3 UNDERSTANDING CELL BIOLOGY FROM INTEGRATED GENE EMBEDDINGS

Having shown that the PLANTAGENET consensus embedding provides a better representation of gene biology, we asked whether it can be used for better representation of cells. As a bridge between these two different scales of biology, we used PanglaoDB's curated cell types' marker gene sets (Franzén et al., 2019) to test if gene embeddings reflect cell lineage's information (Figure S8, Section 3.7) and found that the PLANTAGENET VI consensus embeddings capture cell lineage membership more strongly than any individual model (highest NMI and ARI scores, Table S3). We

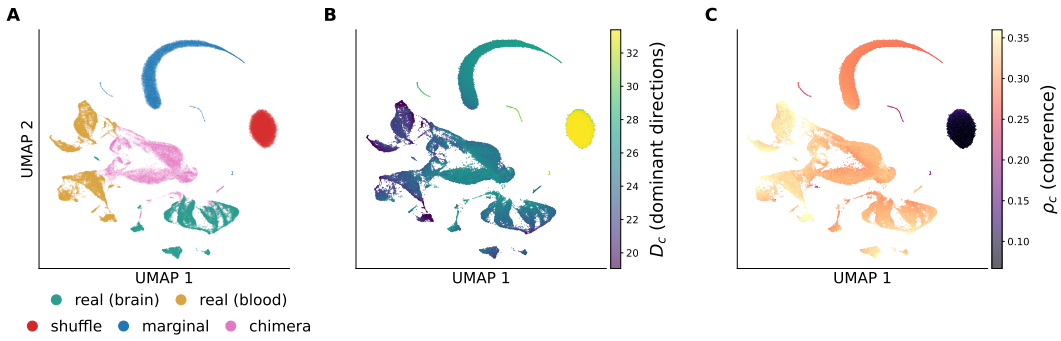


Figure 3: *Landscape of vector cell representations*. **a**–**c**, UMAP visualizations of vector cell representations z_c obtained using PLANTAGENET VI (Procrustes), comprising 20,000 real cells from blood and brain and 20,000 cells from each of the shuffle, marginal, and chimera groups. Cells are colored by group (**a**), number of dominant directions D_c (**b**), and coherence score ρ_c (**c**).

followed the hypothesis that if integrated gene embeddings are biologically meaningful, they should better distinguish biologically realistic cell profiles from implausible synthetic ones.

In single-cell gene-expression, cells are described by the number of copies of mRNA molecules for each gene (Heumos et al., 2023). Given a cell profile c, we construct a vector cell representation $z_c \in \mathbb{R}^d$, and a matrix representation $Z_c \in \mathbb{R}^{d \times |\mathcal{G}_c|}$. The norm of z_c defines the coherence score ρ_c, and the number of dominant gene programs D_c is extracted from Z_c, resembling the first and second moment scores over gene space (Section 3.6, Algorithm 4, Algorithm 3). ρ_c is interpreted as an alignment measure of the expressed genes' embeddings, with the assumption that in contrast to real cells, an unrealistic cell profile will fail to express a coherent et of genes, and will express many different gene programs of similar importance, implying low ρ_c and large D_c, respectively. These scores (ρ_c and D_c) are used to assess biological plausibility of expression profiles.

Using a reference set of 20,000 real profiles of brain and blood cells from the Siletti adult-human-brain (Siletti et al., 2023) and Tabula Sapiens' blood atlases (Tabula Sapiens Consortium, 2022), respectively (Table S1), we constructed three types of implausible biological cell profiles: *shuffle*, per-cell gene permutation, *marginal* counts sampled from empirical distribution of real cells and *chimera* interleaving gene-expression from one blood cell and one brain cell at random (Section 3.8). The former two preserve per-cell or per-gene statistical distributions but destroy biological relations, and should be easily distinguished from real cell profiles. The *chimera* group preserves co-variations of some genes, but presumably creates biologically unrealistic combinations, posing a more difficult discrimination task.

UMAP visualization of the cell representation, z_c, shows that blood and brain cell profile form distinct clusters, chimeras occupy regions neighboring both cell types, while shuffle and marginal cells form distinct isolated clusters (Figure 3 for PLANTAGENET VI Procrustes, Figure S9 for PLANTAGENET VI concatenation, and Figure S10 for individual models). Shuffle cells display the lowest ρ_c and highest D_c while real cells display the highest ρ_c and lowest D_c. Marginal cell profiles also displayed distinctively lower ρ_c and higher D_c than real ones. Finally, chimeras show the most minor shifts. The distribution of ρ_c and D_c across cells reveals the same structure (Figure S11).

Quantifying the separation of real and synthetic cell profiles with AUROC aligns and strengthens the visual evidence (Figure 3, Figure S9, and Figure S10). We separate each corruption type from the blood and brain cells by applying a threshold τ on either the coherence score ρ_c (reject if $\rho_c < \tau$) or the number of dominant directions D_c (reject if $D_c > \tau$) and report the AUROC (Table 2). The PLANTAGENET VI Procrustes space outperforms the PLANTAGENET VI concat space and all individual models on all corruption types. From the individual models, CellPLM performs the best overall (Table 2), as shown with UMAP visualization (Figure S10). The order of complexity across corrupted-cell types is again preserved, with shuffle being trivial and chimera being most challenging.

The PLANTAGENET consensus space remains robust on average, yet different individual models perform better on tissue-specific experiments (Table S4). On blood, Procrustes attains the best mean AUROC (0.848). On brain, although AlphaGenome leads (0.791), the two consensus spaces still

	Shuffle		Marginal		Chimera		
	ρ_c	D_c	ρ_c	D_c	ρ_c	D_c	mean
AlphaGenome	**1.000**	0.999	0.550	0.492	0.580	0.570	0.699
ESM-2	**1.000**	0.994	0.941	<u>0.900</u>	0.607	0.508	0.825
GenePT-protein	0.999	0.939	0.615	0.605	0.722	0.648	0.755
DepMap CRISPR	0.971	0.192	0.770	0.782	0.719	0.561	0.666
scGPT	**1.000**	**1.000**	0.912	0.871	0.638	0.595	0.836
CellPLM	**1.000**	0.704	<u>0.980</u>	0.851	<u>0.803</u>	0.745	0.847
PLANTAGENET VI (concat)	**1.000**	0.991	0.961	0.870	0.791	<u>0.764</u>	<u>0.896</u>
PLANTAGENET VI (Procrustes)	**1.000**	**1.000**	**0.992**	**0.968**	**0.845**	**0.824**	**0.938**

Table 2: Corrupted-cell analysis for blood and brain cells. A cell is rejected if its coherence ρ_c (D_c) are below (above) a threshold. AUROC is obtained by varying the threshold, separating 20,000 real reference cells from 20,000 corrupted cells. Best model is in **bold**, second best is <u>underlined</u>.

rank second and third (concat 0.699, Procrustes 0.684). Thus, even on tissue-specific experiments, the consensus remains competitive and appealing for consistent performance, also captured by its ability to best represent, broader, cross-tissue biology; performing best in the combined blood and brain setup. Overall, we conclude that consensus gene embeddings not only represent gene properties better, but also enable building better representations of higher order objects, such as cells.

5 DISCUSSION

We empirically support the Platonic representation hypothesis in biology for gene representations across modalities, and create a joint PLANTAGENET consensus representation that captures gene properties better than individual embeddings. We further demonstrate that the PLANTAGENET consensus is valuable for representing single-cell gene-expression profiles by evaluating biological plausibility of cellular states. In the near term, such a representation could provide a discriminator for generative models of cells, distinguishing biologically coherent predictions from implausible combinations of cellular features, with utility for both biological discovery and cell engineering, or enable practical applications such as detecting chimeric cell states and improving doublet identification.

These applications point toward a more consequential possibility. As representations incorporate increasingly diverse views of genes and cells, including sequence, expression, regulation, perturbation, and phenotype, their shared structure may increasingly reflect modality-invariant constraints that govern which cellular states are biologically possible. An increasingly comprehensive "Platonic" representation could therefore provide not only a metric for whether a generated cell resembles observed biology, but a framework for exploring cellular states beyond those represented in existing datasets. Such models could ultimately help distinguish states that are merely unobserved from those that are biologically implausible, enabling generative models to propose novel but biologically coherent cell states, including natural states that have not yet been measured, or, eventually, the perturbations required to realize desired states experimentally.

For the broader machine learning community, our results show that looking beyond commonalities into modality-specific information, here captured as complementary biological views, can be useful. A consensus embedding can provide a representation superior to any of the individual parts. We suggest that "Platonic representation," currently used to express the convergence (overlap) between different spaces, should stand for a consensus ("holistic") representation of underlying objects, more accurately capturing true Platonic forms, and preserving distinct world views of different modalities.

Limitations of our work include a non exhaustive set of gene-embedding modalities and models within each modality. Extending these would likely not affect the overall conclusions, but may further improve the PLANTAGENET consensus representation. Another direction of follow-up work could include alternative approaches for consensus embedding construction, including theoretical justifications, and biological interpretation.

AI USE STATEMENT

In this work, we used generative AI tools to polish writing, for retrieval and discovery (identifying potential related work), method execution (assistance in code implementation under strict supervision and approval). We have not used generative AI tools to help develop theoretical models or conceptual frameworks, clean and reformat dataset, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, support qualitative and thematic data analysis, interpret results. Additionally, we used generative AI tools for generating figures, creating or editing software code, creation of artifacts, and searching for information. Generating synthetic data sets, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs are not applicable to this work. We have carefully reviewed all AI-assisted work; we verified all suggested implementations. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C. Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. Multi-Omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6): e8124, 2018. doi: 10.15252/msb.20178124. URL https://doi.org/10.15252/msb.20178124.

Žiga Avsec, Natasha Latysheva, Jun Cheng, Guido Novati, Kyle R. Taylor, Tom Ward, Clare Bycroft, Lauren Nicolaisen, Eirini Arvaniti, Joshua Pan, Raina Thomas, Vincent Dutordoir, Matteo Perino, Soham De, Alexander Karollus, Adam Gayoso, Toby Sargeant, Anne Mottram, Lai Hong Wong, Pavol Drotár, Adam Kosiorek, Andrew Senior, Richard Tanburn, Taylor Applebaum, Souradeep Basu, Demis Hassabis, and Pushmeet Kohli. Advancing regulatory variant effect prediction with AlphaGenome. *Nature*, 649(8099):1206–1218, 2026. doi: 10.1038/s41586-025-10014-0. URL https://doi.org/10.1038/s41586-025-10014-0.

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B Burkhardt, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25): 7045–7063, 2024. doi: 10.1016/j.cell.2024.11.015. URL https://doi.org/10.1016/j.cell.2024.11.015.

Yiqun Chen and James Zou. Simple and effective embedding model for single-cell biology built from ChatGPT. *Nature Biomedical Engineering*, 9(4):483–493, 2025. doi: 10.1038/s41551-024-01284-6. URL https://doi.org/10.1038/s41551-024-01284-6.

David Croft, Gavin O'Kelly, Guanming Wu, Robin Haw, Marc Gillespie, Lisa Matthews, Michael Caudy, Phani Garapati, Gopal Gopinath, Bijay Jassal, Steven Jupe, Irina Kalatskaya, Shahana Mahajan, Bruce May, Nelson Ndegwa, Esther Schmidt, Veronica Shamovsky, Christina Yung, Ewan Birney, Henning Hermjakob, Peter D'Eustachio, and Lincoln Stein. Reactome: A database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(Database issue): D691–D697, 2011. doi: 10.1093/nar/gkq1018. URL https://doi.org/10.1093/nar/gkq1018.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024. doi: 10.1038/s41592-024-02201-0. URL https://doi.org/10.1038/s41592-024-02201-0.

Michael J. Denton, Craig J. Marshall, and Michael Legge. The protein folds as Platonic forms: New support for the Pre-Darwinian conception of evolution by natural law. *Journal of Theoretical Biology*, 219(3):325–342, 2002. doi: 10.1006/jtbi.2002.3128. URL https://doi.org/10.1006/jtbi.2002.3128.

DepMap, Broad. DepMap 24q2 public. Figshare+ dataset, 2024. URL https://doi.org/10.25452/figshare.plus.25880521.v1.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, 2016. doi: 10.1016/j.cell.2016.11.038. URL https://doi.org/10.1016/j.cell.2016.11.038.

Oscar Franzén, Li-Ming Gan, and Johan L M Björkegren. Panglaodb: a web server for exploration of mouse and human single-cell rna sequencing data. *Database*, 2019:baz046, 01 2019. ISSN 1758-0463. doi: 10.1093/database/baz046. URL https://doi.org/10.1093/database/baz046.

Luke Funk, Kuan-Chung Su, Jimmy Ly, David Feldman, Avtar Singh, Brittania Moodie, Paul C. Blainey, and Iain M. Cheeseman. The phenotypic landscape of essential human genes. *Cell*, 185 (24):4634–4653.e22, 2022. doi: 10.1016/j.cell.2022.10.017. URL https://doi.org/10.1016/j.cell.2022.10.017.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the Platonic Representation Hypothesis: An Aristotelian view. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.14486. ICML 2026 camera-ready; arXiv:2602.14486.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855–864. ACM, 2016. doi: 10.1145/2939672.2939754. URL https://doi.org/10.1145/2939672.2939754.

Lukas Heumos, Anna C. Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D. Lücken, Daniel C. Strobl, Juan Henao, Fabiola Curion, Single-cell Best Practices Consortium, Herbert B. Schiller, and Fabian J. Theis. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 24(8):550–572, 2023. doi: 10.1038/s41576-023-00586-w. URL https://doi.org/10.1038/s41576-023-00586-w.

Johannes C. Hingerl, Laura D. Martens, Alexander Karollus, Trevor Manz, Jason D. Buenrostro, Fabian J. Theis, and Julien Gagneur. scooby: Modeling multimodal genomic profiles from DNA sequence at single-cell resolution. *Nature Methods*, 22(11):2275–2285, 2025. doi: 10.1038/s41592-025-02854-5. URL https://doi.org/10.1038/s41592-025-02854-5.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The Platonic Representation Hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John X. Morris. Harnessing the universal geometry of embeddings. In *Advances in Neural Information Processing Systems*, volume 38, pp. 45963–45987. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/4175dee33d6145cb8f0323703d138a53-Abstract-Conference.html.

Kyrylo Kalashnikov. A case for Platonic biology: And implications of Platonism in drug discovery. Kyrylo's Substack, 2026. URL https://kyrylok.substack.com/p/a-case-for-platonic-biology.

Yoav Kan-Tor, Michael Morris Danziger, Eden Zohar, Matan Ninio, and Yishai Shimoni. Does your model understand genes? a benchmark of gene properties for biological and text models, 2024. URL https://arxiv.org/abs/2412.04075.

Takamasa Kudo, Romain Lopez, Ana M. Meireles, Antonio Rios, Paula Coelho, Vineethkrishna Chandrasekar, Dylan C. Lam, Jan-Christian Huetter, Mineto Ota, Orit Rozenblatt-Rosen, Levi Garraway, Kathryn Geiger-Shuller, Avtar Singh, Jonathan K. Pritchard, and Aviv Regev. Scalable multimodal mapping of macrophage regulatory architecture by integrating optical and transcriptomic pooled screens. *bioRxiv*, 2026. doi: 10.64898/2026.05.27.728345. URL https://doi.org/10.64898/2026.05.27.728345.

Avantika Lal, Alexander Karollus, Laura Gunsalus, David Garfield, Surag Nair, Alex M. Tseng, M. Grace Gordon, John Blischak, Bryce Van De Geijn, Tushar Bhangale, Jenna L. Collier, Nathaniel Diamant, Tommaso Biancalani, Hector Corrada Bravo, Gabriele Scalia, and Gokcen Eraslan. Decoding sequence determinants of gene expression in diverse cellular and disease states. *Nature Methods*, 23(6):1138–1151, 2026. doi: 10.1038/s41592-026-03102-0. URL https://doi.org/10.1038/s41592-026-03102-0.

Michael Levin. Ingressing minds: Causal, non-physical patterns in-form natural, synthetic, and hybrid embodiments. *Philosophies*, 11(5):161, 2026. doi: 10.3390/philosophies11050161. URL https://www.mdpi.com/2409-9287/11/5/161.

Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6): 417–425, 2015. doi: 10.1016/j.cels.2015.12.004. URL https://doi.org/10.1016/j.cels.2015.12.004.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://doi.org/10.1126/science.ade2574.

Johannes Linder, Divyanshi Srivastava, Han Yuan, Vikram Agarwal, and David R. Kelley. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. *Nature Genetics*, 57(4):949–961, 2025. doi: 10.1038/s41588-024-02053-6. URL https://doi.org/10.1038/s41588-024-02053-6.

Russell Littman, Jacob Levine, Sepideh Maleki, Yongju Lee, Vladimir Ermakov, Lin Qiu, Alexander Wu, Kexin Huang, Romain Lopez, Gabriele Scalia, Tommaso Biancalani, David Richmond, Aviv Regev, and Jan-Christian Hütter. Gene-embedding-based prediction and functional evaluation of perturbation expression responses with PRESAGE. *bioRxiv*, 2025. doi: 10.1101/2025.06.03.657653. URL https://doi.org/10.1101/2025.06.03.657653.

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

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409, 2022. doi: 10.1093/bib/bbac409. URL https://doi.org/10.1093/bib/bbac409.

Pablo Marcos-Manchón, Rishi Jha, and Lluís Fuentemilla. Platonic representations in the human brain: Unsupervised recovery of universal geometry, 2026. URL https://arxiv.org/abs/2605.20496.

Jonathan M Mudge, Sílvia Carbonell-Sala, Mark Diekhans, Jose Gonzalez Martinez, Toby Hunt, Irwin Jungreis, Jane E Loveland, Carme Arnan, If Barnes, Ruth Bennett, Andrew Berry, Alexandra Bignell, Daniel Cerdán-Vélez, Kelly Cochran, Lucas T Cortés, Claire Davidson, Sarah Donaldson, Cagatay Dursun, Reham Fatima, Matthew Hardy, Prajna Hebbar, Zoe Hollis, Benjamin T James, Yunzhe Jiang, Rory Johnson, Gazaldeep Kaur, Mike Kay, Riley J Mangan, Miguel Maquedano, Laura Martínez Gómez, Nourhen Mathlouthi, Ryan Merritt, Pengyu Ni, Emilio Palumbo, Tamara Perteghella, Fernando Pozo, Shriya Raj, Cristina Sisu, Emily Steed, Dulika Sumathipala, Marie-Marthe Suner, Barbara Uszczynska-Ratajczak, Elizabeth Wass, Yucheng T Yang, Dingyao Zhang, Robert D Finn, Mark Gerstein, Roderic Guigó, Tim J P Hubbard, Manolis Kellis, Anshul Kundaje, Benedict Paten, Michael L Tress, Ewan Birney, Fergal J Martin, and Adam Frankish. GENCODE 2025: Reference gene annotation for human and mouse. *Nucleic Acids Research*, 53(D1):D966–D975, 2025. doi: 10.1093/nar/gkae1078. URL https://doi.org/10.1093/nar/gkae1078.

James D. Pearce, Sara E. Simmonds, Gita Mahmoudabadi, Lakshmi Krishnan, Giovanni Palla, Ana-Maria Istrate, Alexander Tarashansky, Benjamin Nelson, Omar Valenzuela, Donghui Li,

Stephen R. Quake, and Theofanis Karaletsos. TranscriptFormer: A generative cell atlas across 1.5 billion years of evolution. *Science*, 393(6806):aec8514, 2026. doi: 10.1126/science.aec8514. URL https://doi.org/10.1126/science.aec8514.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720. URL https://doi.org/10.1080/14786440109462720.

Plato. *Republic*. c. 375 BC.

Meraj Ramezani, Erin Weisbart, Julia Bauman, Avtar Singh, John Yong, Maria Lozada, Gregory P. Way, Sanam L. Kavari, Celeste Diaz, Eddy Leardini, Gunjan Jetley, Jenlu Pagnotta, Marzieh Haghighi, Thiago M. Batista, Joaquín Pérez-Schindler, Melina Claussnitzer, Shantanu Singh, Beth A. Cimini, Paul C. Blainey, Anne E. Carpenter, Calvin H. Jan, and James T. Neal. A genome-wide atlas of human cell morphology. *Nature Methods*, 22(3):621–633, 2025. doi: 10.1038/s41592-024-02537-7. URL https://doi.org/10.1038/s41592-024-02537-7.

Jennifer E Rood, Anna Hupalowska, and Aviv Regev. Toward a foundation model of causal cell and tissue biology with a perturbation cell and tissue atlas. *Cell*, 187(17):4520–4545, 2024. doi: 10.1016/j.cell.2024.07.035. URL https://doi.org/10.1016/j.cell.2024.07.035.

Kimberly Siletti, Rebecca Hodge, Alejandro Mossi Albiach, Ka Wai Lee, Song-Lin Ding, Lijuan Hu, Peter Lönnerberg, Trygve Bakken, Tamara Casper, Michael Clark, Nick Dee, Jessica Gloe, Daniel Hirschstein, Nadiya V. Shapovalova, C. Dirk Keene, Julie Nyhus, Herman Tung, Anna Marie Yanny, Ernest Arenas, Ed S. Lein, and Sten Linnarsson. Transcriptomic diversity of cell types across the adult human brain. *Science*, 382(6667):eadd7046, 2023. doi: 10.1126/science.add7046. URL https://doi.org/10.1126/science.add7046.

Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L. Gable, Tao Fang, Nadezhda T. Doncheva, Sampo Pyysalo, Peer Bork, Lars J. Jensen, and Christian von Mering. The STRING database in 2023: Protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, 2023. doi: 10.1093/nar/gkac1000. URL https://doi.org/10.1093/nar/gkac1000.

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

Alejandro Tejada-Lapuerta, Anna C. Schaar, Robert Gutgesell, Giovanni Palla, Lennard Halle, Mariia Minaeva, Larsen Vornholz, Leander Dony, Francesca Drummer, Till Richter, Mojtaba Bahrami, and Fabian J. Theis. Nicheformer: A foundation model for single-cell and spatial omics. *Nature Methods*, 22(12):2525–2538, 2025. doi: 10.1038/s41592-025-02814-z. URL https://doi.org/10.1038/s41592-025-02814-z.

The Gene Ontology Consortium. The Gene Ontology knowledgebase in 2026. *Nucleic Acids Research*, 54(D1):D1779–D1792, 2026. doi: 10.1093/nar/gkaf1292. URL https://doi.org/10.1093/nar/gkaf1292.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023. doi: 10.1038/s41586-023-06139-9. URL https://doi.org/10.1038/s41586-023-06139-9.

Hongzhi Wen, Wenzhuo Tang, Xinnan Dai, Jiayuan Ding, Wei Jin, Yuying Xie, and Jiliang Tang. CellPLM: Pre-training of cell language model beyond single cells. In *The Twelfth International Conference on Learning Representations*, pp. 5649–5673, 2024. URL https://openreview.net/forum?id=BKXvPDekud.

Wei Wu, Can Liao, Zizhen Deng, Zhengrui Guo, and Jinzhuo Wang. Neuron Platonic intrinsic representation from dynamics using contrastive learning. In *The*

Thirteenth International Conference on Learning Representations, pp. 39826–39842, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/63403963ebabae8b5cc05e14b5761c4c-Abstract-Conference.html.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana V. Davuluri, and Han Liu. DNABERT-2: Efficient foundation model and benchmark for multi-species genomes. In *The Twelfth International Conference on Learning Representations*, pp. 41642–41665, 2024. URL https://openreview.net/forum?id=oMLQB4EZE1.

A SUPPLEMENTARY MODEL DETAILS AND ANALYSES TABLES

Table S1: Gene embedding representations. The first 31 entries are non-integrated representations; the final two are integrated PRESAGE representations.

Embedding	# genes	Dim.	Data or model input	Gene-level representation
Natural-language representations				*3 representations*
GenePT-protein (Chen & Zou, 2025)	19,091	3,072	UniProt protein-product summaries	Text-embedding-3-large representation of the protein summary associated with each gene
GenePT-ada (Chen & Zou, 2025)	19,091	1,536	NCBI gene summaries	Text-embedding-ada-002 representation of the textual summary associated with each gene
BioGPT (Luo et al., 2022)	9,675	1,600	Biomedical literature	BioGPT gene embeddings distributed with PRESAGE (Littman et al., 2025); the precise extraction procedure was not reported
Protein-sequence representations				*1 representations*
ESM-2 (Lin et al., 2023)	19,204	5,120	Protein sequences	Mean-pooled residue representation of each gene's protein product
DNA-sequence representations				*5 representations*
AlphaGenome (Avsec et al., 2026)	20,020	3,072	DNA sequence and multimodal genomic tracks	Mean-pooled encoder representation from a 524,288-bp sequence centered on each transcription start site
Borzoi (Linder et al., 2025)	20,020	1,920	DNA sequence and RNA-seq coverage tracks	Final hidden representation before the species-specific track heads, mean-pooled across genomic bins
SCOOBY (Hingerl et al., 2025)	20,020	1,920	DNA sequence with single-cell RNA and chromatin-accessibility profiles	Conv1D refinement-layer representation without cell conditioning, mean-pooled across genomic bins
Decima (Lal et al., 2026)	16,024	8,856	DNA sequence and expression profiles from approximately 22 million single cells or nuclei	Predicted expression of each gene across 8,856 cell-type pseudobulks
DNABERT-2 (Zhou et al., 2024)	20,020	768	Multi-species DNA sequences	Final-layer representation of a 1,024-bp sequence centered on each transcription start site
Pathway and interaction-network representations				*4 representations*
STRING PPI	2,681	128	STRING protein-association network (Szklarczyk et al., 2023)	node2vec (Grover & Leskovec, 2016) gene representation followed by projection onto 128 principal components; distributed with PRESAGE (Littman et al., 2025)
Reactome	2,100	128	Reactome pathway graph (Croft et al., 2011)	node2vec (Grover & Leskovec, 2016) gene representation followed by projection onto 128 principal components; distributed with PRESAGE (Littman et al., 2025)
GO-BP	2,694	128	Gene Ontology biological-process graph (The Gene Ontology Consortium, 2026)	node2vec (Grover & Leskovec, 2016) gene representation followed by projection onto 128 principal components; distributed with PRESAGE (Littman et al., 2025)
Hallmark	1,106	128	MSigDB Hallmark gene-set graph (Liberzon et al., 2015)	node2vec (Grover & Leskovec, 2016) gene representation followed by projection onto 128 principal components; distributed with PRESAGE (Littman et al., 2025)
Cell-viability perturbation representations				*1 representations*
DepMap CRISPR	18,426	128	DepMap 24Q2 CRISPR gene-effect profiles (DepMap, Broad, 2024)	Gene loadings from a 128-component PCA of cell-line viability profiles; distributed with PRESAGE (Littman et al., 2025)
Atlas-derived transcriptomic representations				*8 representations*

Continued on next page

Table S1 — continued

Embedding	# genes	Dim.	Data or model input	Gene-level representation
PCA (brain) (Pearson, 1901)	17,970	128	Siletti adult-human-brain atlas (Siletti et al., 2023)	Gene loadings on the first 128 principal components
PCA (blood) (Pearson, 1901)	18,595	128	Tabula Sapiens blood atlas (Tabula Sapiens Consortium, 2022)	Gene loadings on the first 128 principal components
scVI-encoder (brain) (Lopez et al., 2018)	2,338	128	Siletti adult-human-brain atlas (Siletti et al., 2023)	Gene weights from the first scVI encoder layer
scVI-decoder (brain) (Lopez et al., 2018)	2,338	128	Siletti adult-human-brain atlas (Siletti et al., 2023)	Gene weights from the final scVI decoder layer
scVI-encoder (blood) (Lopez et al., 2018)	2,151	128	Tabula Sapiens blood atlas (Tabula Sapiens Consortium, 2022)	Gene weights from the first scVI encoder layer
scVI-decoder (blood) (Lopez et al., 2018)	2,151	128	Tabula Sapiens blood atlas (Tabula Sapiens Consortium, 2022)	Gene weights from the final scVI decoder layer
MOFA (brain) (Argelaguet et al., 2018)	2,450	30	Siletti adult-human-brain atlas (Siletti et al., 2023)	Gene factor loadings learned by MOFA
MOFA (blood) (Argelaguet et al., 2018)	2,444	30	Tabula Sapiens blood atlas (Tabula Sapiens Consortium, 2022)	Gene factor loadings learned by MOFA
Single-cell and spatial foundation-model representations				*5 representations*
scGPT (Cui et al., 2024)	19,201	512	Multi-tissue human single-cell transcriptomes	Learned gene-embedding lookup table
TranscriptFormer (Pearce et al., 2026)	11,086	248	Multi-species, multi-tissue single-cell transcriptomes	Contextual gene representations mean-pooled over a fixed PBMC panel
Geneformer (Theodoris et al., 2023)	20,007	768	Multi-tissue human single-cell transcriptomes	Learned gene-embedding lookup table
Nicheformer (Tejada-Lapuerta et al., 2025)	19,290	512	Dissociated single-cell and spatial transcriptomes	Learned gene-embedding lookup table
CellPLM (Wen et al., 2024)	19,325	1,024	Multi-tissue single-cell and spatial transcriptomes	Learned gene-embedding lookup table
Imaging-perturbation representations				*4 representations*
OPS (Funk) (Funk et al., 2022)	4,950	312	Optical pooled CRISPR screen of essential genes in HeLa cells	CellProfiler morphology features aggregated for each knocked-out gene
CellPaint HeLa–HPLM (Ramezani et al., 2025)	18,479	1,705	Genome-wide Cell Painting CRISPR screen in HeLa cells cultured in HPLM	Image-derived morphology features aggregated for each knocked-out gene
CellPaint HeLa–DMEM (Ramezani et al., 2025)	18,479	1,597	Genome-wide Cell Painting CRISPR screen in HeLa cells cultured in DMEM	Image-derived morphology features aggregated for each knocked-out gene
CellPaint A549 (Ramezani et al., 2025)	18,479	1,520	Genome-wide Cell Painting CRISPR screen in A549 cells	Image-derived morphology features aggregated for each knocked-out gene
Integrated representations				*2 representations*
PRESAGE knowledge (Littman et al., 2025)	2,835	8,768	ESM-2, BioGPT, GenePT-ada, STRING, GO-BP, Reactome, and Hallmark	Concatenation of the separately L2-normalized source representations
PRESAGE all (Littman et al., 2025)	2,835	9,208	PRESAGE knowledge, DepMap CRISPR, and Funk OPS	PRESAGE knowledge concatenated with the normalized perturbation representations

Table S2: Models ranked by the number of genes represented in the 20,020-gene canonical set.

Rank	Model	Genes in 20,020 canonical set
1	SCOOBY	20020
2	DNABERT-2	20020
3	Borzoi	20020
4	AlphaGenome	20020
5	Geneformer	20007
6	CellPLM	19325
7	Nicheformer	19290
8	ESM-2	19204
9	scGPT	19201
10	GenePT-protein	19091
11	GenePT-ada	19091
12	PCA (blood)	18595
13	CellPaint HeLa-HPLM	18479
14	CellPaint HeLa-DMEM	18479
15	CellPaint A549	18479
16	DepMap CRISPR	18426
17	PCA (brain)	17970
18	Decima	16024
19	TranscriptFormer	11086
20	BioGPT	9675
21	OPS (Funk)	4950
22	PRESAGE knowledge	2835
23	PRESAGE all	2835
24	GO-BP	2694
25	STRING PPI	2681
26	MOFA (brain)	2450
27	MOFA (blood)	2444
28	scVI-enc (brain)	2338
29	scVI-dec (brain)	2338
30	scVI-enc (blood)	2151
31	scVI-dec (blood)	2151
32	Reactome	2100
33	Hallmark	1106

Table S3: The gene embeddings (128 PCs reduced views for the individual models) of the 3,073 genes labeled by cell lineage using PanglaoDB (Franzén et al., 2019, Section 3.7) are 10-mean clustered and scored against the lineage labels by ARI and NMI.

	NMI	ARI
AlphaGenome	0.124	0.083
ESM-2	0.028	0.011
GenePT-protein	0.080	0.057
DepMap CRISPR	0.008	0.002
scGPT	0.130	0.065
CellPLM	0.043	0.016
PLANTAGENET VI (concat)	0.125	0.113
PLANTAGENET VI (Procrustes)	**0.156**	**0.123**

Table S4: Corrupted-cell rejection for the blood, brain, and blood and brain experiments. A cell is rejected if its coherence ρ_c is below (or its D_c is above) a threshold. The AUROC is obtained by varying the threshold, separating 20,000 real reference cells from 20,000 corrupted cells. Best model per column within each tissue block is in **bold**.

		Shuffle		Marginal		Chimera		
Tissue	Representation	ρ_c	D_c	ρ_c	D_c	ρ_c	D_c	mean
Blood	AlphaGenome	**1.000**	**1.000**	0.725	0.580	0.507	0.531	0.724
	ESM-2	**1.000**	0.989	0.935	**0.914**	0.591	0.531	0.827
	GenePT-protein	**1.000**	0.948	0.658	0.700	0.581	0.541	0.738
	DepMap CRISPR	0.967	0.164	0.738	0.770	0.559	0.516	0.619
	scGPT	**1.000**	**1.000**	0.904	0.903	0.641	**0.621**	0.845
	CellPLM	**1.000**	0.895	**0.994**	0.726	**0.725**	0.568	0.818
	PLANTAGENET VI (concat)	**1.000**	0.993	0.915	0.822	0.616	0.584	0.822
	PLANTAGENET VI (Procrustes)	**1.000**	**1.000**	0.949	0.891	0.632	0.617	**0.848**
Brain	AlphaGenome	**1.000**	0.999	**0.908**	**0.827**	0.501	0.511	**0.791**
	ESM-2	**1.000**	**1.000**	0.404	0.454	0.525	0.516	0.650
	GenePT-protein	**1.000**	0.965	0.464	0.516	0.565	0.550	0.677
	DepMap CRISPR	**1.000**	0.231	0.419	0.435	0.534	0.502	0.520
	scGPT	**1.000**	**1.000**	0.370	0.388	0.551	0.541	0.642
	CellPLM	**1.000**	0.432	0.703	0.765	**0.607**	**0.585**	0.682
	PLANTAGENET VI (concat)	**1.000**	**1.000**	0.543	0.521	0.567	0.563	0.699
	PLANTAGENET VI (Procrustes)	**1.000**	**1.000**	0.508	0.487	0.554	0.554	0.684
Blood & Brain	AlphaGenome	**1.000**	0.999	0.550	0.492	0.580	0.570	0.699
	ESM-2	**1.000**	0.994	0.941	0.900	0.607	0.508	0.825
	GenePT-protein	0.999	0.939	0.615	0.605	0.722	0.648	0.755
	DepMap CRISPR	0.971	0.192	0.770	0.782	0.719	0.561	0.666
	scGPT	**1.000**	**1.000**	0.912	0.871	0.638	0.595	0.836
	CellPLM	**1.000**	0.704	0.980	0.851	0.803	0.745	0.847
	PLANTAGENET VI (concat)	**1.000**	0.991	0.961	0.870	0.791	0.764	0.896
	PLANTAGENET VI (Procrustes)	**1.000**	**1.000**	**0.992**	**0.968**	**0.845**	**0.824**	**0.938**

B SUPPLEMENTARY FIGURES

Before evaluating the PLANTAGENET consensus space, we examined the impact of using a compacted 128 PCs view over the native gene-embedding space. The cumulative-variance spectra (Figure S1) show that 128 PCs sit past the variance elbow for all models: it captured 87% of the variance for AlphaGenome and CellPLM, 73% for GenePT-protein, and 68% for ESM-2 and 49% for scGPT. Since DepMap is natively 128-dimensional, it naturally captured all its variance. The 128-d reduction did not decrease performance on the gene-property prediction tasks, and even slightly improved it (Figure S2). DepMap performs poorly on gene-property prediction benchmark (AUROC of 0.62-0.63, Figure S2), but this is expected rather than alarming, as it is built from CRISPR viability screens and thus captures information of a different nature than needed for these gene-property prediction tasks.

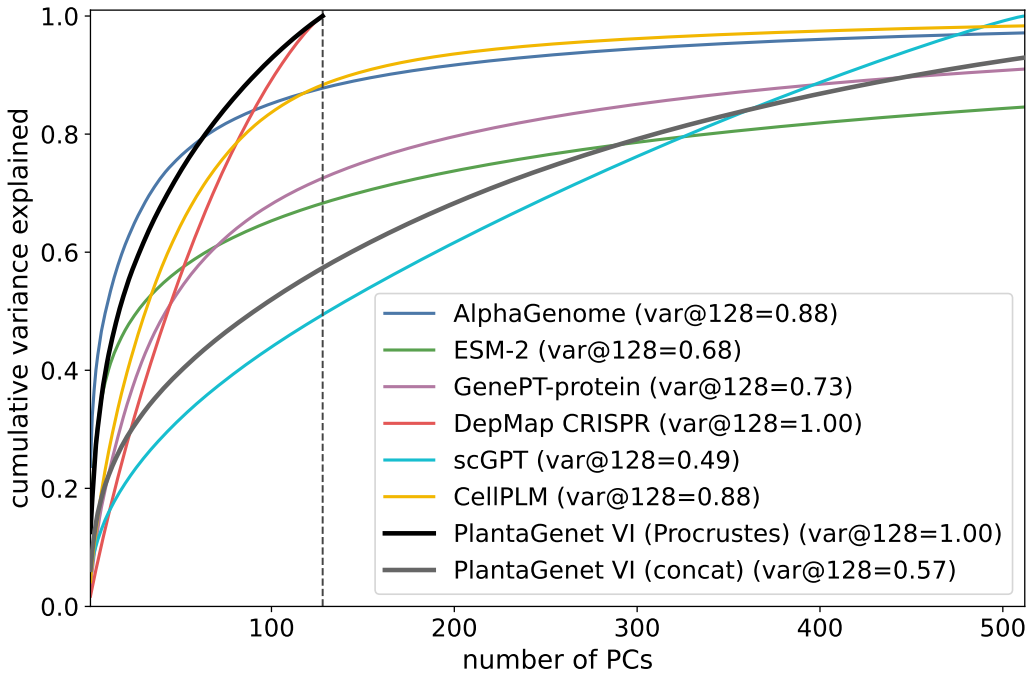


Figure S1: *Cumulative variance explained versus the number of PCs for the* PLANTAGENET VI consensus *and the individual models it is built on.*

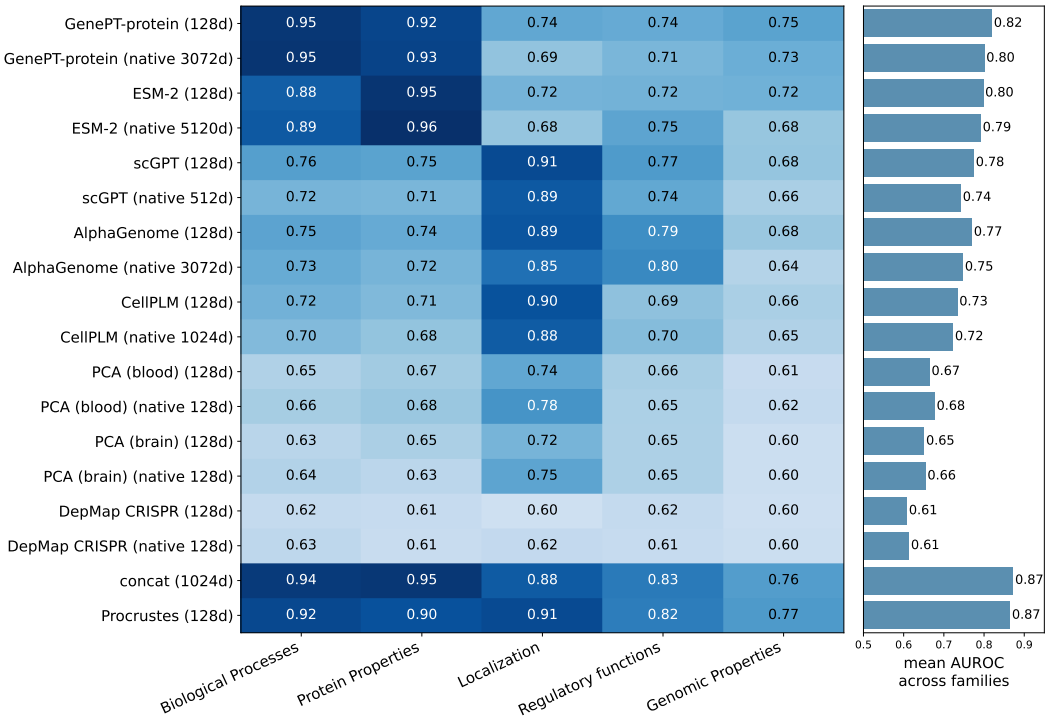


Figure S2: *Gene property prediction benchmark performance across models, using the reduced 128 PCs embeddings or the native embeddings.* The listed gene views are the ones contributing to PLANTAGENET VIII consensus. The concat and Procrustes rows refer to PLANTAGENET VIII. The entries are the AUROC (Section 3.3).

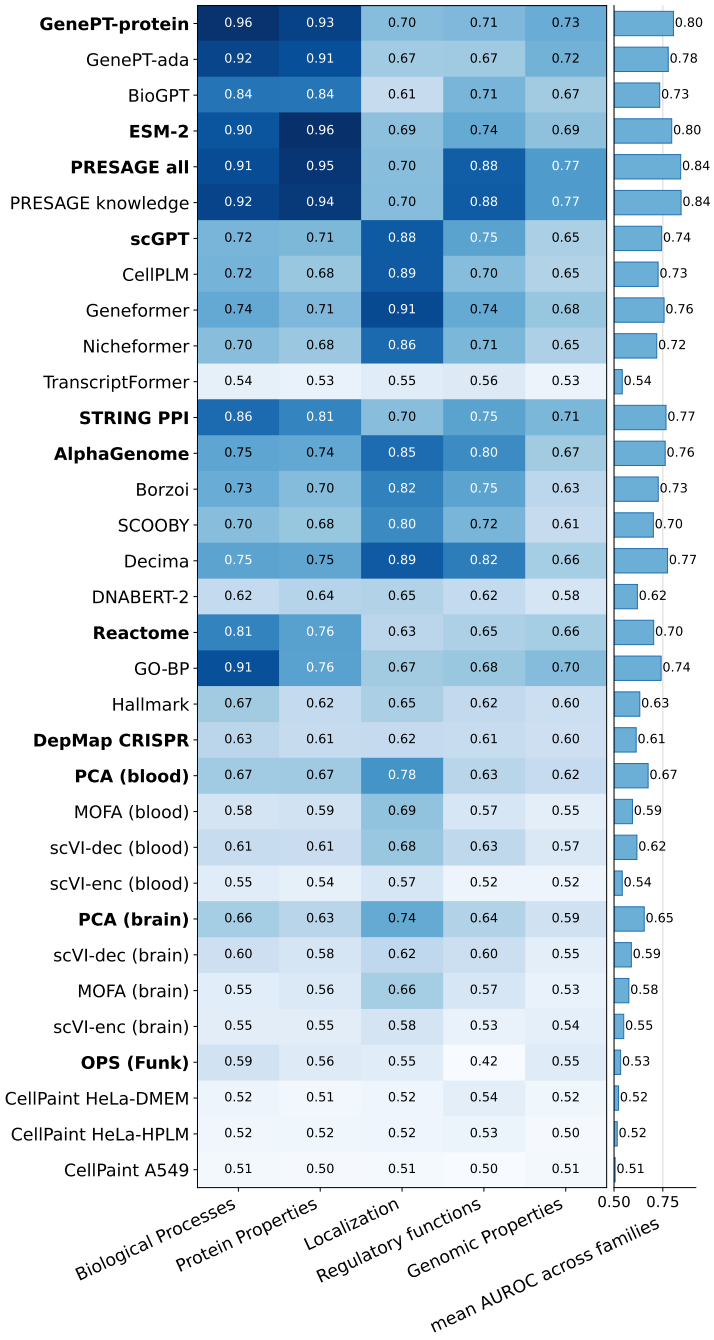


Figure S3: *Gene property prediction benchmark performance on different biological tasks across models.* The entries correspond to the AUROC (Section 3.3)

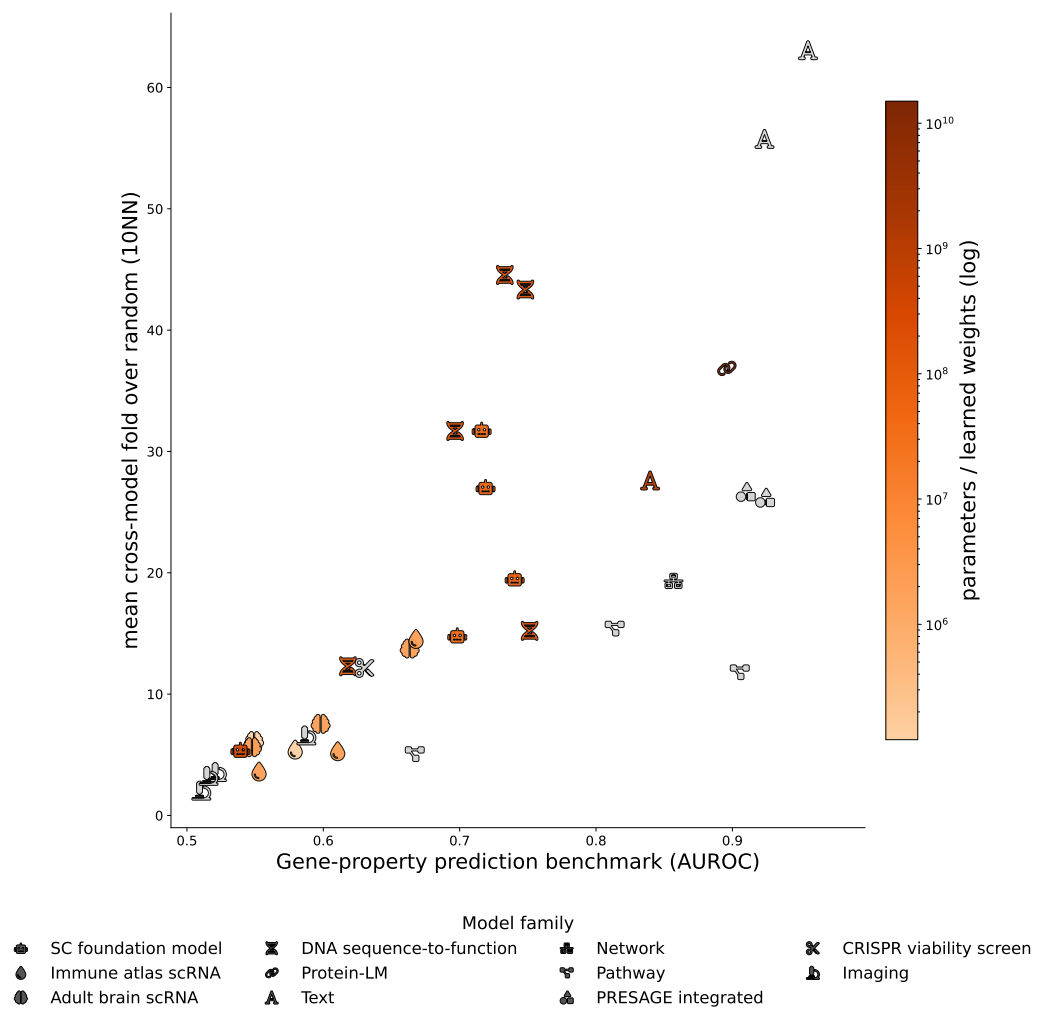


Figure S4: *Relationship between gene property prediction performance and mutual-nearest-neighbor agreement.* Color indicates the number of learnable weights in each gene embedding model. Pearson $r = 0.75$, p-value $= 6.6 \times 10^{-7}$, $n = 33$.

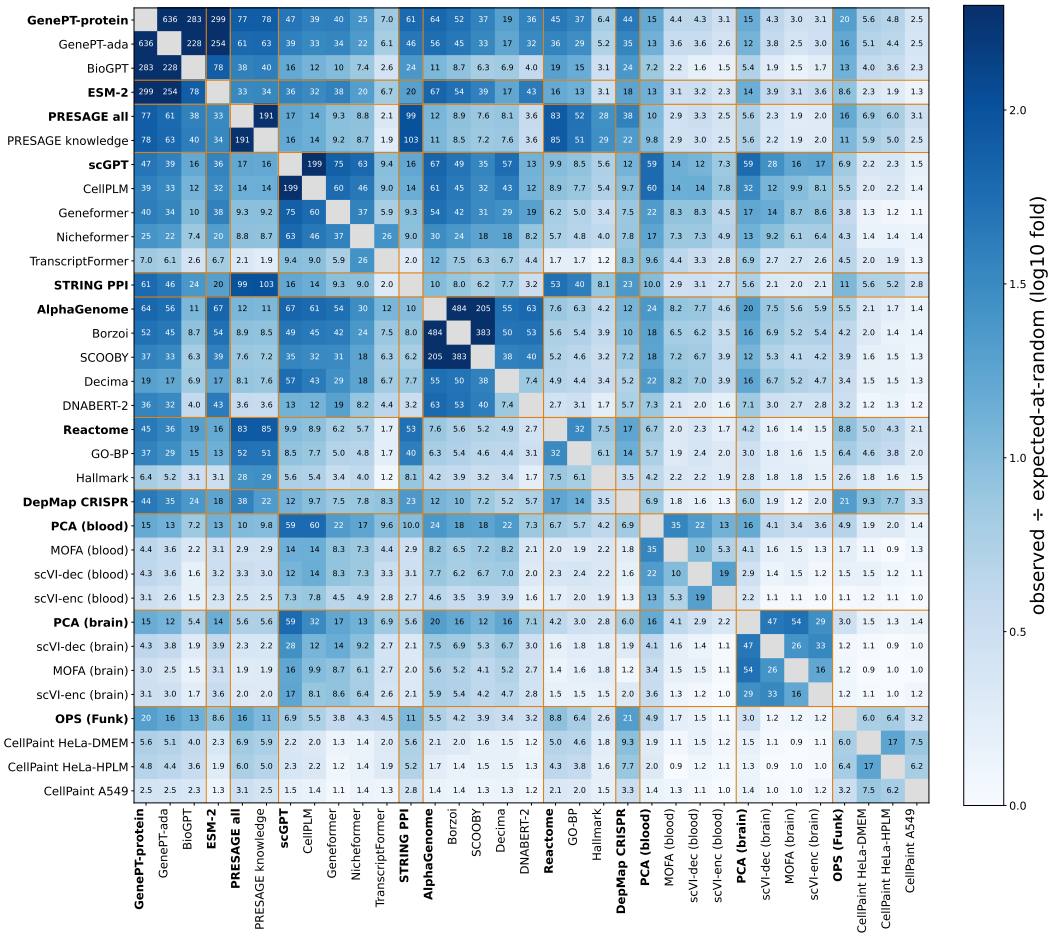


Figure S5: *Per-model mutual-nearest-neighbor agreement of gene embeddings divided by random expectation.* The entries correspond to mNN_{excess} (Section 3.2).

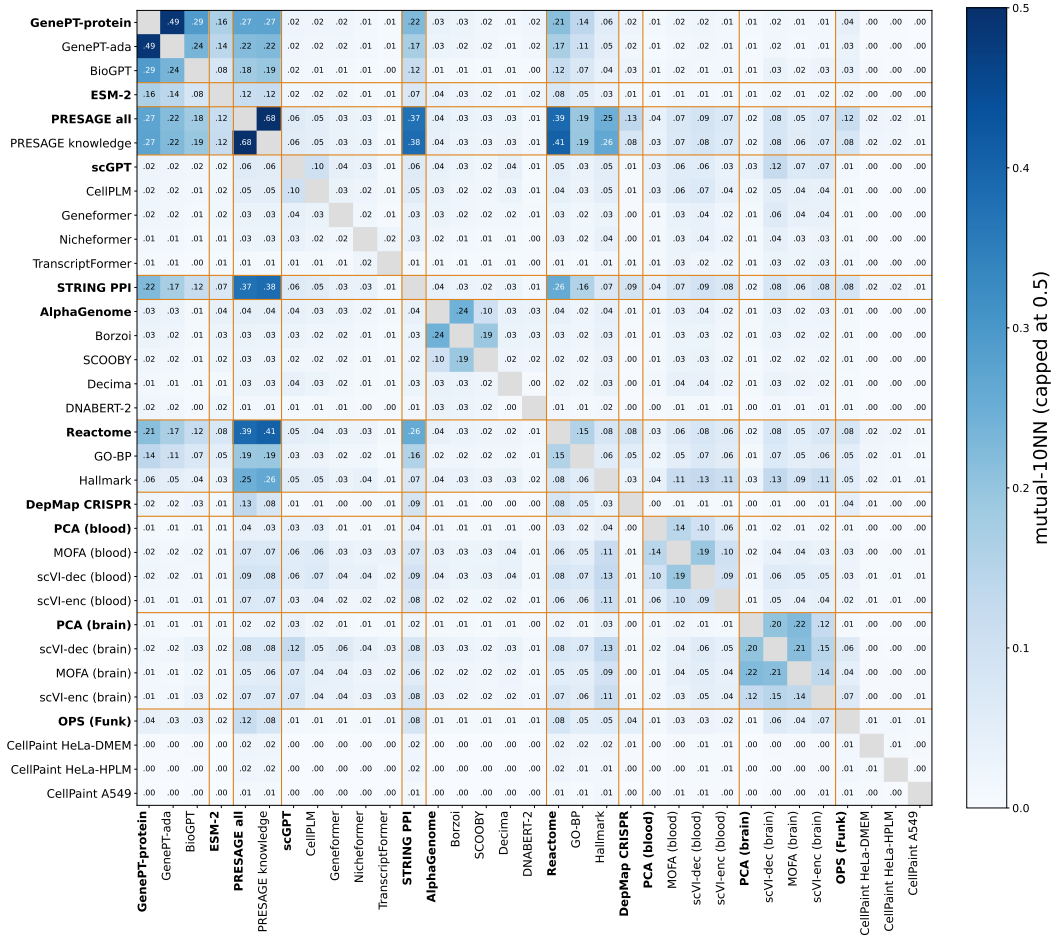


Figure S6: *Per-model mutual-nearest-neighbor agreement of gene embeddings*. Because models do not all contain the full gene set, the values are not directly comparable. The entries correspond to mNN (Section 3.2). See Figure S5 for the normalized version (mNN_{excess}).

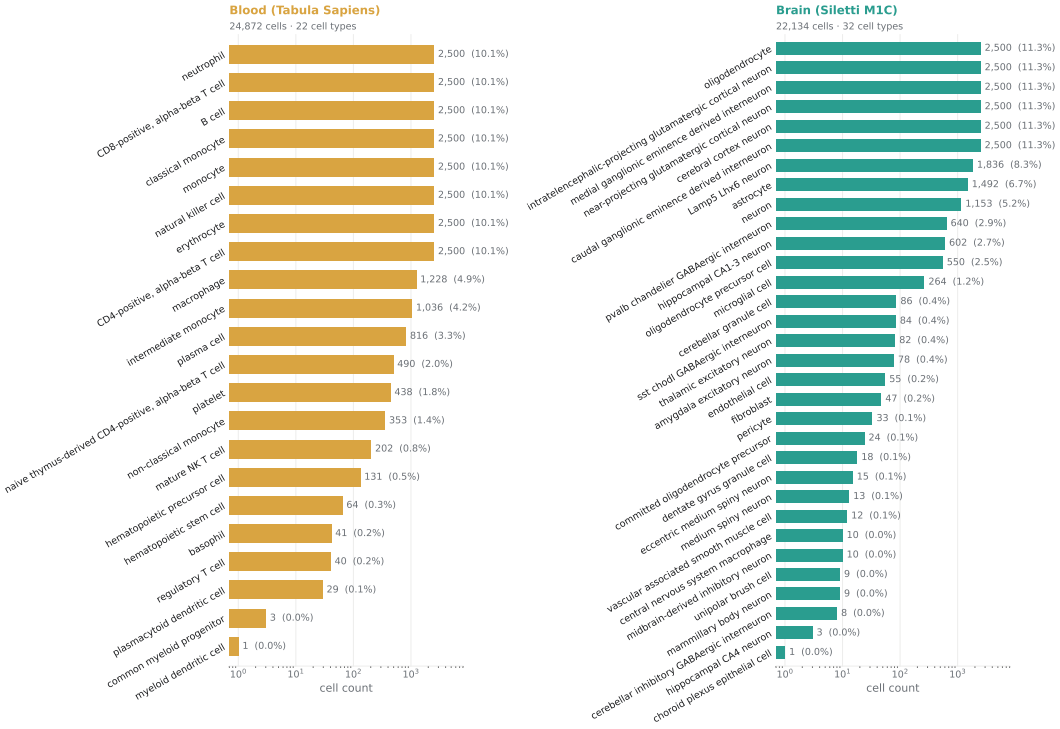


Figure S7: *Cell types distribution per tissue.* We sampled each atlas, the Siletti's adult-human-brain (Siletti et al., 2023) and Tabula Sapiens' blood cell atlases (Tabula Sapiens Consortium, 2022), with a per-cell type cap of 2,500.

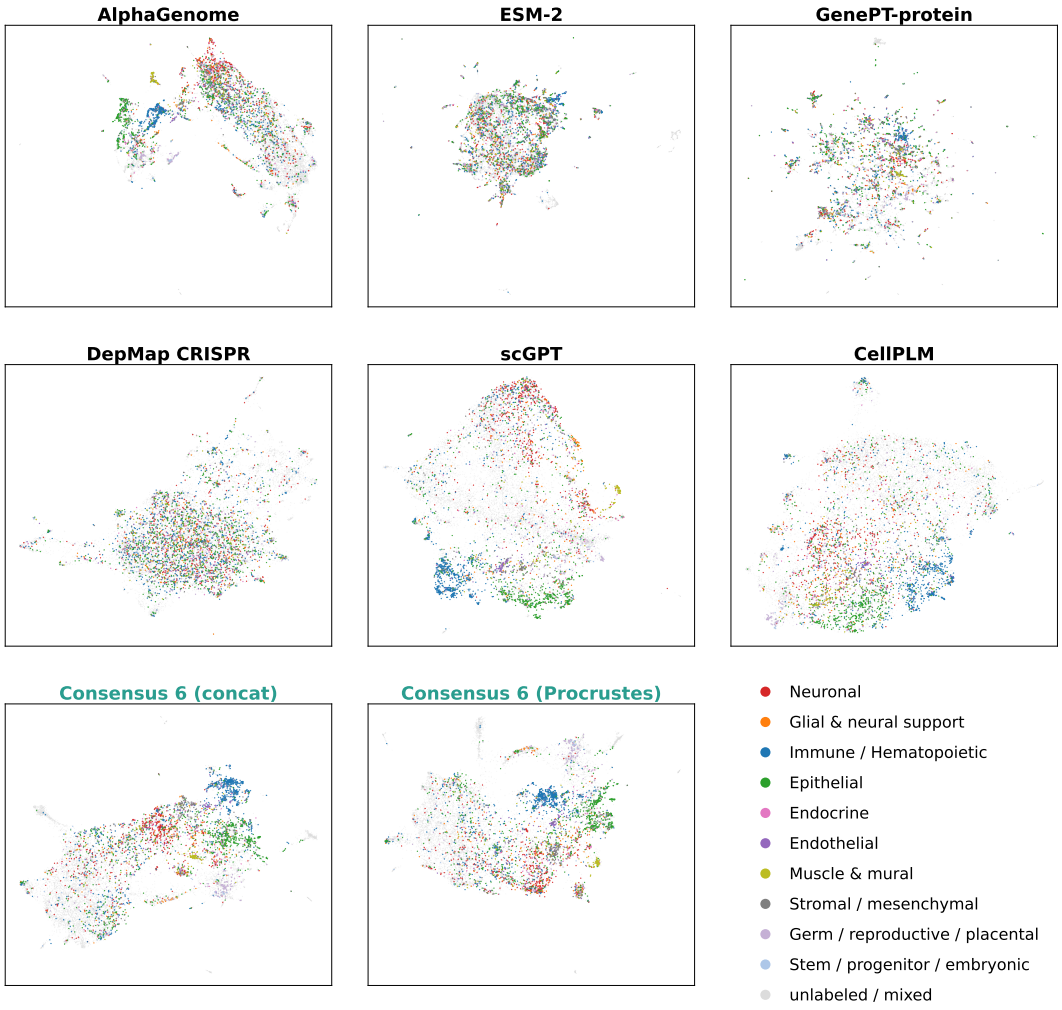


Figure S8: *Landscape of gene embeddings.* UMAP representation of the $18,217$ gene embeddings z_g from the PLANTAGENET VI consensus gene universe using the Procrustes, the concat, or the individual models' views contributing to the consensus, labeled by cell lineage using cell types marker gene sets from PanglaoDB (Franzén et al. (2019)) (Section 3.7).

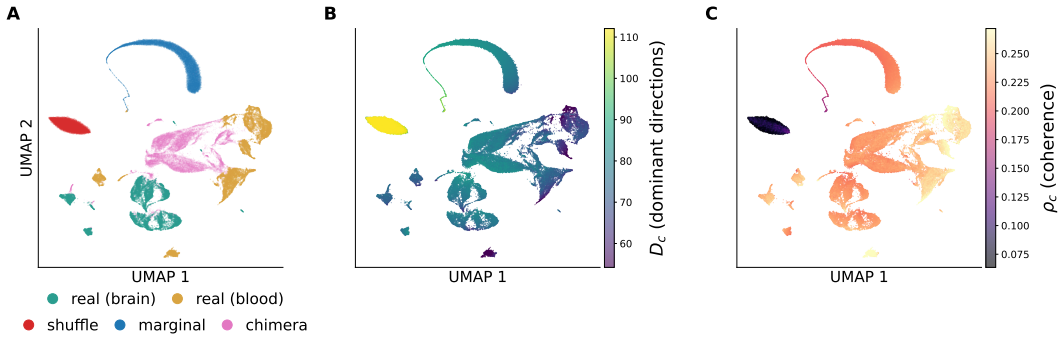


Figure S9: *Landscape of vector cell representations*. **a–c**, UMAP visualizations of vector cell representations z_c obtained using PLANTAGENET VI (Concat), comprising 20,000 real cells from blood and brain and 20,000 cells from each of the shuffle, marginal, and chimera groups. Cells are colored by group (**a**), number of dominant directions D_c (**b**), and coherence score ρ_c (**c**).

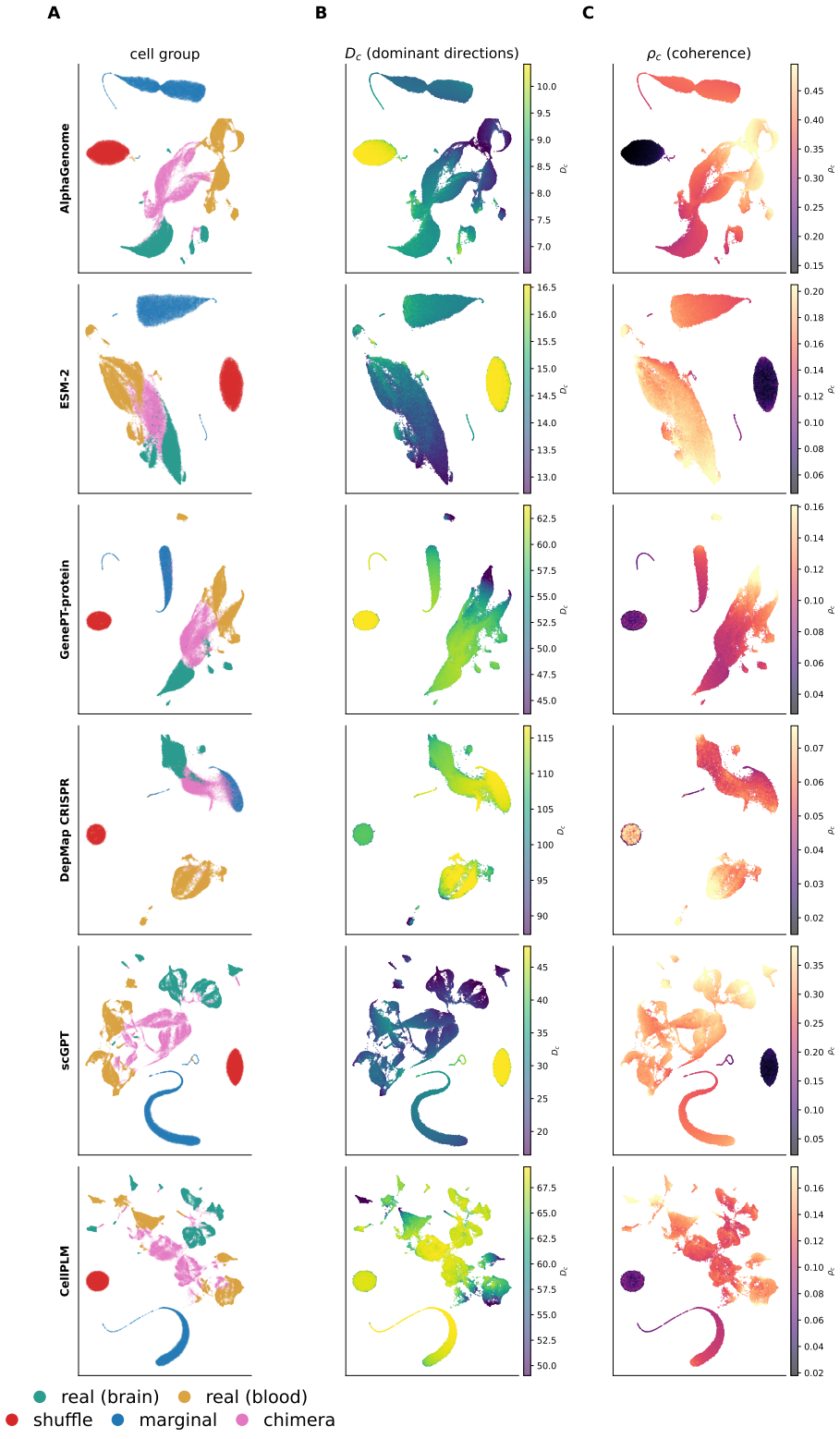


Figure S10: *Landscape of vector cell representations.* UMAP visualizations of vector cell representations z_c obtained using the individual models contributing to PLANTAGENET VI, comprising 20,000 real cells from blood and brain and 20,000 cells from each of the shuffle, marginal, and chimera groups. Cells are colored by group (left), number of dominant directions D_c (middle), and coherence score ρ_c (right).

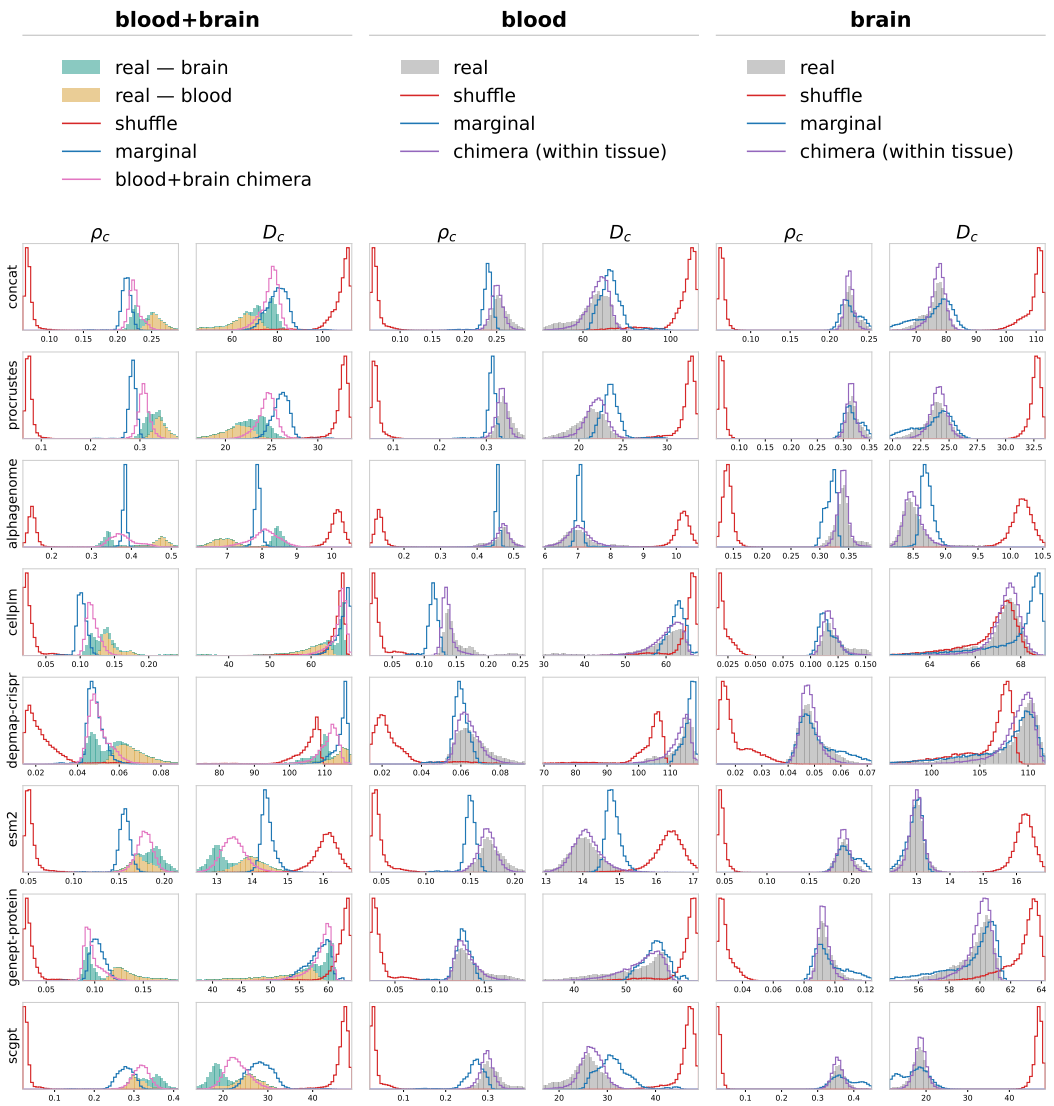


Figure S11: *Distribution of coherence ρ_c scores and number of dominant directions D_c in real cells and corrupted cells*. For each of the blood, brain, or blood and brain experiment, there are 20,000 real cells and 20,000 cells from each of the shuffle, marginal and chimera groups.