Bioinformatics, 2025, **41(11)**, btaf595
https://doi.org/10.1093/bioinformatics/btaf595
Advance Access Publication Date: 30 October 2025
Review

Systems biology

The cell as a token: high-dimensional geometry in language models and cell embeddings

William Gilpin[1,2,*]

[1]Department of Physics, The University of Texas at Austin, Austin, TX 78712, United States
[2]Medici Therapeutics, Boston, MA 02114, United States

*Corresponding author: Department of Physics, The University of Texas at Austin, 2515 Speedway PMA 14.202, Austin, TX 78712, United States.
E-mail: wgilpin@utexas.edu

Abstract

Motivation: Single-cell sequencing technology maps cells to a high-dimensional space encoding their internal activity. Recently-proposed virtual cell models extend this concept, enriching cells' representations based on patterns learned from pretraining on vast cell atlases.

Results: This review explores how advances in understanding the structure of natural language embeddings informs ongoing efforts to analyze single-cell datasets. Both fields process unstructured data by partitioning datasets into tokens embedded within a high-dimensional vector space. We discuss how the context of tokens influences the geometry of embedding space, and how low-dimensional manifolds shape this space's robustness and interpretation. We highlight how new developments in foundation models for language, such as interpretability probes and in-context reasoning, can inform efforts to construct cell atlases and train virtual cell models.

Availability and Implementation: Code is available at https://github.com/williamgilpin/celltoken.

1 Introduction

Modern single-cell experiments *decompile* the cell—abstracting it away from its squishy context, and rendering it as a single point in a high-dimensional vector space. Computational workflows attempt to invert this process: spatial transcriptomics recovers information about a cell's position, while lineage tracing reconstructs developmental stages. Recent efforts to construct virtual cells—massive machine learning models built upon language model architectures—represent the next step of this process. Trained on vast amounts of genomic data, these models aim to provide richer, more informative representations than the raw count matrices produced by single cell experiments.

How do we know if the representations learned by virtual cell models are meaningful? If a single-cell embedding exactly matches known regulatory and developmental mechanisms, an embedding space may be accurate yet uninformative for making new discoveries. Conversely, if this space fails to recapitulate known relationships, it is difficult to attribute this to novelty or inaccuracy. A similar problem arises in statistical learning. Large language models are trained on vast, unannotated volumes of text. To represent diverse text sources consistently, these models initially split input text into discrete tokens: minimal units consisting of words or word fragments. They then convert these tokens into vectors in a high-dimensional space, a representation that enables further processing by modern, continuously-valued learning models.

The success of language models stems, in part, from the ability of language embeddings to accurately encode syntactic and semantic structure in high-dimensional spaces. The unique properties of high-dimensional geometry allow embeddings to effectively encode the semantic structure of language along low-dimensional manifolds, mirroring findings from single-cell biology in identifying developmental pathways and rare cell types. What insights can single-cell embeddings gain from the statistical learning community? Here, we review recent developments and commonalities between these fields, highlighting general principles of language embeddings that may inform ongoing work in single-cell genomics.

2 Context shapes the geometry of embeddings

Modern large language models owe their scale to self-supervised training, which obviates the need to collect expensive labeled training data. Given a sequence of words, ”The parliamentarian led the assembly.”, a single word is masked and the model is trained to fill in the blank: ”The `[TOK]` led the assembly.” When trained at scale, models learn to group certain words (e.g. ”leader,” ”president,” ”speaker”) that appear in similar contexts. Theoretical motivation for context masking comes from the *distributional hypothesis*, which equates distances between vector representations of different words in embedding space, with distances between distributions of co-occuring tokens within the training corpus (Harris 1954; Firth 1957). The distributional hypothesis typically describes co-occurrence statistics: words like ”president” or ”parliamentarian” often appear with similar other words. However, the distribution may be further conditioned on language type, historical era, domain-specific register, or other latent variables that modulate word usage. In systems biology, the distributional hypothesis motivates efforts to train

Received: 30 April 2025; **Revised:** 19 September 2025; **Accepted:** 21 October 2025

self-supervised foundation models from single cell data, often termed ”virtual cells” (Bunne *et al.* 2024). An implicit assumption of such approaches is that models can learn informative, predictive knowledge purely from training to be self-consistent. Such approaches inherently invoke a distributional hypothesis—that cells occurring in the same tissues, interactions, or regulatory roles ought to retain that similarity when represented in a single-cell workflow, and that this similarity can be exploited for self-supervised training.

Predating modern large language models, `word2vec` language embeddings introduced an early notion of *context* to word representations. During training, `word2vec` directly optimizes an objective function motivated by the distributional hypothesis, producing an embedding that maximizes the posterior probability of word-context pairs seen in the corpus, while minimizing the probability of randomly-generated pairs (Mikolov 2013). This approach represents contrastive learning, which allows an embedding space to be constructed for data that otherwise lacks a well-defined distance metric. A typical fully-trained `word2vec` model maps each of 10^7 distinct words to a point in a 300-dimensional continuous vector space. Because the distributional objective is optimized only during training on the text corpus, `word2vec` produces *static embeddings*: after training, any appearance of a given token always maps to the same point in embedding space.

Cell gene expression profiles lack an an obvious distance metric, and the results of computational workflows like cell type clustering vary depending on the choice of cell-cell distance metric such as Euclidean distance, correlation, or t-statistic (Ji *et al.* 2023). Raw expression profiles are typically context-independent. After isolation, sequencing, and demultiplexing, a cell becomes a collection of RNA transcripts, each of which may be considered a vector approximating the transcript counts per gene per cell (Stuart and Satija 2019). The expression levels of each gene thus uniquely determine the embedding, decoupling a given cell's representation from others. Thus, in principle, raw count data do not invoke the distributional hypothesis: a cell's embedding is an innate property, rather than a property relative to a corpus of cells. Many preprocessing schemes applied to count matrices—such as batch or cell cycle correction—enforce static, context-free structure in embedding space (Korsunsky *et al.* 2019). However, data reduction methods like principal component analysis for visualization, or unsupervised clustering for cell type identification, produce context-dependent representations that depend on relative differences among cells. Context-dependence also arises when multiple datasets are merged, or when end-to-end embedding models are trained across many datasets. However, such approaches stop short of invoking the distributional hypothesis, because they do not enforce a notion of context tied to co-occurrence statistics. In contrast, the extensive pretraining used in modern single-cell foundation models aims to learn a distance metric among expression profiles based on statistical patterns in expression across the training data (Heimberg *et al.* 2025).

3 The geometry of embedding spaces

Theoretical analysis of `word2vec` and its variants shows that these methods, in practice, factorize a matrix representing the mutual information between the distribution of each token across the corpus, and the distribution of its context (Fig. 1A) (Levy and Goldberg 2014). Linguistic structures, such as synonym clusters, lead to low-rank structure in this matrix (Dhillon *et al.* 2015; Allen *et al.* 2019), similar to the low-rank structure that forms in single-cell count matrices due to statistical similarities in the expression vectors of cells belonging to the same type (Fig. 1B) (Nitzan *et al.* 2019). Low-dimensional manifolds in single-cell embeddings typically arise from highly-coordinated biological processes, such as differentiation, which exhibit predominantly deterministic dynamics. In language embeddings, low-rank structure arises due to overparametrization—the highest-rank word-context matrix would simply represent an isotropic Gaussian distribution. Anisotropy in high-dimensional embeddings thus indicates structure in the underlying generative process, whether linguistic or biological.

A key limitation of static language embeddings stems from polysemy, in which the same token has multiple meanings (Garí Soler and Apidianaki 2021; Liu *et al.* 2020). For example, ”bank” may refer to the side of a riverbed, or to a financial institution. Static word embeddings like `word2vec` place polysemous tokens at intermediate positions in embedding space, between positions associated with their divergent meanings. Such compromises distort and curl embedding space, reducing the space's ability to represent large-scale structure by making distances between vectors less meaningful (Jakubowski *et al.* 2020; Neelakantan *et al.* 2014; Goel *et al.* 2022). As a result, static embedding models tend to underestimate differences among strongly-distinct concepts, limiting their ability to recognize hierarchies among words (Nickel and Kiela 2017). In gene expression, curvature due to polysemy may arise due to biological processes, rather than artifacts. Cellular differentiation datasets exhibit low curvature in regions associated with stereotyped cell states, punctuated by high-curvature regions associated with transitions (Sritharan *et al.* 2021). These transition states, such as differentiating stem cells, occupy intermediate locations in embedding space (Wang *et al.* 2020). However, spurious polysemy can also arise due to technical errors, leading to unresolved cell subtypes or tissue groups. This effect becomes more pronounced at low read depth, resulting in missing genes or greater sampling error in counts. For example, blood vascular endothelial cells share relatively consistent transcriptional profiles, due to their similar structural roles across different tissues (Kalucka *et al.* 2020). Endothelial cells from different tissues often map to the same area in embedding space, despite their anatomical separation. Resolving this ambiguity either requires additional marker genes and greater sequencing resolution, or assays that barcode transcripts with additional information. For example, cell painting produces a high-dimensional vector of morphological features extracted from fluorescence microscopy (Bray *et al.* 2016), while CITE-Seq augments each transcript with information about cell surface proteins—thus avoiding cellular polysemy (Stoeckius *et al.* 2017).

Contemporary language models use dynamic token embeddings, in which a given token's embeddings varies based on its context after training (Devlin *et al.* 2019; Liu *et al.* 2020; Reisinger and Mooney 2010). The standard mechanism, self-attention, combines a token's static representation, neighboring context tokens, and a positional encoding (Vaswani *et al.* 2017). The resulting joint representation thus varies even after training when it appears in new contexts. Thus, while static embeddings associate each token with a single

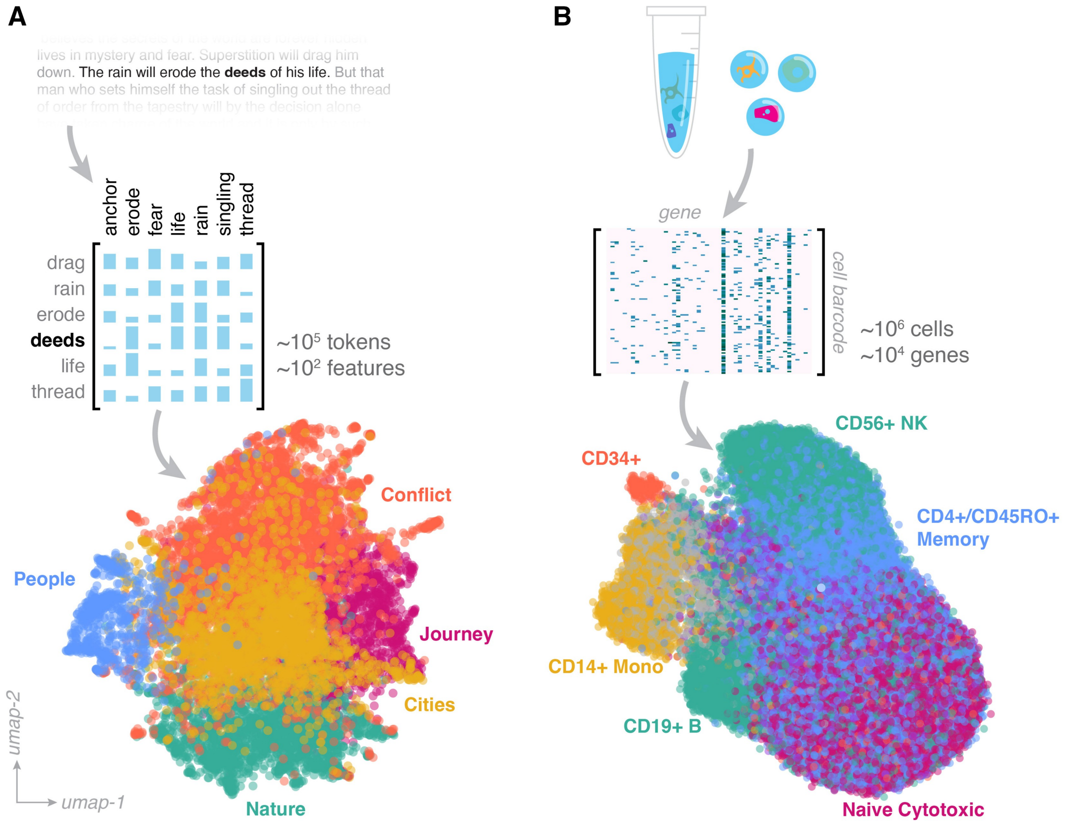

Figure 1. Low-rank structure in high-dimensional embeddings. (A) An embedding of the full text of the novel *Blood Meridian* (McCarthy 1985) using a word2vec model originally trained on a dataset of 10^{11} words drawn from Google News articles (Mikolov *et al.* 2013a). Vectors are clustered using K-means partitioning, and then summarized into metagroups with a topic embedding model (colors and annotations). (B) An embedding of 6×10^4 human peripheral blood mononuclear cells based on single-cell RNA sequencing of 1.6×10^4 genes. Colors correspond to immune cell subtypes, as determined by marker genes for characteristic cell surface proteins like CD4, CD8, etc.

embedding point, dynamic embeddings map each token to a cloud of points capturing the diverse contexts in which it appears. The distance between the same token in different contexts is smaller than the distance between tokens, consistent with low-dimensional, anisotropic structure (Ethayarajh 2019).

In large-scale gene expression datasets like cell atlases, dynamic cell embeddings improve the global structure of representations. Spatial transcriptomics augments each transcript with information about the cell's absolute spatial position, or relative position among neighboring cells. As a result, embeddings learned by these methods encode an underlying metric, and so both local and global distances are meaningful (Tian *et al.* 2023; Nitzan *et al.* 2019). More abstract context information, such as organ group or tissue annotations, improves embeddings by disambiguating similar transcriptional profiles arising in distinct contexts (Xu *et al.* 2021). Conceptually, these approaches resemble language models that combine tokenization with queries to an external database that provides richer context (Borgeaud *et al.* 2022; Khandelwal *et al.* 2020). This can include structured sources of information about tokens, like encyclopedias or human-curated concept maps (Speer *et al.* 2017; Zhang *et al.* 2019). The resulting models capture global relationships among concepts, without requiring substantial additional training. For single-cell data, similar approaches enrich transcript information with tissue or preexisting cell type annotations (Brbić *et al.* 2020; Lin *et al.* 2022; Lotfollahi *et al.* 2022), gold-standard experimental associations or transcription factors (Lee *et al.* 2024), gene ontologies (Yuan *et al.* 2024), or even topic information from scientific literature databases (Zhao *et al.* 2021; Istrate *et al.* 2024). Other efforts pair each cell with gene-level context (such as sequence position or chromatin accessibility) to highlight mechanistic relationships (Chen *et al.* 2024; Fu *et al.* 2025).

When single cell foundation models are trained using self-supervision, their internal representations can be extracted and used as dynamic embeddings of expression vectors (Fang *et al.* 2024; Song *et al.* 2021; Eraslan *et al.* 2019; Fu *et al.* 2025; Zhao *et al.* 2021; Istrate *et al.* 2024). Multimodal models produce richer representations by training on both expression data and external information, such as known regulatory hierarchies (Zhao *et al.* 2021; Istrate *et al.* 2024; Bunne *et al.* 2024; Lopez *et al.* 2018). Many such approaches use self-attention to dynamically process tokens, as well as contrastive learning, producing representations with an underlying similarity metric—representing a form of distributional hypothesis for expression vectors (Cui *et al.* 2024; Theodoris *et al.* 2023; Vaswani *et al.* 2017; Han *et al.* 2022).

4 Are cells or genes the "words" of single-cell biology?

Many large-scale pretrained models for genomic data directly adapt language architectures, treating the genome as a large body of text, with nucleotides acting as an alphabet and genes as words (Consens *et al.* 2025; Ji *et al.* 2021; Levine *et al.* 2024; Rizvi *et al.* 2025; Rosen *et al.* 2023; Cui *et al.*

2024). Genes thus may seem to be a more natural analogue to words in statistical learning frameworks. However, this equivalence has limits: genes do not recur within a single genome, and so identifying variations in their function across cells or individuals requires expression information, a quantity without an obvious analogue in natural language. Instead, from the perspective of the distributional hypothesis, cells, not genes, represent minimal tokens, because similarity among cells can be inferred from recurring patterns across different biological contexts. As a result, many proposed applications of virtual cell models, such as cell type identification or lineage tracing, implicitly treat cells as words (Bunne *et al.* 2024; Pearce *et al.* 2025). Context thus arises from neighboring cells, tissue microenvironments, or developmental stages, while genes represent the fixed vocabulary describing each cell token. Ambiguities about the correct unit of tokens also exist in language models: while many language models treat words as tokens, others use finer-grained units such as characters (Boukkouri *et al.* 2020) or even raw byte sequences (Xue *et al.* 2022). Similarly, in biological settings, the choice of ”token” is not fixed *a priori*, but should be defined at the level where meaningful context recurs.

5 Analogies as manifolds in embedding space

Effective language embeddings encode semantic relationships as distances (Mikolov *et al.* 2013b). For example, the vector from ”Sacramento” to ”California” in embedding space may match the vector from ”Austin” to ”Texas.” As a result, vector arithmetic in `word2vec` solves unseen analogy problems from college admissions exams, even without retraining (Fig. 2A)(Liu *et al.* 2017; Turney and Littman 2005). Embedding space thus unfolds computation into a higher-dimensional space in which reasoning coincides with distances (Ushio *et al.* 2021). In this sense, early word embeddings foreshadowed modern works on in-context learning and zero-shot inference, phenomena in which sufficiently-large models are able to perform new tasks not seen during training (Kojima *et al.* 2022).

Organismal cell atlases exhibit well-defined clusters associated with cell and tissue types. However, cells with different compositions but similar functions can nonetheless occupy similar relative locations in embedding space. For example, immune cells form subtypes within different organ groups, such as Kupffer cells in the liver or microglia in the brain. In whole-organism cell atlases, these cells typically appear in separate clusters associated with their primary organ groups. However, within each organ cluster, they occupy similar positions relative to other cells, underscoring their analogous roles (Wang *et al.* 2022; Suo *et al.* 2022; Gautier *et al.* 2012).

Language embeddings also capture continuous relationships among tokens. For example, escalating sequences like ”good,” ”better,” or ”best” map to linear sequences in `word2vec` (Mikolov *et al.* 2013b). Generally, word embeddings exhibit high anisotropy (Mimno and Thompson 2017; Ethayarajh 2019), with embeddings spanning low-dimensional manifolds within the higher-dimensional representation space. Depending on the language, this manifold has effective dimensionality $\sim 10^1$, even when the feature dimension is $\sim 10^2$ (Mu *et al.* 2018). These manifolds capture gradations in meanings among similar words, shifts in a word’s meaning over time, singular-plural pairs, or even groups of synonyms (Hamilton *et al.* 2016). For example, in dynamic embeddings produced by large language models, days of the week and calendar months map onto circular manifolds, while colors and years map onto linear manifolds (Engels *et al.* 2025; Modell *et al.* 2025). Consistent with these manifolds representing informative subspaces, the performance of embeddings in downstream tasks initially increases with the embedding dimension, but it eventually plateaus at a fixed multiple of the manifold dimension (Yin and Shen 2018).

Similar low-dimensional manifolds arise in cell embeddings. Across different datasets, cell replication cycles and circadian rhythms form rings (Kowalczyk *et al.* 2015), spatially-extended tissues form grids (Adler *et al.* 2019; Nitzan *et al.* 2019), and cell differentiation hierarchies form branches (Fig. 2B) (Paul *et al.* 2015). In one well-known case, populations of blood cells of mixed maturity form a pitchfork in embedding space, illustrating a continuous progression from

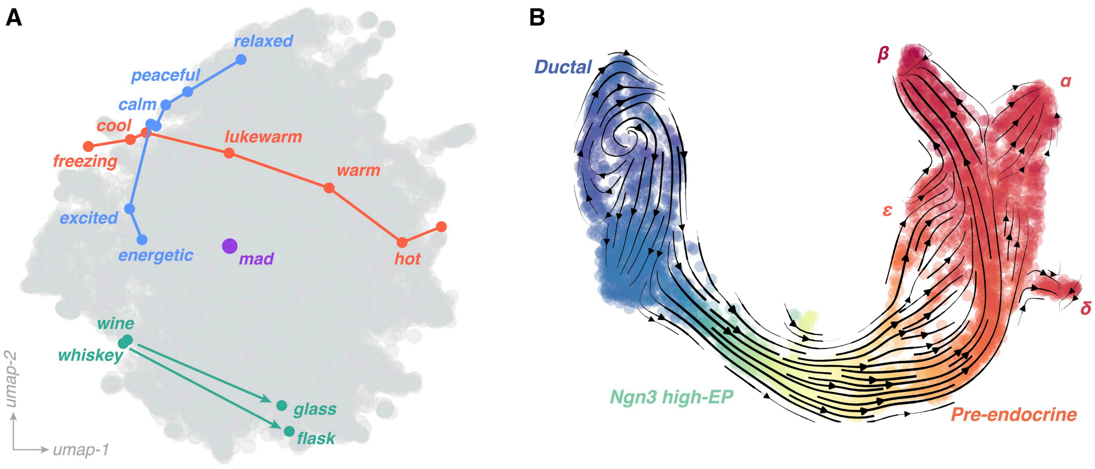

Figure 2. Analogies and low-dimensional manifolds. (A) Embeddings of particular sequences of tokens using the model of Fig. 1, with examples of escalating manifolds (red and blue lines), which overlap in regions with similar meaning (weak polysemy). A token with strong polysemy appears at an intermediate location (purple circle). An example of an analogy relationship encoded as nearly-congruent difference vectors (turquoise arrows). While nonlinear embedding methods like UMAP distort the local metric over large scales (Chari and Pachter 2023), the nearby position of the two analogy vectors’ heads and tails protects their congruency. (B) RNA Velocity applied to developing endocrine cells in the pancreas (Bastidas-Ponce *et al.* 2019; La Manno *et al.* 2018). Vectors correspond to development direction, and color corresponds to pseudotime assigned via diffusion components. Cell types along the differentiation axis are overlaid.

stem cells to different types of blood cells (Paul *et al.* 2015). Gene expression manifolds have a typical intrinsic dimensionality $\sim 10^1$, compared to the $\sim 10^5$ genes measured in a typical single-cell experiment (Sritharan *et al.* 2021). Just as word embeddings trace their properties to low-rank structure in the word-context mutual information matrix, biological processes confer low-rank structure on count matrices (Nitzan and Brenner 2021; Thibeault *et al.* 2024).

Theoretical models of word embeddings frame text generation as a stochastic dynamical system, with sentence formation as a random walk in token embedding space (Arora *et al.* 2016, 2015; Hashimoto *et al.* 2016). Under this framework, semantic manifolds act as kinetic traps for the walk, with synonymous tokens acting as basins, and connective phrases acting as bridges. In single-cell analysis, diffusion maps simulate the action of many random walks through expression space (Coifman and Lafon 2006; Wolf *et al.* 2019). These methods represent a standard approach to calculating pseudotime, which orders unsorted cell embeddings to identify temporal progressions of cells (like developmental stages) (Haghverdi *et al.* 2016; Setty *et al.* 2019). In dynamic word embeddings, or in static embeddings trained on corpora from different historical periods, the relative positions of words gradually shift over time. This semantic drift may be quantified using a calculation resembling pseudotime (Bamler and Mandt 2017).

6 Cross-lingual embeddings

Many features of natural language, such as parts-of-speech, intensifiers, and modifiers, recur across languages. For example, while English and Sanskrit have different inflections and character sets, they exhibit similar verb conjugations and noun declensions. Machine translation models must disentangle these distinctions to construct maps between different languages' embedding spaces. A common approach is an encoder-decoder translation model, which trains a model to map sentences onto a universal representation in a latent space (Wu *et al.* 2016). For example, a Sanskrit encoder maps a sentence into the latent space, and an English decoder then translates it. Syntax information, such as word ordering or inflection, is typically distinct among languages and thus not necessarily preserved in the latent space. In contrast, manifolds associated with semantic content remain conserved, and the low-dimensional latent coordinates capture information such as token positions and parts-of-speech (Artetxe and Schwenk 2019; Chang *et al.* 2022). Taking this approach even further, cross-lingual translation models construct a single shared latent space from many languages. These models typically outperform single-pair translation models, particularly for languages with less available training data, like Swahili or Urdu (Conneau 2019).

Could the same effect hold for rare cell types? One analogy for cross-lingual latent spaces is shared embeddings of cell types across distinct organisms. Statistical alignment methods may be used to combine cell type populations across species with similar tissue groups (Fig. 3A) (Butler *et al.* 2018; Stein-O'Brien *et al.* 2019; Song *et al.* 2023; Tarashansky *et al.* 2021; Kriebel and Welch 2022; Yang *et al.* 2024; Pearce *et al.* 2025). Like understudied languages, rarer cell types benefit from integrated analysis; for example, in a joint embedding of human and mouse pancreatic cells, a combined embedding better resolves subpopulations associated with stress response during protein assembly (Butler *et al.* 2018). Recent works extend this concept by proposing universal cell embeddings, in which a single foundation model is trained on data spanning subjects, species, and even sequencing modalities (Rosen *et al.* 2023; Lopez *et al.* 2018; Rosen *et al.* 2024). The resulting embedding exhibits emergent properties, including zero-shot embedding of new species or tissues without retraining.

As single-cell foundation models become more contextual and high-dimensional, the geometry of their embeddings may encode nontrivial biological structure, such as indirect regulatory grammars. One approach to probing embedding geometry is topological data analysis, a set of tools for analyzing high-dimensional point clouds. In language models, these methods detect cusp-like singularities that form due to polysemy (Jakubowski *et al.* 2020). On gene expression data, topological methods quantify the degree to which processes like developmental lineages produce low-dimensional manifolds or branches (Palande *et al.* 2023; Korem *et al.* 2015). Newly-

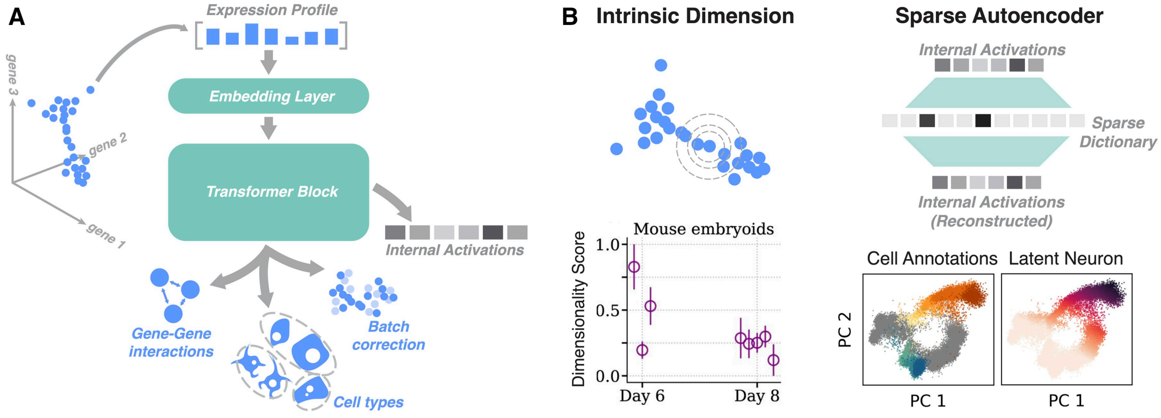

Figure 3. Mechanistic interpretability in single-cell foundation models. (A) Common architectural features and target tasks for single-cell foundation models. (B) Mechanistic interpretability methods for single cell embeddings. (Left) Intrinsic dimensionality may be calculated directly from expression profiles, or from internal activations of the model. Inset shows the intrinsic dimensionality of staged expression profiles from developing mice. Panel adapted from Ref. (Biondo *et al.* 2024). (Right) Sparse autoencoders are trained in an unsupervised manner to reconstruct internal activations of foundation models, by mapping activations to sparse combinations of features in a latent dictionary. Inset shows application of sparse autoencoders to the activations of the Universal Cell Embedding model on a dataset of human bone marrow. The left subpanel corresponds to annotated cell types, while the right corresponds to the decoding of a single latent unit. Panels adapted from Ref. (Schuster 2024).

introduced robust statistical estimators of the intrinsic dimensionality of point clouds may be used to probe internal representations in artificial neural networks (Fig. 3B, left) (Facco *et al.* 2017). Recent results use these estimators to show that the intrinsic dimensionality of gene expression correlates with pluripotency, across diverse taxa ranging from mice to zebrafish (Biondo *et al.* 2024). Several approaches directly constrain representations to enforce particular topological features. For example, imposing hyperbolic structure on embeddings improves resolution of branching processes associated with differentiation (Kuang *et al.* 2025; Schlüter and Uhler 2025; Klimovskaia *et al.* 2020; Zhou and Sharpee 2021; Ding and Regev 2021; Bhasker *et al.* 2025).

The emerging field of *mechanistic interpretability* examines the reasoning and internal representations of large language models. One such approach, linear probes, trains small linear regression models to predict particular linguistic features, like parts-of-speech or subject-verb agreement, from latent states or internal activations of layers (Alain and Bengio 2016; Mamou *et al.* 2020; Belinkov and Glass 2019). This approach quantifies how explicitly different features are represented, and can identify where semantic versus grammatical information resides within the model. In single-cell foundation models, linear probes identify gene families that the model weighs particularly highly in making predictions, such as by highlighting inflammation and heatshock genes in immune cell datasets (Pedrocchi *et al.* 2024) However, linear probes typically require a supervision signal, such as a ground truth dataset showing known effects of knockdowns, motivating the need for unsupervised methods. Sparse autoencoders train shallow, wide neural networks to encode the activations of individual layers of large models (Fig. 3B, right). The width of the latent space, coupled with a strong sparsity penalty, encourages sparse autoencoders to map single concepts onto each latent dimension, thus unfolding polysemous activations in the original large model (Gao *et al.* 2025). In single-cell foundation models, sparse autoencoders isolate cell types that otherwise would be difficult to distinguish in embeddings (Schuster 2024).

7 Task-independence and amortization of reasoning

Word embeddings derive their utility from their independence from downstream tasks. Training frontier models typically requires access to large amounts of computing resources, with contemporary models like RoBERTa-large optimizing as many as 10^9 parameters over 10^{12} language tokens (Liu *et al.* 2019). However, once trained, these models may be used as a preprocessing step for downstream tasks like sentiment classification. Single-cell technologies share a goal of identifying general representations that foreground relevant biological variables, while removing uninformative variation like batch or technical effects (Rosen *et al.* 2023). Embeddings thus represent one motivation for the emerging foundation model paradigm in both language modelling and single-cell analysis, which argues that large-scale pretraining on diverse datasets leads to simpler starting representations for smaller-scale tasks. Task-independent embeddings thus serve to amortize computation.

Large-scale pretrained models exhibit *inference-time computation*, in which they spontaneously solve new tasks without additional training (Kojima *et al.* 2022). For example, large language models can be prompted to produce poetry with meter and scansion that are unseen in their training corpus (Walsh *et al.* 2024). The underlying mechanism, *in-context learning*, exploits the emergent ability of large models not only to retrieve, but also process, information during inference. Inference-time symbolic reasoning appears to improve with model scale, with language models recently advancing from solving elementary-school word problems to standardized mathematics exams for undergraduates (Wu *et al.* 2024; Liu *et al.* 2024; Ahn *et al.* 2024). Achieving similar results for biological datasets represents a frontier for single-cell foundation models. Several recent models exhibit forms of inference-time reasoning, such as zero-shot embedding of novel cell types, prediction of protein interactions, and anticipation of responses to genetic perturbations (Cui *et al.* 2024; Theodoris *et al.* 2023). However, these tasks have unclear difficulty compared to language modeling tasks like standardized exams, leading to conflicting results regarding the efficacy of current single-cell foundation models (Kedzierska *et al.* 2023; Csendes *et al.* 2025; Ahlmann-Eltze *et al.* 2025; Wenteler *et al.* 2024). A better test may be the ability of large models to decipher the indirect, multiscale, and highly nonlinear logic of many regulatory circuits. For example, the immune system implements elaborate combinatorial receptor-ligand interactions, phosphorylation cascades, and feedback loops, in order to discriminate self from nonself antigens (Germain 2001; Chakraborty and Weiss 2014). Parsing these logical circuits is akin to solving a complex mathematical reasoning problem, requiring models that can effectively process symbolic information.

8 Conclusion: limitations of the analogy

Drawing parallels between language models and single-cell embeddings reveals shared principles in how high-dimensional spaces encode structured, context-dependent information. However, the analogy between cells and word tokens has natural limits, presaging potential limitations of foundation models for single-cell genomics.

In natural languages, ”context” arises from discrete, ordered sequences, where exact position and co-occurrence of tokens convey meaning (Harris 1954; Firth 1957). In contrast, a cell's context arises from a web of spatial relationships, signaling interactions, lineage history, and environmental conditions—most of which are not naturally represented as ordered sequences. No two cells are exact replicates, and their surrounding biochemical and environmental context can never be fully reproduced (Stuart and Satija 2019; Hicks *et al.* 2018). Furthermore, unlike words in a corpus, cells cannot be resampled from the same underlying distribution without perturbing the system, limiting the robustness and stationarity assumptions of statistical analogies. A key challenge for emerging virtual cell models will thus be their ability to distill informative context in order to resolve polysemy in cell states while still finding concise representations. Other challenges include integrating diverse experimental contexts without loss of biological specificity, and capturing nonlinear regulatory logic within embedding spaces (Fang *et al.* 2024; Eraslan *et al.* 2019; Fu *et al.* 2025; Cui *et al.* 2024).

In contrast to language, where token context is explicit and uniformly structured, biological context is multiscale, incomplete, and often indirect. Moreover, while neighboring words directly inform language token context, a cell's relevant

”neighbors” may be defined in multiple, potentially conflicting ways (physical proximity, developmental stage, functional similarity). Resolving this ambiguity requires contextual cell embeddings, integrating heterogeneous modalities such as spatial transcriptomics, proteomics, chromatin accessibility, or lineage tracing to derive a unified representation. Truly multimodal foundation models offer a potential solution, by treating auxiliary information—like gene ontologies, medical literature, or known regulatory hierarchies—on an even footing with expression data, thus decoupling modality-specific factors from informative biological variation (Hu *et al.* 2025; Levine *et al.* 2024; Rizvi *et al.* 2025; Theodoris *et al.* 2023). However, combining modalities at scale raises technical challenges: batch effects, inconsistent coverage among different modalities, and the difficulty of defining context windows across different samples (Armingol *et al.* 2021; Bastidas-Ponce *et al.* 2019). Even if relevant auxiliary information is available, its incorporation into embeddings can amplify biases in the experimental design, leading to overfitting to specific tissue types, organisms, or experimental protocols (Rosen *et al.* 2023; Cui *et al.* 2024). Identifying such effects will be necessary in future virtual cell models, and represents an area where mechanistic interpretability and low-dimensional manifold discovery may prove informative.

The cell token analogy also breaks down when considering the dynamical nature of biological systems. In languages, dynamic embeddings model variability in token meaning without altering the underlying corpus. Yet in biology, a cell’s ”meaning” irreversibly changes over time through differentiation, signaling, senescence, and adaptation (La Manno *et al.* 2018; Haghverdi *et al.* 2016; Bergen *et al.* 2020). Capturing these processes requires embedding models that are temporally aware, capable of representing continuous trajectories, and robust to sparse or noisy longitudinal data. Moreover, truly contextual embeddings for cells must incorporate causal relationships—distinguishing between correlation and regulatory influence—a level of mechanistic grounding without an obvious equivalence in grammatical rules. Improved benchmarks, which test the ability of foundation models to parse complex and indirect regulatory logic, will help help transform cell representations from descriptive maps into predictive, reasoning-ready representations for biology.

Acknowledgements

W.G. thanks Ski Krieger and the members of the Chan-Zuckerberg theory in biology group. W.G. thanks Medici Therapeutics.

Author contributions

W.G. wrote and revised the manuscript.

Conflict of interest: No competing interest is declared.

Funding

This project has been made possible in part by Grant No. DAF2023-329596 from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation. W.G. is supported by NSF DMS 2436233 and NSF CMMI 2440490.

References

Adler M, Kohanim YK, Tendler A *et al.* Continuum of gene-expression profiles provides spatial division of labor within a differentiated cell type. *Cell Syst* 2019;**8**:43–52.e5.

Ahlmann-Eltze C, Huber W, Anders S. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nat Methods* 2025;**22**:1657–61.

Ahn J, Verma R, Lou R *et al.* Large language models for mathematical reasoning: Progresses and challenges. arXiv preprint. arXiv: 2402.00157, 2024, preprint: not peer reviewed.

Alain G, Bengio Y. Understanding intermediate layers using linear classifier probes. In: *The Fourth International Conference on Learning Representations (ICLR)*, 2016.

Allen C, Balazevic I, Hospedales T. What the vec? towards probabilistically grounded embeddings. *Adv Neural Inf Process Syst* 2019:32.

Armingol E, Officer A, Harismendy O *et al.* Deciphering cell–cell interactions and communication from gene expression. *Nat Rev Genet* 2021;**22**:71–88.

Arora S, Li Y, Liang Y *et al.* Random walks on context spaces: towards an explanation of the mysteries of semantic word embeddings. arXiv preprint. arXiv: 150203520, 2015:385–399, preprint: not peer reviewed.

Arora S, Li Y, Liang Y *et al.* A latent variable model approach to pmi-based word embeddings. *TACL* 2016;**4**:385–99.

Artetxe M, Schwenk H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans Assoc Comput Linguist*, 2019;**7**:597–610.

Bamler R, Mandt S. Dynamic word embeddings. In: *International Conference on Machine Learning*, PMLR, 2017, 380–9.

Bastidas-Ponce A, Tritschler S, Dony L *et al.* Comprehensive single cell mrna profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development* 2019;**146**:dev173849.

Belinkov Y, Glass J. Analysis methods in neural language processing: a survey. *Transactions of the Association for Computational Linguistics* 2019;**7**:49–72.

Bergen V, Lange M, Peidli S *et al.* Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol* 2020; **38**:1408–14.

Bhasker N, Chung H, Boucherie L *et al.* Uncovering developmental lineages from single-cell data with contrastive poincaré maps. bioRxiv, 2025:2025–08.

Biondo M, Cirone N, Valle F *et al.* The intrinsic dimension of gene expression during cell differentiation. bioRxiv, 2024:2024–08.

Borgeaud S, Mensch A, Hoffmann J *et al.* Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, PMLR, 2022, 2206–40.

Boukkouri HE, Ferret O, Lavergne T *et al.* Characterbert: Reconciling elmo and bert for word-level open-vocabulary representations from characters. In: *Proceedings of the 28th International Conference on Computational Linguistics. International Committee on Computational Linguistics*, 2020.

Bray M-A, Singh S, Han H *et al.* Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat Protoc* 2016;**11**:1757–74.

Brbić M, Zitnik M, Wang S *et al.* Mars: discovering novel cell types across heterogeneous single-cell experiments. *Nat Methods* 2020; **17**:1200–6.

Bunne C, Roohani Y, Rosen Y *et al.* How to build the virtual cell with artificial intelligence: priorities and opportunities. *Cell* 2024; **187**:7045–63.

Butler A, Hoffman P, Smibert P *et al.* Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 2018;**36**:411–20.

Chakraborty AK, Weiss A. Insights into the initiation of tcr signaling. *Nat Immunol* 2014;**15**:798–807.

Chang T, Tu Z, Bergen B. 2022. The geometry of multilingual language model representations. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 119–36.

Chari T, Pachter L. The specious art of single-cell genomics. *PLoS Comput Biol* 2023;**19**:e1011288.

Chen H, Ryu J, Vinyard ME *et al.* Simba: single-cell embedding along with features. *Nat Methods* 2024;**21**:1003–13.

Coifman RR, Lafon S. Diffusion maps. *Appl Comput Harmon Anal* 2006;**21**:5–30.

Conneau A, Khandelwal K, Goyal N et al. Unsupervised cross-lingual representation learning at scale. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, 8440–51.

Consens ME, Dufault C, Wainberg M *et al.* Transformers and genome language models. *Nat Mach Intell* 2025;**7**:346–62.

Csendes G, Sanz G, Szalay KZ *et al.* Benchmarking foundation cell models for post-perturbation rna-seq prediction. *BMC Genomics* 2025;**26**:393.

Cui H, Wang C, Maan H *et al.* Scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nat Methods* 2024;**21**:1470–80.

Devlin J, Chang M-W, Lee K *et al.* 2019. Bert: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–86.

Dhillon PS, Foster DP, Ungar LH. Eigenwords: spectral word embeddings. *J Mach Learn Res* 2015;**16**:3035–78.

Ding J, Regev A. Deep generative model embedding of single-cell rna-seq profiles on hyperspheres and hyperbolic spaces. *Nat Commun* 2021;**12**:2554.

Engels J, Michaud EJ, Liao I *et al.* Not all language model features are one-dimensionally linear. In: *The Thirteenth International Conference on Learning Representations*, 2025.

Eraslan G, Simon LM, Mircea M *et al.* Single-cell rna-seq denoising using a deep count autoencoder. *Nat Commun* 2019;**10**:390.

Ethayarajh K. 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)*. Association for Computational Linguistics, 2019.

Facco E, d'Errico M, Rodriguez A *et al.* Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Sci Rep* 2017;**7**:12140.

Fang Z, Zheng R, Li M. Scmae: a masked autoencoder for single-cell rna-seq clustering. *Bioinformatics* 2024;**40**:btae020.

Firth J. A synopsis of linguistic theory, 1930–1955. *Studies in Linguistic Analysis* 1957:10–32.

Fu X, Mo S, Buendia A *et al.* A foundation model of transcription across human cell types. *Nature* 2025;**637**:965–73.

Gao L, la Tour TD, Tillman H *et al.* Scaling and evaluating sparse autoencoders. In: *The Thirteenth International Conference on Learning Representations*, 2025.

Garí Soler A, Apidianaki M. Let's play Mono-poly: bert can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics* 2021;**9**:825–44.

Gautier EL, Shay T, Miller J, Immunological Genome Consortium *et al.* Gene-expression profiles and transcriptional regulatory pathways that underlie the identity and diversity of mouse tissue macrophages. *Nat Immunol* 2012;**13**:1118–28.

Germain RN. The art of the probable: system control in the adaptive immune system. *Science* 2001;**293**:240–5.

Goel A, Sharma C, Kumaraguru P. An unsupervised, geometric and syntax-aware quantification of polysemy. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022,10565–74.

Haghverdi L, Büttner M, Wolf FA *et al.* Diffusion pseudotime robustly reconstructs lineage branching. *Nat Methods* 2016;**13**:845–8.

Hamilton WL, Leskovec J, Jurafsky D. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*. 2016, 2116.

Han W, Cheng Y, Chen J *et al.* Self-supervised contrastive learning for integrative single cell rna-seq data analysis. *Brief Bioinform* 2022; **23**:bbac377.

Harris ZS. Distributional structure. *Word* 1954;**10**:146–62.

Hashimoto TB, Alvarez-Melis D, Jaakkola TS. Word embeddings as metric recovery in semantic spaces. *TACL* 2016;**4**:273–86.

Heimberg G, Kuo T, DePianto DJ *et al.* A cell atlas foundation model for scalable search of similar human cells. *Nature* 2025; **638**:1085–94.

Hicks SC, Townes FW, Teng M *et al.* Missing data and technical variability in single-cell rna-sequencing experiments. *Biostatistics* 2018; **19**:562–78.

Hu L, Qiu P, Qin H *et al.* Regformer: a single-cell foundation model powered by gene regulatory hierarchies. *bioRxiv*, 2025:2025–01.

Istrate A-M, Li D, Karaletsos T. scgenept: Is language all you need for modeling single-cell perturbations? *bioRxiv*, 2024:2024–0.

Jakubowski A, Gasic M, Zibrowius M. 2020. Topology of word embeddings: singularities reflect polysemy. In: *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, 103–13.

Ji Y, Green TD, Peidli S *et al.* Optimal distance metrics for single-cell rna-seq populations. *bioRxiv*, 2023:2023–12.

Ji Y, Zhou Z, Liu H *et al.* Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics* 2021;**37**:2112–20.

Kalucka J, de Rooij LPMH, Goveia J *et al.* Single-cell transcriptome atlas of murine endothelial cells. *Cell* 2020;**180**:764–79.e20.

Kedzierska KZ, Crawford L, Amini AP *et al.* Assessing the limits of zero-shot foundation models in single-cell biology. bioRxiv, 2023:2023-10.

Khandelwal U, Levy O, Jurafsky D *et al.* Generalization through memorization: nearest neighbor language models. In: *International Conference on Learning Representations*, 2020.

Klimovskaia A, Lopez-Paz D, Bottou L *et al.* Poincaré maps for analyzing complex hierarchies in single-cell data. *Nat Commun* 2020; **11**:2966.

Kojima T, Gu SS, Reid M *et al.* Large language models are zero-shot reasoners. *Adv Neural Inf Process Syst* 2022;**35**:22199–213.

Korem Y, Szekely P, Hart Y *et al.* Geometry of the gene expression space of individual cells. *PLoS Comput Biol* 2015;**11**:e1004224.

Korsunsky I, Millard N, Fan J *et al.* Fast, sensitive and accurate integration of single-cell data with harmony. *Nat Methods* 2019; **16**:1289–96.

Kowalczyk MS, Tirosh I, Heckl D *et al.* Single-cell rna-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res* 2015;**25**:1860–72.

Kriebel AR, Welch JD. Uinmf performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. *Nat Commun* 2022;**13**:780.

Kuang D, Qiu G, Kim J. Reconstructing cell lineage trees from phenotypic features with metric learning. In: *Forty-second International Conference on Machine Learning*, 2025.

La Manno G, Soldatov R, Zeisel A *et al.* Rna velocity of single cells. *Nature* 2018;**560**:494–8.

Lee S, Lin C, Chen C-Y *et al.* Chrombert: uncovering chromatin state motifs in the human genome using a bert-based approach. bioRxiv 2024:2024–07.

Levine D, Rizvi SA, Lévy S *et al.* Cell2sentence: teaching large language models the language of biology. In: *International Conference on Machine Learning*, PMLR, 2024, 27299–325.

Levy O, Goldberg Y. Neural word embedding as implicit matrix factorization. *Adv Neural Inf Process Syst* 2014:27.

Lin Y, Wu T-Y, Wan S *et al.* Scjoint integrates atlas-scale single-cell rna-seq and atac-seq data with transfer learning. *Nat Biotechnol* 2022;**40**:703–10.

Liu H, Wu Y, Yang Y. Analogical inference for multi-relational embeddings. In: *International Conference on Machine Learning*, PMLR, 2017, 2168–78.

Liu H, Zheng Z, Qiao Y *et al.* Mathbench: evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark. In: *Findings of the Association for Computational Linguistics ACL 2024*, 2024, 6884–915.

Liu Q, Kusner MJ, Blunsom P. A survey on contextual embeddings. arXiv preprint. arXiv: 2003.07278, 2020, preprint: not peer reviewed.

Liu Y, Ott M, Goyal N *et al.* Roberta: a robustly optimized bert pretraining approach. arXiv preprint. arXiv: 1907.11692, 2019, preprint: not peer reviewed.

Lopez R, Regier J, Cole MB *et al.* Deep generative modeling for single-cell transcriptomics. *Nat Methods* 2018;**15**:1053–8.

Lotfollahi M, Naghipourfar M, Luecken MD *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nat Biotechnol* 2022; **40**:121–30.

Mamou J, Le H, Del Rio M *et al.* Emergence of separable manifolds in deep language representations. In: *International Conference on Machine Learning*, PMLR, 2020, 6713–23.

Mikolov T. Efficient estimation of word representations in vector space. arXiv preprint. In: *International Conference on Learning Representations*, 2013.

Mikolov T, Sutskever I, Chen K *et al.* Distributed representations of words and phrases and their compositionality. *Adv Neural Inf Process Syst* 2013a;**26**.

Mikolov T, Yih W-t, Zweig G. Linguistic regularities in continuous space word representations. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013b, 746–51.

Mimno D, Thompson L. The strange geometry of skip-gram with negative sampling. In: *Empirical Methods in Natural Language Processing*, 2017.

Modell A, Rubin-Delanchy P, Whiteley N. The origins of representation manifolds in large language models. arXiv preprint. arXiv: 2505.18235, 2025, preprint: not peer reviewed.

Mu J, Bhat S, Viswanath P. All-but-the-top: simple and effective post-processing for word representations. In: *International Conference on Learning Representations*, 2018.

Neelakantan A, Shankar J, Passos A *et al.* Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, 1059–69.

Nickel M, Kiela D. Poincaré embeddings for learning hierarchical representations. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, 6341–50.

Nitzan M, Brenner MP. Revealing lineage-related signals in single-cell gene expression using random matrix theory. *Proc Natl Acad Sci USA* 2021;**118**:e1913931118.

Nitzan M, Karaiskos N, Friedman N *et al.* Gene expression cartography. *Nature* 2019;**576**:132–7.

Palande S, Kaste JAM, Roberts MD *et al.* Topological data analysis reveals a core gene expression backbone that defines form and function across flowering plants. *PLoS Biol* 2023;**21**:e3002397.

Paul F, Arkin Y, Giladi A *et al.* Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* 2015;**163**:1663–77.

Pearce JD, Simmonds SE, Mahmoudabadi G *et al.* A cross-species generative cell atlas across 1.5 billion years of evolution: the transcriptformer single-cell model. bioRxiv, 2025:2025–04.

Pedrocchi F, Stark S, Ratsch G *et al.* Identifying biological priors and structure in single-cell foundation models. In: *ICML 2024 Workshop on Efficient and Accessible Foundation Models for Biological Discovery*, 2024.

Reisinger J, Mooney RJ. Multi-prototype vector-space models of word meaning. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, 109–17.

Rizvi SA, Levine D, Patel A *et al.* Scaling large language models for next-generation single-cell analysis. bioRxiv, 2025:2025–04.

Rosen Y, Brbić M, Roohani Y *et al.* Toward universal cell embeddings: integrating single-cell rna-seq datasets across species with saturn. *Nat Methods* 2024;**21**:1492–500.

Rosen Y, Roohani Y, Agarwal A *et al.* Universal cell embeddings: a foundation model for cell biology. bioRxiv 2023:2023–11.

Schlüter HM, Uhler C. Integrating representation learning, permutation, and optimization to detect lineage-related gene expression patterns. *Nat Commun* 2025;**16**:1062.

Schuster V. Can sparse autoencoders make sense of latent representations? arXiv Preprint. arXiv: 2410.11468. 2024, preprint: not peer reviewed.

Setty M, Kiseliovas V, Levine J *et al.* Characterization of cell fate probabilities in single-cell data with palantir. *Nat Biotechnol* 2019; **37**:451–60.

Song Q, Su J, Zhang W. Scgcn is a graph convolutional networks algorithm for knowledge transfer in single cell omics. *Nat Commun* 2021;**12**:3826.

Song Y, Miao Z, Brazma A *et al.* Benchmarking strategies for cross-species integration of single-cell rna sequencing data. *Nat Commun* 2023;**14**:6495.

Speer R, Chin J, Havasi C. Conceptnet 5.5: an open multilingual graph of general knowledge. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31, 2017, 4444–51.

Sritharan D, Wang S, Hormoz S. Computing the riemannian curvature of image patch and single-cell RNA sequencing data manifolds using extrinsic differential geometry. *Proc Natl Acad Sci USA* 2021;**118**: e2100473118.

Stein-O'Brien GL, Clark BS, Sherman T *et al.* Decomposing cell identity for transfer learning across cellular measurements, platforms, tissues, and species. *Cell Syst* 2019;**8**:395–411.e8.

Stoeckius M, Hafemeister C, Stephenson W *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* 2017; **14**:865–8.

Stuart T, Satija R. Integrative single-cell analysis. *Nat Rev Genet* 2019; **20**:257–72.

Suo C, Dann E, Goh I *et al.* Mapping the developing human immune system across organs. *Science* 2022;**376**:eabo0510.

Tarashansky AJ, Musser JM, Khariton M *et al.* Mapping single-cell atlases throughout metazoa unravels cell type evolution. *Elife* 2021; **10**:e66747.

Theodoris CV, Xiao L, Chopra A *et al.* Transfer learning enables predictions in network biology. *Nature* 2023;**618**:616–24.

Thibeault V, Allard A, Desrosiers P. The low-rank hypothesis of complex systems. *Nat Phys* 2024;**20**:294–302.

Tian L, Chen F, Macosko EZ. The expanding vistas of spatial transcriptomics. *Nat Biotechnol* 2023;**41**:773–82.

Turney PD, Littman ML. Corpus-based learning of analogies and semantic relations. *Mach Learn* 2005;**60**:251–78.

Ushio A, Anke LE, Schockaert S *et al.* Bert is to nlp what alexnet is to cv: can pre-trained language models identify analogies? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021, 3609–24.

Vaswani A, Shazeer N, Parmar N *et al.* Attention is all you need. *Adv Neural Inf Process Syst* 2017;**30**.

Walsh M, Preus A, Gronski E. Does chatgpt have a poetic style? 2024.

Wang F, Ding P, Liang X *et al.* Endothelial cell heterogeneity and microglia regulons revealed by a pig cell landscape at single-cell level. *Nat Commun* 2022;**13**:3620.

Wang S, Drummond ML, Guerrero-Juarez CF *et al.* Single cell transcriptomics of human epidermis identifies basal stem cell transition states. *Nat Commun* 2020;**11**:4239.

Wenteler A, Occhetta M, Branson N *et al.* Perteval-scfm: benchmarking single-cell foundation models for perturbation effect prediction. bioRxiv, 2024:2024–10.

Wolf FA, Hamey FK, Plass M *et al.* Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol* 2019;**20**:59.

Wu Y, Schuster M, Chen Z *et al.* Google's neural machine translation system: bridging the gap between human and machine translation. arXiv preprint. arXiv: 1609.08144, 2016, preprint: not peer reviewed.

Wu Y, Sun Z, Li S *et al.* Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. arXiv preprint. arXiv: 2408.00724, 2024, preprint: not peer reviewed.

Xu C, Lopez R, Mehlman E *et al.* Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol Syst Biol* 2021;**17**:e9620.

Xue L, Barua A, Constant N *et al.* Byt5: towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics* 2022;**10**:291–306.

Yang X, Liu G, Feng G, X-Compass Consortium *et al.* Genecompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. *Cell Res* 2024;**34**:830–45.

Yin Z, Shen Y. On the dimensionality of word embedding. *Adv Neural Inf Process Syst* 2018;**31**.

Yuan X, Zhan Z, Zhang Z *et al.* Cell ontology guided transcriptome foundation model. *Adv Neural Inf Process Syst* 2024.

Zhang Z, Han X, Liu Z *et al.* Ernie: enhanced language representation with informative entities. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 1441. Association for Computational Linguistics, 2019.

Zhao Y, Cai H, Zhang Z *et al.* Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nat Commun* 2021;**12**:5261.

Zhou Y, Sharpee TO. Hyperbolic geometry of gene expression. *iScience* 2021;**24**:102225.

Bioinformatics, 2025, 41, 1–10
https://doi.org/10.1093/bioinformatics/btaf595
Review