bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Revealing a coherent cell state landscape across single cell datasets with CONCORD

Qin Zhu[1]*, Zuzhi Jiang[2], Matt Thomson[3], Zev Gartner[1,4,5]*

[1] University of California San Francisco, Department of Pharmaceutical Chemistry, San Francisco, CA 94158, USA

[2] Tetrad Graduate Program, University of California San Francisco, San Francisco, CA 94158, USA.

[3] Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 94158, USA

[4] Chan Zuckerberg Biohub, University of California San Francisco, San Francisco, CA 94158, USA

[5] Center for Cellular Construction, University of California San Francisco, CA 94158, USA

*Correspondence to Q. Zhu at qin.zhu@ucsf.edu and Z. J. Gartner at zev.gartner@ucsf.edu

Author information

Qin Zhu, qin.zhu@ucsf.edu

Zuzhi Jiang, zuzhi.Jiang@ucsf.edu

Matt Thomson, mthomson@caltech.edu

Zev J. Gartner, zev.gartner@ucsf.edu

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Abstract

Resolving the intricate structure of the cellular state landscape from single-cell RNA sequencing (scRNAseq) experiments remains an outstanding challenge, compounded by technical noise and systematic discrepancies—often referred to as batch effects—across experimental systems and replicate. To address this, we introduce CONCORD (COntrastive learNing for Cross-dOmain Reconciliation and Discovery), a self-supervised contrastive learning framework designed for robust dimensionality reduction and data integration in single-cell analysis. The core innovation of CONCORD lies in its probabilistic, dataset- and neighborhood-aware sampling strategy, which enhances contrastive learning by simultaneously improving the resolution of cell states and mitigating batch artifacts. Operated in a fully unsupervised manner, CONCORD generates denoised cell encodings that faithfully preserve key biological structures, from fine-grained distinctions among closely related cell states to large-scale topological organizations. The resulting high-resolution cell atlas seamlessly integrates data across experimental batches, technologies, and species. Additionally, CONCORD's latent space capture biologically meaningful gene programs, enabling the exploration of regulatory mechanisms underlying cell state transitions and subpopulation heterogeneity. We demonstrate the utility of CONCORD on a range of topological structures and biological contexts, underscoring its potential to extract meaningful insights from both existing and future single-cell datasets.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Introduction

Cells express thousands of genes to perform specialized functions and maintain homeostasis. These gene expression patterns are governed by gene regulatory networks and cell–cell interactions, confining cell distributions to a low-dimensional state space (or ‘state landscape’) within the high-dimensional gene expression space[1,2]. Waddington’s landscape is often invoked as a conceptual model of the state landscape, illustrating how cells traverse branching developmental trajectories toward distinct fates[3]. Modern single-cell technologies such as scRNA-seq capture snapshots of these processes, revealing diverse topological structures - including clusters, bifurcations, convergences and loops—indicating more nuanced structures than that depicted in Waddington’s original diagram[4]. Accurately capturing these structures, which are typically unknown a priori, is therefore essential for elucidating developmental processes, homeostatic regulation, and disease progression.

Dimensionality reduction, or representation learning in machine learning, is a common approach for uncovering such structures. By projecting high-dimensional data into a lower-dimensional space, key structural patterns become more tractable to visualize and analyze. However, standard techniques—such as principal component analysis, non-negative matrix factorization[5], and factor analysis[6]—tend to overemphasize broad cell type separation, overlook subtle states, and can confound processes like differentiation with cell cycle progression. Compounding these challenges is the continuous nature of cell state changes, yet scRNA-seq profiles typically sample only discrete snapshots. Researchers often integrate multiple datasets capturing distinct developmental stages, biological replicates, perturbations, or other conditions to assemble a more comprehensive view of the underlying state landscape. Yet, such integration efforts are often impeded by technical and biological batch effects, which can obscure genuine biological signals. Although an array of batch-correction tools—such as Harmony[7], Scanorama[8], Seurat[9], scVI[10], LIGER[11] and

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

MNN[12] —have been developed, they can distort the biological structure by over- or under-correcting batch effects[13], and many face scalability issues when applied to massive atlas-level datasets.

Recently, contrastive learning—a technique proven effective in image and natural language processing[14-16]—has gained attention in single-cell genomics[17-23]. This approach learns meaningful representations by contrasting similar (positive) cells against dissimilar (negative) ones within mini-batches—small, randomly sampled subsets of the dataset that are iteratively contrasted during training. By building an understanding of what makes cells similar and different, contrastive learning extracts features underly distinct cellular states, thereby improving clustering and annotation[17-19,21-23]. Some methods adopt a supervised strategy, using known cell type labels to define positive and negative pairs[22,23]. However, this approach requires substantial manual annotation effort and limits generalization to novel cell states or continuous trajectories. Others employ self-supervised strategies, treating each cell in a mini-batch as a negative for all others[17-20]. Yet, with negatives uniformly drawn from the global distribution, these methods often emphasize global differences - such as major cell type clusters - while underutilizing the fine resolution of single-cell data to detect subtle cellular states. Although recent work suggests that mini-batch composition heavily influences contrastive outcomes[24], existing single-cell approaches have yet to fully explore the implications and generalizability of this insight.

Applying contrastive learning to multiple datasets presents additional challenges: naively contrasting cells across different batches can amplify batch-specific artifacts rather than isolating biologically relevant variation. Various techniques - such as generative adversarial networks (GANs)[20,25,26], unsupervised domain adaptation via backpropagation[27], and conditional variational autoencoders (CVAEs)[28] - have been employed to mitigate batch effects. However, in a contrastive learning setting, maximizing differences between dissimilar cells

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

conflicts with the goal of minimizing dataset-specific differences, frequently resulting in incomplete removal of batch effects. This dilemma raises a key question: can contrastive learning fully capture cellular diversity while minimizing batch artifacts?

In this work, we introduce CONCORD (COntrastive learNing for Cross-dOmain Reconciliation and Discovery), a novel contrastive learning framework designed to denoise and reveal the detailed structure of the cellular state landscape across single or multiple datasets. Central to CONCORD is a probabilistic, neighborhood-aware and dataset-aware sampling strategy that harnesses mini-batch composition to achieve two critical goals: enhancing the resolution of cell states in the latent space, and removing batch-associated biases and noise. Neighborhood-aware sampling prompts the model to contrast nearby cells, thereby capturing subtle distinctions among closely related states. Meanwhile, dataset-aware sampling ensures comparisons occur within the same dataset, prioritizing biological variation while minimizing dataset-specific artifacts. Crucially, these strategies are unified within a single probabilistic sampler, enabling contrastive learning to simultaneously improve resolution and mitigate batch effects. Using simulated and real datasets, we demonstrate that CONCORD produces high-resolution, denoised encodings that robustly capture diverse structures—including loops, trajectories, trees, and specialized cell states—reflecting bona fide biological processes even when the data originate from multiple technologies, time points, or species.

Results

The CONCORD framework

Analysis of single-cell sequencing data suggest that gene expression is not randomly sampled; rather, the mechanism of gene regulation imposes strong constraints, producing dynamically changing gene co-expression patterns

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

reflected as intricate structures in the low dimensional embedding of cells[1,2,4]. For example, adult cells at homeostasis often form discrete clusters corresponding to stable cell types or states, with adjacent clusters representing closely related cell states (Figure 1A, left). In contrast, developmental or pathological contexts—such as early embryogenesis, tissue repair, or tumorigenesis—tend to exhibit trajectories branching from progenitor cells to multiple terminal fates, with semi-stable transitional states forming denser clusters (Figure 1A, middle). Loop structures are also frequently observed[4,29], typically associated with cyclic gene expression programs, such as those governing cell cycle progression[26] (Figure 1A, right). Despite these rich patterns, standard dimensionality reduction methods like principal component analysis (PCA) or nonnegative matrix factorization (NMF) capture only partial representations of the cell state landscape, either oversimplifying complex structures or disproportionately emphasizing certain features while obscuring others.

We hypothesized that a latent representation method capable of intrinsically modeling relationships among cells at the level of gene co-expression programs could yield a more comprehensive view of the state landscape. Recent evidence suggests that self-supervised contrastive learning, particularly SimCLR-type approaches[14], significantly improves clustering and cell-type classification performance[18-20]. In this framework, positive pairs are typically generated by perturbing a given cell—such as by randomly masking gene expression values—to create two augmented versions. Contrastive learning then operates through two complementary forces: differentiation and alignment. By contrasting each cell against all others in the mini-batch, the model identifies features that differentiate between distinct cell states. Simultaneously, aligning the two masked versions of the same cell encourages the model to capture gene co-expression patterns, rather than relying on the expression of individual genes[30]. Theoretical results demonstrate that contrastive learning with random masking and ReLU networks effectively captures sparse signals while suppressing noise[30] (see Methods). In the context of gene expression, these sparse signals correspond to gene co-

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

expression programs, where sets of co-expressed genes define functional cell states[6,31,32]. We incorporated similar conditions—LeakyReLU activation and random masking—into our model architecture (see Methods, Supplemental Figure 1) and observed an encoding that effectively captures gene co-expression programs with suppressed universal features and noise (Figure 1B). By prioritizing sets of co-expressed genes over individual gene expression fluctuations, this representation is also inherently more robust to dropouts – a common artifact in single-cell RNA-seq[33].

In standard contrastive learning, training proceeds iteratively: in each step, a mini-batch is drawn from the dataset, cells are augmented, encodings are computed, and a contrastive loss is applied based on similarities between augmentations and differences among samples. This cycle repeats over multiple iterations, gradually refining the latent encodings. Since contrastive objectives are computed within each mini-batch[24], the sampling strategy – which dictates mini-batch composition - plays a pivotal role in shaping the learned representation. Most methods rely on uniform sampling across the entire dataset, leading to two key limitations. First, uniform sampling emphasizes broad differences, such as major cell types, while underrepresenting rare subpopulations, leading to poor resolution of fine-scale cellular states (Figure 1B). Second, mixing cells from different datasets within the same mini-batch can amplify dataset-specific differences, inadvertently encoding batch effects rather than isolating biologically meaningful variation.

To address the first issue, we developed a neighborhood-aware sampler, inspired by k-nearest-neighbor (kNN) based sampling[24]. In this approach, cells are sampled probabilistically from both the global distribution and local neighborhoods (Figure 1B). Local sampling, guided by a coarse graph approximation of the cellular state landscape, compels the model to contrast cells against their neighbors, allowing it to capture subtle differences between closely

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

related states. Meanwhile, global sampling preserves a broad perspective of major cell types, ensuring the model robustly encodes large-scale distinctions. By iteratively presenting the model with local neighborhoods (e.g., T cells in one mini-batch, epithelial cells in another) alongside the global distribution, the model allocates capacity to represent both large-scale distinctions and nuanced local details, leading to improved resolution of biologically structures in the learned latent space (Figure 1B).

When applied to a single dataset, contrastive learning effectively distills biological variation into a meaningful latent representation (Figure 1C). However, with uniform sampling across multiple datasets, both biological and dataset-specific variations are encoded, leading to latent spaces that separate by dataset as well as cell type (Figure 1D). To mitigate these batch effects, we additionally developed a dataset-aware sampling approach. This method leverages the fact that contrastive learning primarily captures biologically variable signals within each mini-batch, while suppressing global, non-informative features and noises[30] (Figure 1B). By restricting mini-batches to a single dataset, the model learns only biological variations at each training step, mirroring the effect of applying contrastive learning within a single dataset. Meanwhile, batch-associated biases and noise are progressively reduced through mini-batch shuffling, which randomly interleaves mini-batches from different datasets. If the contrastive loss induces dataset-specific encodings in one mini-batch, subsequent mini-batches from other datasets disrupt and overwrite these encodings due to their inconsistency across datasets. As a result, only biologically meaningful signals, such as gene co-expression patterns, persist across training iterations, leading to a latent space that accurately reflects cell type and state distributions with significantly reduced batch effects (Figure 1E). In cases where datasets have minimal or no overlap, a "leaky" dataset-aware sampler enables soft alignment without imposing artificial harmonization, allowing for a flexible integration scheme that generalizes to both fully and partially overlapping datasets (Figure 1F). Unlike traditional batch-correction approaches used in machine learning

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

models—such as conditional variational autoencoders (CVAEs)[28], unsupervised domain adaptation by backpropagation[27], and generative adversarial networks (GANs)[20,25,26] —which often struggle in contrastive learning due to competing objectives, CONCORD directly leverages the contrastive objective and training process to learn a batch-effect-mitigated latent representation.

Both the neighborhood-aware and dataset-aware samplers follow a unified design principle: probabilistically structuring mini-batches to balance global biological variations with local and dataset-specific differences. In the neighborhood-aware sampler, most cells are selected to maintain the global distribution of cell types, ensuring stable representation of large-scale biological variation, while a smaller proportion is drawn from local neighborhoods to enhance fine-grained resolution. Similarly, in the dataset-aware sampler, mini-batches primarily consist of cells from a single dataset to minimize dataset-specific biases, with a smaller subset incorporating cross-dataset samples to retain biologically meaningful variation. By integrating both samplers, we establish a joint probabilistic sampling framework, where the likelihood of selecting a given cell reflects the combined probabilities of dataset-aware and neighborhood-aware sampling (Figure 1F). This generalized sampling strategy enables both robust dataset integration and improved resolution within a self-supervised contrastive learning framework, forming the foundation of the CONCORD model (Supplemental Figure 1). Notably, we found that a simple two-layer encoder with Leaky ReLU activation effectively captures rich biological structures without requiring additional layers or a decoder, significantly reducing training data requirements and improving robustness. This minimalistic architecture also enhances interpretability, as a Leaky ReLU activation after the first linear layer enables the second layer to encode context-dependent gene expression programs, effectively compressing complex cell states into a moderate number of latent neurons.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

CONCORD learns denoised latent representations that preserve underlying structures

Recovering biologically meaningful insights from single-cell data relies on preserving key features of the gene expression state space, including both its geometric organization and topological structure. To assess whether CONCORD meets this requirement, we evaluated its performance on both simulated and real-world single-cell datasets. While existing simulation packages (e.g., splatter[34]) generate discrete cell-type clusters, they fail to recapitulate the diverse structures observed in real biological data, including branching trajectories, loops, and multi-scale hierarchies. To address this limitation, we developed a custom simulation workflow capable of producing a wide range of structures that closely resemble real biological data, while allowing flexible control over noise distributions and batch effects (Figure 2A).

In parallel, we designed an evaluation pipeline that integrates both geometric and topological metrics to quantitatively assess the quality of latent representations (Figure 2B). Traditional benchmarking frameworks, such as scIB[35], primarily focus on biological label preservation and batch correction, but provide less insight into whether a representation retains the intrinsic geometrical relationships and topological structures of the original data. To fill this gap, we incorporated geometric metrics, including trustworthiness[36] and distance correlation, alongside topological data analysis (TDA) methods based on persistent homology and Betti numbers (Figure 2B). Trustworthiness quantifies how well local neighborhood structures in high-dimensional space are preserved in the latent representation, with high scores indicating that nearby cells in the original data remain neighbors in the learned embedding. Persistent homology, in contrast, provides a more global perspective by tracking the emergence and persistence of topological features—such as connected components (i.e. clusters; Betti-0), loops (Betti-1), and enclosed voids (Betti-2)—across varying distance scales. These features are visualized in persistence diagrams and Betti

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

curves, where stable structures appear as long-lived features in the persistence diagram and extended plateaus in the Betti curve, whereas transient, noise-induced features vanish quickly.

We first evaluated CONCORD on a simple simulated dataset consisting of three well-separated clusters corrupted by cluster-specific Gaussian noise (Figure 2C, Supplemental Figure 2A). We compared CONCORD against a diverse set of dimensionality reduction and embedding methods, including diffusion map, NMF, Factor Analysis (FA), FastICA, Latent Dirichlet Allocation (LDA), ZIFA, scVI, and PHATE. While visualizations with the UMAP algorithm[37] provide a qualitative comparison (Figure 2C), all benchmarking statistics were computed directly on the latent space, ensuring that evaluations reflect the true structure learned by each model rather than artifacts introduced by UMAP. Many methods, including NMF, FA, FastICA, ZIFA, and scVI, struggled to fully separate the clusters, while others, such as PHATE, introduced spurious trajectory-like structures (Figure 2C). By contrast, CONCORD cleanly distinguished the three clusters and produced a denoised distance matrix that closely matched the ground truth (Figure 2C). Persistent homology further confirmed these results: CONCORD's persistence diagram correctly identified the expected three-cluster topology (Betti-0 = 3), showing a stable plateau in the Betti curve that closely matched the noise-free dataset.

For more complex structures, such as a self-connecting trajectory with three loops and multiple branching points, CONCORD was the only method that accurately recovered the full underlying structure (Figure 2D, Supplemental Fig. 2B). Other methods either collapsed trajectories into clusters (e.g., DiffusionMap, NMF, FactorAnalysis, FastICA, LDA) or failed to generate a stable persistence diagram, likely due to excessive noise retention (e.g., ZIFA, scVI). While PHATE produced a relatively smooth latent space (Figure 2D), its Betti curve analysis

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

revealed only a single persistent loop rather than the expected three, indicating an incomplete preservation of topological structure.

We summarized these findings using geometric and topological metrics and found that CONCORD consistently outperformed alternative methods across a broad range of simulated structures (Figure 2E, 2F). Importantly, despite its denoising capability, CONCORD preserves relative noise levels in the latent space, as demonstrated by the strong correlation between latent variance and input variance in the cluster simulation (Figure 2E). This is particularly important because biological noise—such as transcriptional variability—often carries biologically relevant information, distinguishing dynamic cellular states from stable ones. Furthermore, CONCORD consistently maintains high trustworthiness scores across neighborhood sizes from 10 to 100, underscoring its ability to preserve local structure over a significant radius (Supplemental Figure 2C, 2D). In contrast, many alternative methods exhibit sharp declines in trustworthiness, indicating a loss of fine-scale geometric relationships in the latent space.

To investigate how neighborhood-aware sampling influences performance, we simulated a three-level branching tree that mimics hierarchical tissue-cell-type-cell-state relationships (Figure 2G, Supplemental Figure 2E). Without the neighborhood-aware sampler, the model failed to resolve sub-branches; by contrast, higher local sampling uncovered more subtle differences (Supplemental Figure 2E, 2F), leading to substantially improved resolution in the latent representation (Figure 2G). However, excessive local enrichment had drawbacks. When intra-kNN sampling probability exceeded 0.6, global distinctions at tissue and cell-type levels were suppressed (Figure 2G, Supplemental Figure 2F), consistent with our earlier findings that contrastive learning harmonizes inconsistent global signals. Based on these results, we

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

recommend an intra-kNN sampling probability below 0.5 to maintain a balance between fine-grained resolution and global structure preservation.

CONCORD learns a coherent, batch-effect-mitigated latent representation

Batch effects often manifest as global signals that impact all cells within a dataset but differ across datasets. When mini-batches are enriched for cells from a single dataset, these dataset-specific signals rapidly diminish during training due to their inconsistent nature (Figure 1E, Figure 3A). This batch-correction mechanism fundamentally differs from conventional methods, which typically assume a specific model of batch distortion and force the alignment between datasets. By contrast, CONCORD prioritizes the learning of biologically informative and coherent variation over inconsistent batch signals, making minimal assumptions about the nature of batch effects. As a result, it preserves biological structures more effectively while mitigating technical artifacts.

To evaluate this, we tested CONCORD on a simulated dataset containing five clusters corrupted by both batch effects and noise. Among all methods tested, CONCORD was the only approach that accurately identified all clusters; alternative methods struggled to separate closely related cell types or introduced alignment artifacts, such as the “ring” structure produced by Scanorama[8]. Notably, applying a standard contrastive learning sampler without dataset-aware sampling resulted in pronounced batch effects (Figure 3A). While significantly reducing noise, CONCORD preserved cluster-specific variance differences rather than over-smoothing them—a critical aspect often overlooked by batch-correction algorithms (Supplemental Figure 3A). Moreover, CONCORD demonstrated consistent denoising and batch-effect correction performance across different noise distributions (e.g. Gaussian, negative-binomial, Poisson) and batch-effect models (e.g. batch-specific gene expression differences, sequencing depth variability, and dropout rate differences), underscoring its independence from specific noise and batch effect assumptions (see Methods; data not shown).

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Beyond aligning discrete clusters, a major challenge in single-cell data integration is constructing a unified representation from datasets capturing related but distinct conditions—such as different developmental stages, tissues, species, or perturbations. Many integration techniques struggle in cases where datasets exhibit limited overlap in cell states, particularly those relying on matched clusters (e.g., Harmony[7]) or mutual nearest neighbors (e.g., Seurat[9], Scanorama[8], MNN[12]). Aligning datasets solely based on overlapping states can also distort the structure when these regions are sparse or imbalanced. As the number of datasets increases, these limitations become more pronounced, often leading to fragmented or warped embeddings.

To systematically assess these challenges, we simulated batch effects across a range of structures: clusters (Supplemental Figure 3B), trajectories (Figure 3B, Supplemental Figure 3C), loops (Figure 3C, Supplemental Figure 3D), and trees (Figure 3D, Supplemental Figure 3E). On a trajectory simulation where batches were fully overlapping, most methods achieved some degree of alignment, though several exhibited partial or suboptimal correction (e.g., scVI, LIGER). However, as overlapping regions were progressively reduced, most methods deteriorated. Some failed to align batches entirely (e.g., LIGER, Harmony), while others introduced spurious structures, such as Scanorama producing an artificial loop instead of a linear trajectory, or scVI collapsing trajectories with partial overlap into a single cluster. Across all conditions, CONCORD consistently recovered the underlying structure with significantly reduced noise, even when shared cell states were minimal or absent (Figure 3A–D, Supplemental Figure 3B–E, Supplemental Table 1).

CONCORD’s robustness likely arises from its focus on learning biologically meaningful gene co-expression programs (see Methods, Figure 1B, Supplemental Figure 2A, 2B, 2E) rather than forcing direct cell-state alignment

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

across batches—a fundamental distinction from many existing batch-correction approaches. By encoding gene co-expression programs, CONCORD positions cells with similar transcriptomic states together in the latent space, eliminating the need for explicit reference points. This property allows CONCORD to better preserve biological signals, as reflected in its high biological label conservation scores in the scIB benchmarking metrics (Figure 3E, 3F). However, because CONCORD does not explicitly merge batches, it achieves high but not the highest batch-correction scores. Interestingly, CONCORD more effectively preserves local distance relationships than global ones, leading to high trustworthiness scores but lower global distance correlations, which affects its overall geometric score ranking (Figure 3F, Supplemental Table 1). This outcome reflects an intrinsic trade-off in manifold learning, where maintaining local neighborhood structure is often favored over enforcing global distance relationships[38,39]. Nonetheless, CONCORD consistently ranks among the top-performing methods for topological structure preservation and biological label conservation, as well as in overall rankings (Figure 3F). These results indicate that CONCORD effectively mitigates batch effects while preserving the intrinsic structure of the data. Given the complexity of real single-cell datasets, these findings suggest that CONCORD provides a reliable, generalizable solution for both dimensionality reduction and batch-effect mitigation, even in cases with uncertain underlying structures or limited batch overlap.

CONCORD aligns whole-organism developmental atlases and resolves high-resolution lineage trajectories

To evaluate CONCORD's ability to capture biologically meaningful structures across datasets generated by different technologies, we benchmarked it against popular batch correction methods using lung and pancreas single-cell atlases[35] (Supplemental Figure 4A, B). While CONCORD excelled at identifying discrete cell-type clusters, these datasets and their annotations do not reflect the hierarchical organization of biological systems and lack the continuous or

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

branching trajectories characteristic of developmental and pathological processes. A rigorous assessment of integration methods requires datasets with diverse structures and ground-truth lineage annotations, enabling a fine-scale evaluation of both dataset integration and biological structure conservation.

Caenorhabditis elegans (*C. elegans*) embryogenesis provides an ideal benchmark, as it follows a well-characterized, nearly invariant lineage tree from the fertilized egg to the 558 cells present at hatching[40]. This lineage tree is highly conserved in closely related species, such as *C. briggsae*[41]. Packer et al. previously generated a lineage-resolved single-cell atlas of *C. elegans* embryogenesis[42], which was recently expanded by Large et al. to include over 200,000 *C. elegans* cells and 190,000 *C. briggsae* cells[41]. The authors annotated progenitor and terminal cell fates through extensive subset-specific projections, clustering and fluorescent reporter imaging. These datasets serve as ideal resources for evaluating whether dimensionality reduction and integration algorithms can faithfully reconstruct and align developmental trajectories across species.

We first tested CONCORD on the *C. elegans* dataset[42] and observed that its UMAP embedding recapitulated known developmental trajectories (Supplemental Figure 5A). The effect of neighborhood-aware sampling mirrored trends observed in our simulations: moderate local enrichment improved resolution of subtle differences among neurons, while excessive local sampling disrupted the global structure (Supplemental Fig. 5A, B). This is because local sampling enhances mini-batches with underrepresented subpopulations, while adequate global sampling is necessary to capture broader variation (Supplemental Figure 5C).

Running CONCORD on the larger, cross-species dataset yielded a unified developmental atlas that closely matches original cell-type and lineage

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

annotations (Figure 4A). The embedding effectively separated broad cell classes (Supplemental Figure 6A) and arranged cells along continuous trajectories from early progenitors to terminal states in alignment with estimated embryo time (Figure 4B). While UMAP is well-suited for visualizing complex latent representations, the resulting latent representation proved too complex for two-dimensional UMAP to disentangle fully— likely due to extensive lineage bifurcations and convergences—necessitating a three-dimensional UMAP to mitigate trajectory cross-overs (Figure 4A). Running CONCORD with or without a decoder yielded similar UMAP embeddings (Figure 4B), and we present the with-decoder version due to slightly less entangled trajectories. However, we strongly encourage readers to explore the interactive 3D UMAP visualizations (https://qinzhu.github.io/Concord_documentation/galleries/cbce_show/#__tabbed_1_1), which offer a more detailed view of developmental trajectories than can be captured in static 2D projections.

To systematically compare integration methods, we moved beyond UMAP-based visual assessment and directly analyzed the 300-dimensional latent space produced by each algorithm. Standard benchmarking tools, such as scIB[35], could not scale to this dataset's size and complexity, requiring us to adapt previous evaluation strategies. To assess species alignment, we computed species composition within randomly sampled latent space neighborhoods, evaluating whether each neighborhood's species fraction matched expectations, similar to the kBET algorithm[43] (Figure 4C). Given the dataset's time-dependent sampling biases, with *C. elegans* cells progressively increasing over time due to experimental bias or true biological variation, we stratified the analysis by developmental time bins. If species were well-aligned at a fine scale, neighborhood species fractions should remain close to expected proportions with minimal variation. Indeed, CONCORD maintained this pattern, whereas other methods exhibited alignment failures: Scanorama produced neighborhoods dominated by a single species (fractions of 0 or 1, also evident in the UMAP in Figure 4B), scVI and LIGER showed excessive species fraction variability,

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

indicating incomplete species alignment, and Seurat and Harmony deviated significantly from expected proportions, suggesting local density mismatches between species.

Besides species alignment, another key benchmark is how well each method preserves biological structure without sacrificing resolution. As seen in UMAP visualizations, Harmony and scVI successfully aligned species but lost resolution, obscuring subtypes and lineage trajectories (Figure 4B). Only Seurat and CONCORD successfully recovered divergent terminal fates, and among them, only CONCORD maintained continuous, fine-grained trajectories from progenitors to terminal cell types (Figure 4B, D). Seurat's reciprocal PCA approach captured later-stage variation but overshadowed subtle early lineage differences, likely because later stages contain more cells and greater transcriptional diversity. In contrast, CONCORD's neighborhood-aware sampling preserved both early and late states, distributing representational capacity across the entire developmental spectrum.

The rich lineage and cell-type annotations in this dataset enabled us to map the *C. elegans* lineage tree onto UMAP embeddings, revealing how progenitor lineages give rise to distinct terminal fates[40,44] (Supplemental 6B, C). Strikingly, on many occasions, CONCORD's UMAP trajectories aligned with underlying lineage subtrees. For example, ciliated amphid neurons ASE, ASJ, AUA—arising from AB-derived progenitors—formed branching trajectories in CONCORD's UMAP that closely mirrored the actual lineage sub-tree (Figure 4D). In contrast, scVI and Seurat introduced large gaps between parent and daughter lineages or obscured distinct terminal fates (Figure 4D). Beyond capturing lineage relationships, CONCORD's latent activation patterns effectively distinguished terminal cell states (Figure 4E). Hierarchical clustering of CONCORD's latent space revealed two distinct ASE subpopulations, characterized by differential expression of GCY receptors—gcy-5, gcy-19, gcy-22 in one group, and gcy-6,

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

gcy-7, gcy-14 in the other (Figure 4E). These receptors distinguish ASE-left (ASEL) from ASE-right (ASER) neurons, which are morphologically symmetric but functionally asymmetric in their salt-sensing responses[45,46].

To systematically assess the preservation of lineage structures in the latent space, we repeated the lineage distance analysis from Packer et al.[42], correlating latent space distances with lineage distances for AB lineages which give rise to ~70% of terminal embryonic cells (Figure 4F). Consistent with the original study, where transcriptome distances correlated with lineage distances, we found that CONCORD's latent distances exhibited strong correlation with lineage distances, even in early generations where transcriptomic differences were minimal. This demonstrates CONCORD's ability to capture subtle, progressive molecular changes underlying lineage bifurcations. Notably, CONCORD significantly outperformed other integration methods in overall correlation (Figure 4G), underscoring its potential for trajectory inference in developmental studies.

The improvement in resolving fine-scale structures became even more pronounced when zooming into subsets of cells. In early embryonic cells, Scanorama, Harmony, and scVI failed to fully align species or lost resolution, whereas CONCORD revealed extensive lineage bifurcation patterns (Supplemental Figure 6D). On muscle formation, CONCORD showed the MS, C, and D lineages converge into sub-branches of body wall muscle, positioned from the head (anterior) to the tail (posterior) in an orientation reflecting genuine spatial gene expression gradients (Figure 4H, Supplemental Figure 6E). CONCORD also captured rare lineage convergence events, such as ABplp/ABprp-derived muscle integrating with MS-derived counterparts to form intestinal muscles (mu_int) (Figure 4H). Pharyngeal development, involving complex branching and convergence of AB- and MS-derived cells, was likewise resolved by CONCORD (e.g., pm3–5 deriving from both AB and MS lineages, and pm1–2, 6–8 specific to AB/MS lineage), whereas scVI and Seurat UMAPs

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

displayed fewer fine-grained details of such lineage convergence and divergence (Figure 4I, Supplemental Figure 6F). Importantly, all of these zoom-in analyses were performed directly on the global latent space learned by CONCORD, without requiring subset-specific variable gene (VEG) selection and re-alignment – steps that are often recommended for other methods.

Finally, CONCORD demonstrated superior efficiency and scalability on this 400,000-cell, 20-batch dataset (Figure 4J). It completed integration in ~30 minutes on an NVIDIA A100 GPU, outperforming LIGER, Seurat, and scVI, which required several hours and significantly more memory. These results confirm that CONCORD excels in scalability and accuracy, making it an ideal tool for reconstructing complex developmental trajectories.

CONCORD captures cell cycle and differentiation trajectories in mammalian intestinal development

Unlike *C. elegans*, where embryonic divisions are largely governed by maternally supplied mRNAs[47], mammalian development involves producing many more cells in each lineage, often coupling cell cycle activity with ongoing differentiation. To evaluate whether CONCORD can recover these intertwined processes, we applied it to a single-cell atlas of embryonic mouse intestinal development[48]. This dataset encompasses multiple developmental stages, spatial segments, and diverse cell types (Figure 5A). Moreover, it contains batches enriched for specific populations—such as mesenchymal cells—providing a challenging scenario where incomplete coverage across batches could hinder integration and resolution of fine-grained cell states.

CONCORD not only integrated these datasets significantly faster than most other methods (data not shown), but it also resolved finer substructures within each cell type (Figure 5A, Supplemental Fig. 7A, Supplemental Figure 8). Batches

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

derived from different developmental stages and enrichment strategies merged seamlessly into a cohesive cell atlas (Supplemental Figure 7A). Notably, CONCORD revealed numerous loop-like patterns within each cell type, corresponding to cell cycle progression (Figure 5A-D). Erythrocytes—known to lack proliferative capacity—did not form such loops, matching established biology (Supplemental Figure 8). As discussed earlier, because CONCORD can capture complex structures, a standard 2D UMAP does not fully represents these loops or breaks trajectories. We therefore focused our analysis on 3D UMAPs, (Figure 5A-D, Supplemental Figure 7B) and strongly recommend readers explore an interactive visualization of CONCORD embedding (https://qinzhu.github.io/Concord_documentation/galleries/huycke_show/).

In intestinal epithelial cells, CONCORD not only delineated detailed cell subtype structures, including rare enteroendocrine cells (EECs), but also revealed two parallel differentiation trajectories, each forming its own cell cycle loop (Figure 5B). Coloring the UMAP by zonation suggests that these trajectories correspond to proximal and distal segment identities, reinforced by the expression of regional markers such as *Bex1* and *Onecut2*[49](Figure 5B). While zonation in adult mouse and human intestines has been documented[49], our data suggest that segment-specific epithelial subsets arise as early as embryonic day 13.5 (E13.5), implying that CONCORD can detect subtle positional signatures well before maturity.

In enteric nervous system (ENS) cells, CONCORD uncovered a clear cell cycle loop alongside two bifurcating differentiation branches (Figure 5C). Morarach et al.[50] previously characterized ENS progenitor states in mouse embryos, showing progressive cell cycle phases and bifurcations marked by genes such as *Etv1* and *Bnc2*. CONCORD successfully recapitulated these branching patterns, appropriately anchoring them at a neuroblast population defined by *Cck*, which arises from *Sox10*$^+$ progenitor cells. The convergence of the two branches appears to be driven by a set of neuronal maturation genes that are broadly

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

expressed across both trajectories (Supplemental Figure 7C). Notably, CONCORD was the only method that preserved both the cell cycle loop and the bifurcation, whereas other methods introduced discontinuities or misplaced bifurcation points (Supplemental Figure 8).

Mesenchymal cells comprise a major fraction of this dataset, yet prior studies primarily focused on the *Pdgfra*$^+$ mesenchymal subset, which plays a key role in villus formation[48]. In CONCORD's embedding, we not only recapitulated the previously described *Pdgfra*$^+$ trajectory (Supplemental Figure 8), but also uncovered highly heterogeneous substructures within the *Pdgfra*$^-$ mesenchymal subset. This population, along with smooth muscle cells, formed four consecutive cell cycle loops, each marked by distinct expression of *Ebf1*, *Slit2*, *Kit*, and *Acta2*, with varying degrees of gradual transitions between loops (Figure 5D). Interestingly, *Ebf1*, *Slit2* have been previously associated with mesenchymal cell multipotency[51,52], while Kit marks interstitial cells of Cajal (ICC) and its progenitors[53,54]. Unlike traditional approaches, where cell cycle effects often obscure cell-state annotations, CONCORD preserves both cell cycle and differentiation signals, allowing for the identification of these previously unrecognized subpopulations.

The rich structures revealed by CONCORD is also evident in its latent space, where each neuron is activated in one or more cell populations—ensuring no latent capacity is wasted (Figure 5E). In contrast, the PCA results from Seurat's integrated dataset compress variation into only a few principal components, while scVI leaves nearly half of its latent dimensions minimally utilized. Furthermore, CONCORD's latent space can be interrogated with gradient-based attribution methods[55], a standard approach for interpreting the contribution of input features to a neuron's activation (Figure 5F). Concretely, this technique produces an attribution map for a given input cell, indicating each gene's contribution to that cell's activation pattern. This framework enables single-cell or cell-state–level

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

gene explanations in a context-dependent manner. For instance, neuron 46 (N46) is activated in both epithelial cells and the ENS cells (Figure 5F), but its activity is attributed to two distinct sets of highly co-expressed genes for each context (Figure 5F, Supplemental Figure 7C). In the epithelial context, the top genes for N46 activation are differentially expressed in goblet cells and enriched for glycosylation pathways linked to goblet cell function. Meanwhile, in the ENS context, the activating genes are highly co-expressed in late-stage neurons with neuronal functions. Neither gene set shows strong expression outside its respective context, indicating that CONCORD latent captures interpretable, context-dependent gene co-expression programs.

Discussion

Contrastive learning has proven to be a powerful approach in image and language processing, and recent work has begun to explore its adaptation for single-cell sequencing data[17-23]. While some approaches use a supervised paradigm[22,23], relying on annotated positive and negative samples to guide cell-type separation, such strategies require substantial labeling and can be prone to inaccuracies—particularly in single-cell contexts where cell identities may not be completely defined. Other methods[17-20] adopt self-supervised contrastive learning but employ a uniform sampling scheme that struggles to capture local neighborhood structure and may inadvertently amplify batch effects. Additionally, many of these methods use deep neural network architectures that require large training datasets, limiting their applicability to single-cell studies with moderate sample sizes.

In this work, we adapt self-supervised contrastive learning as a general framework for dimensionality reduction and data integration in single-cell analysis, operating in a fully unsupervised manner without requiring extensive priors or large-scale training corpora. Central to our approach is a probabilistic sampler that strengthens the contrastive learning process. Rather than drawing

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

cells entirely at random, our sampler selects from both global distributions and local neighborhoods, enabling the model to observe broad variability across cell types while also focusing on finer distinctions within closely related states. When combined with masking and contrastive objectives, this design effectively learns both large-scale and fine-grained gene co-expression patterns, filters out inconsistent noise, and produces a high-resolution latent representation that significantly improve downstream analyses, such as clustering, trajectory inference, and UMAP visualizations.

A second key contribution of CONCORD is its capacity to integrate data across multiple batches—an essential feature for large-scale efforts like the Human Cell Atlas[56]. Conventional approaches often assume overlapping states or particular batch-distortion models, leading to over- or under-correction that can disrupt continuous trajectories or obscure biological signals[13]. In contrast, CONCORD mitigates batch effects directly through its sampling strategy: the dataset-aware sampler ensures that mini-batches primarily contain cells from a single dataset, allowing the model to learn biologically meaningful variation while minimizing dataset-specific artifacts. As a result, CONCORD generates a batch-effect-mitigated latent representation even when overlap in cell states is minimal or absent. Rather than stitching datasets together based on mutual-nearest-neighbor-based ‘anchors’ or predefined reference points—a strategy prone to accumulating distortions—CONCORD aligns cells based on underlying gene co-expression structures, preserving biological continuity and structural integrity. Moreover, its generalized sampling framework unifies neighborhood-aware and dataset-aware sampling, enabling both effective data integration and enhanced resolution. Crucially, these improvements demonstrate that contrastive learning can be significantly enhanced without requiring complex model architectures or additional training objectives, but rather through principled sampling and training.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Simulations and real-data benchmarks illustrate CONCORD's effectiveness in capturing fine-scale manifold structures. In simulated datasets, with or without batch effects, CONCORD consistently produces a denoised latent space that preserves critical topological and geometric characteristics. Importantly, these findings extend to real biological data. For instance, in embryonic atlases of *C. elegans* and *C. briggsae*, whose lineage trees are highly conserved, CONCORD not only integrated data from multiple species and batches spanning embryogenesis, but also accurately reconstructed intricate fate bifurcations and lineage convergences. Other methods either fell short in aligning these datasets, overlooked detailed structures, or fractured continuous trajectories. While narrowing the dataset to specific cell types and re-running integration can partially address these issues, it fragments any organism-wide perspective. CONCORD, by contrast, offers a unified, high-resolution atlas of developmental processes, enabling detailed tracing from progenitor cells to terminal states. Similar advantages emerged in mouse intestinal development, where CONCORD revealed complex hierarchies of subtypes, spatial zonation patterns, and cell cycle loops—all within a single integrated analysis. Unlike traditional workflows that remove cell cycle effects, CONCORD's embedding preserves both cell cycle loops and differentiation trajectories, allowing researchers to investigate the interplay between cell cycle and fate specification. In the intestine, analysis of the CONCORD latent revealed that differentiation trajectories are better described as complex topological structures where loops and cylinders are linked together by filaments. Furthermore, the simplicity of the CONCORD model enhances interpretability. Gradient-based attribution techniques allow users to identify the gene programs driving latent neuron activations, conferring mechanistic insights at single-cell or cell-type resolution.

The current CONCORD implementation employs a simple two-layer neural network, which significantly reduces memory requirements, improves computational efficiency, and scales to atlas-scale datasets while remaining robust for small datasets. However, its effectiveness may be limited in cases

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

where gene co-expression structures are highly distorted—such as in whole-cell versus single-nucleus scRNA-seq—or when dropout rates vary significantly across technologies. As with most neural network models, hyperparameter selection is crucial. While CONCORD's minimalistic design reduces the number of tunable parameters, key factors—including neighborhood size, neighborhood enrichment probability, masking fraction, and contrastive temperature—can substantially influence the final latent representation. Our benchmarking analysis, along with prior studies[57], offers guidance on optimizing these parameters to balance local resolution with global structural preservation and we offer a set of default parameters that have performed well across multiple contexts. Additionally, detailed tutorials on the CONCORD website (https://qinzhu.github.io/Concord_documentation/) support users in making informed hyperparameter choices. Future improvements to our sampling strategy, particularly neighborhood-aware sampling, could incorporate adaptive sampling probabilities that scale with distance, eliminating the need for predefined neighborhood radii and enrichment probabilities. In its current form, CONCORD supports a decoder for gene-level batch correction and a classifier for tasks such as cell-type annotation or doublet detection. Early results suggest these modules benefit from CONCORD's robust latent representation, though additional optimization and benchmarking are required.

While the present benchmarks focused on single-cell RNA-seq, we have observed promising outcomes in other single-cell modalities, including spatial transcriptomics and scATAC-seq. Owing to its domain-agnostic design and generalized sampler framework, CONCORD may extend beyond single-cell transcriptional biology, offering a flexible, powerful approach for other single-cell modalities and beyond.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Methods

Self-supervised contrastive learning and sparse coding

We implemented CONCORD in PyTorch, using a self-supervised contrastive learning approach inspired by SimCLR[14] and SimCSE[15], but with a unique dataset- and neighborhood-aware sampler design. The primary goal is to obtain a denoised, batch-effect-corrected latent representation that accurately captures essential characteristics of the underlying cell-state landscape. To achieve this, we apply the Normalized Temperature-scaled Cross Entropy Loss (NT-Xent loss)[14,15,58] directly to masking-augmented cell representations, following the design principles of unsupervised SimCSE[15].

A key feature of our approach is the integration of LeakyReLU activation functions with random masking augmentation, which enhances the capture of highly correlated gene co-expression patterns while suppressing spurious noise. Theoretical results[30] suggest that, if gene expression data x can be approximated by a sparse coding model:

$$x = Mz + \varepsilon$$

where Mz represents the sparse signal with $\|z\|_0 = \tilde{O}(1)$, and ε denotes noise, then contrastive learning provably recovers the sparse features when trained with ReLU networks and random masking augmentation.

This sparse coding framework provides a generalized approach for modeling gene expression compared to widely used methods such as non-negative matrix factorization (NMF), principal component analysis (PCA), and factor analysis. Unlike these methods, sparse coding:

- Does not enforce orthogonality on M (as in PCA),
- Does not require non-negativity constraints (as in NMF),
- Does not assume a probabilistic generative model (as in factor analysis).

At its core, this framework assumes an intrinsic low-rank latent structure, where gene co-expression programs define functional states—a phenomenon widely observed in single-cell transcriptomic studies[6,31,32]. Since contrastive learning has

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

been theoretically proven to capture such low-rank structures—or co-expressed gene programs—we adopt a similar architecture in CONCORD and observe an effective encoding of gene co-expression modules alongside noise suppression. By relaxing constraints on orthogonality, non-negativity, and Gaussian noise assumptions, the model may better capture diverse gene regulatory programs, including those that do not conform to conventional constraints. Additionally, random masking-based contrastive learning enhances robustness to dropout artifacts and improves biological interpretability, ensuring that the learned latent space faithfully represents gene programs underlying both discrete cell states and continuous trajectories.

Model architecture

A schematic overview of the architecture is illustrated in Supplemental Figure 1 and described below:

1. Encoder:
   The encoder takes randomly masked gene-expression vectors as input and produces a latent encoding. By default, it is a two-layer fully connected network (though the user can adjust the number of layers and neurons). Optionally, a feature masking module with learnable weights can be placed before the encoder to weight genes differently and encourage sparse feature usage. Typically, single-cell data are normalized by total count and log-transformed prior to training, though we have observed that CONCORD remains robust to various normalization schemes as long as no additional negative or zero values are introduced, and normalization is applied consistently across datasets. During training, each cell is masked at the gene level with a user-defined dropout probability (commonly 0.3–0.6). Two masked versions of the same cell are fed through the encoder, and their resulting embeddings are used in the contrastive loss.

2. Layer normalization and activation:

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Each linear layer in the encoder is followed by layer normalization and a user-selectable activation function (default: Leaky ReLU). Layer normalization normalizes activations within each sample, making it more robust to cross-batch variability than batch normalization[59] (which is also implemented in CONCORD and available to user).

3. Contrastive objective:

   We employ the Normalized Temperature-scaled Cross Entropy Loss (NT-Xent loss)[14,15,58] operating on mini-batches of N cells. Each cell is randomly masked and encoded twice, yielding two different latent representations: z_i and z_j, then the contrastive loss is:

   $$l_{i,j} = -\log(\frac{\exp(sim(z_i, z_i')/\tau)}{\sum_{k=1}^{2N} 1_{[k\neq i]}\exp(sim(z_i, z_k)/\tau)})$$

   where $sim(z_i, z_i') = \frac{z_i^T z_j}{\|z_i\|\|z_j\|}$ is the cosine similarity, and τ (default 0.5) is a customizable temperature hyperparameter that controls the extent of local separation and global uniformity of the embeddings[57].

4. Decoder and classifier (optional):

   For tasks requiring gene-expression reconstruction, an optional decoder can be attached to the latent embeddings. In such cases, the latent vector is combined with a learnable dataset embedding before being fed to the decoder. This design separates batch information from the core latent representation, preventing re-introduction of dataset-specific artifacts.

   A classification head (i.e., a multi-layer perceptron) with cross-entropy loss can be appended to the encoder for downstream tasks such as cell-type prediction and doublet detection. The classifier can be trained on a pre-trained encoder or jointly with the encoder to simultaneously guide cell-type separations in the latent space. However, the latter approach may impose a strong prior on the latent structure, potentially disrupting the

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

continuity of cellular trajectories. To mitigate overfitting in classification tasks, a standard train-validation split with early stopping can be employed.

Dataset and neighborhood-aware probabilistic sampler

The key innovation of CONCORD is the probabilistic mini-batch sampler, which determines how cells are grouped and contrasted during training. Instead of sampling cells uniformly at random, we introduce a generalizable sampler framework that simultaneously (i) enriches local neighborhoods and (ii) focuses each mini-batch primarily on a single dataset. This approach improves resolution of detailed biological structures and yields a coherent, batch-effect-mitigated cell atlas.

We begin by approximating the global data manifold with a k-nearest neighbors (kNN) graph, where k is user-defined (and typically moderately large). This graph can be built from normalized gene-expression data, a PCA projection, or a preliminary CONCORD embedding. For large datasets, we employ the Faiss library[60] for efficient neighbor searches. The kNN graph then facilitates sampling at a user-specified local neighborhood enrichment probability, p_{kNN} (default 0.3).

To construct mini-batches that are both dataset- and neighborhood-enriched, we partition each mini-batch into four subsets—in-dataset neighbors, in-dataset global samples, out-of-dataset neighbors, and out-of-dataset global samples (Figure 1F). We randomly select a “core sample” from one dataset to anchor both neighborhood selection and dataset enrichment. Then, each partition is sampled based on P_d (default 0.95) and P_{kNN} as follows:

- In-dataset neighbors ($P_d \times P_{kNN}$): cells from the same dataset as the core sample and lie in its KNN neighborhood.
- In-dataset global samples ($P_d \times (1 - P_{kNN})$): cells from the same dataset but sampled uniformly from the entire data.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

- Out-of-dataset neighbors ($(1 - P_d) \times P_{kNN}$):): cells from other datasets that appear in the core sample's local kNN neighborhood.
- Out-of-dataset global samples ($(1 - P_d) \times (1 - P_{kNN})$): Cells from other datasets, drawn uniformly across the entire data.

The sampler uses vectorized operations in PyTorch and NumPy for neighbor retrieval, shuffling, and batch construction, thereby minimizing computational overhead.

Model training

During each training epoch, CONCORD generates mini-batches using the dataset and neighborhood-aware sampler, and randomly shuffles the mini-batches. The Adam optimizer[61] was used to optimize the core contrastive objective (NT-Xent), with optional loss terms—including reconstruction loss (MSE) for the decoder, cross-entropy loss for classification, or feature-masking penalties (L1/L2)—which can be incorporated as needed. By default, all loss components are weighted equally but can be adjusted based on user-defined preferences. A learning rate scheduler gradually reduces the learning rate to enhance stability and convergence.

Simulation pipeline

We developed a versatile simulation pipeline to generate synthetic single-cell gene expression data with diverse underlying structures. Unlike conventional simulators that predominantly produce discrete clusters, our pipeline accommodates a broad range of topologies, including linear trajectories, branching trees, loops, and intersecting paths frequently observed in real single-cell datasets.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

In the first stage the state simulator constructs data according to a user-defined structure:

- Clusters: Cells form discrete groups characterized by unique gene programs, optionally including shared or ubiquitously expressed genes.
- Trajectories: Cells exhibit gradual shifts in gene expression, emulating cell differentiation processes.
- Loops and intersecting paths: Continuous trajectories that close into loops or intersect, representing cyclic biological processes.
- Trees: Hierarchical, branching lineages representing progenitor-to-terminal fate differentiation, configurable by branching factor and tree depth.

With the chosen structure, the pipeline first generates a *noise-free* data matrix with customizable cell and gene numbers. Expression values are then sampled from selected distributions (e.g., Normal, Poisson, Negative Binomial), introducing realistic variability and dropout patterns. Users can precisely control parameters including mean baseline expression, dispersion (noise level), dropout probability, and can enforce non-negativity or integer rounding of the generated values.

In the second step, an optional batch simulator introduces dataset-specific technical variability. This stage enables simulation of batch effects through scaling factors, differential sampling rates, batch-specific gene subsets, and expression-dependent dropout mechanisms. Multiple simulated batches are then concatenated into a single dataset, with customizable proportions and varying degrees of batch overlap to mimic real-world sampling scenarios.

By combining diverse gene expression structures with realistic noise models and customizable batch effects, this simulation pipeline can approximate a broad

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

spectrum of biological and technical scenarios. As such, it provides a powerful testbed for benchmarking data-integration techniques, trajectory-inference algorithms, and manifold-learning methods under controlled yet biologically realistic conditions.

Benchmarking pipeline

To comprehensively evaluate the performance of CONCORD and other dimensionality reduction or data integration methods, we designed a robust benchmarking pipeline that integrates geometric, topological, biological, and batch-mixing metrics. This multifaceted assessment framework consists of the following components:

1. Topological assessments:
   To quantify the preservation of intrinsic topological features, we employed persistent homology analysis implemented via Giotto-TDA[62]. Persistent homology captures structural properties of the data across multiple scales, using Vietoris-Rips complexes constructed over increasing radii to generate persistence diagrams and Betti curves. Persistence diagrams reveal the lifespan of topological features such as connected components (Betti-0), loops (Betti-1), and voids (Betti-2). We summarized these diagrams through Betti curves and compared their mode—representing the most persistent Betti number across all scales—to the known topological ground truths. Additionally, we computed the entropy of Betti curves to quantify the stability and complexity of the inferred topology, with lower entropy reflecting stable and distinct structures, and higher entropy indicating noisy or unstable topologies. These metrics were scaled between 0 and 1 using min-max normalization to facilitate comparisons among methods.

2. Geometric assessments:

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

We evaluated the preservation of geometric relationships by calculating distance correlations between embeddings and the corresponding noise-free reference data, averaging Pearson, Spearman, and Kendall's tau correlations to robustly quantify global geometric similarity. For local neighborhood preservation, we employed trustworthiness[36], a metric assessing how faithfully high-dimensional neighborhood structures are maintained in lower-dimensional embeddings. Trustworthiness scores range from 0 (poor preservation) to 1 (perfect preservation), and we computed average trustworthiness scores across neighborhood sizes (k-values) from 10 to 100 in increments of 10. Additionally, we visualized trustworthiness as a function of k to reveal how each method performs at different local scales. In cluster simulations with cluster-specific noise, we further assessed the correlation of variance between the latent embedding and the noisy input data, quantifying how accurately each method preserves relative noise levels.

3. Batch mixing and biological label conservation:
We adopt established metrics from the scIB-metrics package[35] to systematically evaluate biological label conservation and batch mixing. Biological label conservation was quantified using metrics such as Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI), which measure the correspondence between known biological labels and inferred clusters. Batch mixing quality was assessed using silhouette scores and graph metrics to determine how effectively embeddings integrate cells from different batches or datasets.

A key limitation of the scIB pipeline is that the pipeline does not fully accommodate the hierarchical and continuous nature of many biological systems. Consequently, for simulations with continuous trajectories or loops, we first apply Leiden clustering to noise-free data to define “clusters” as ground truth, or use “branch” labels as a proxy for cell states

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

in tree simulations. Under these conditions, the scIB metrics are applied in a more coarse-grained manner, offering an approximate assessment in these more complex scenarios.

Data Availability

The human lung and pancreas datasets were compiled by Luecken et al.[35], and obtained from the scIB-metrics website (https://scib-metrics.readthedocs.io/en/stable/notebooks/lung_example.html) and the Open Problems in Single-Cell Analysis website (https://openproblems.bio/datasets/openproblems_v1/pancreas), respectively. The *C. elegans* embryogenesis atlas was downloaded from the Gene Expression Omnibus (GEO) (www.ncbi.nlm.nih.gov/geo) under accession code GSE126954. The joint *C. elegans* and *C. briggsae* dataset was obtained via email request from the authors of Large et al[41]. The mouse intestinal developmental atlas was acquired from GEO under accession code GSE233407.

Code Availability

Concord is available at https://github.com/Gartner-Lab/Concord under a Creative Commons Attribution 4.0 International License. All benchmarking codes to generate results in this manuscript are deposited to https://github.com/Gartner-Lab/Concord_benchmark. Full documentation of Concord can be found at https://github.com/Gartner-Lab/Concord_documentation and available online at: https://qinzhu.github.io/Concord_documentation/.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

References

1 Wagner, D. E. & Klein, A. M. Lineage tracing meets single-cell omics: opportunities and challenges. *Nature Reviews Genetics* **21**, 410-427 (2020).

2 Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331-338 (2017).

3 Waddington, C. H. *The strategy of the genes*. (Routledge, 2014).

4 Flores-Bautista, E. & Thomson, M. Unraveling cell differentiation mechanisms through topological exploration of single-cell developmental trajectories. *bioRxiv*, 2023.2007. 2028.551057 (2023).

5 Johnson, J. A. *et al.* Inferring cellular and molecular processes in single-cell data with non-negative matrix factorization using Python, R and GenePattern Notebook implementations of CoGAPS. *Nature protocols* **18**, 3690-3731 (2023).

6 Kunes, R. Z., Walle, T., Land, M., Nawy, T. & Pe'er, D. Supervised discovery of interpretable gene programs from single-cell data. *Nature Biotechnology* **42**, 1084-1095 (2024).

7 Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature methods* **16**, 1289-1296 (2019).

8 Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature biotechnology* **37**, 685-691 (2019).

9 Stuart, T. *et al.* Comprehensive integration of single-cell data. *cell* **177**, 1888-1902. e1821 (2019).

10 Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature methods* **15**, 1053-1058 (2018).

11 Welch, J. D. *et al.* Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**, 1873-1887. e1817 (2019).

12 Haghverdi, L., Lun, A. T., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology* **36**, 421-427 (2018).

13 Zhang, Z. *et al.* Recovery of biological signals lost in single-cell batch integration with CellANOVA. *Nature Biotechnology*, 1-17 (2024).

14 Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. in *International conference on machine learning*. 1597-1607 (PMLR).

15 Gao, T., Yao, X. & Chen, D. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821* (2021).

16 He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729-9738.

17 Richter, T., Bahrami, M., Xia, Y., Fischer, D. S. & Theis, F. J. Delineating the effective use of self-supervised learning in single-cell genomics. *Nature Machine Intelligence*, 1-11 (2024).

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

18 Ciortan, M. & Defrance, M. Contrastive self-supervised clustering of scRNA-seq data. *BMC bioinformatics* **22**, 280 (2021).

19 Wang, J., Xia, J., Wang, H., Su, Y. & Zheng, C.-H. scDCCA: deep contrastive clustering for single-cell RNA-seq data based on auto-encoder network. *Briefings in Bioinformatics* **24**, bbac625 (2023).

20 Zhao, B., Song, K., Wei, D.-Q., Xiong, Y. & Ding, J. scCobra allows contrastive cell embedding learning with domain adaptation for single cell data integration and harmonization. *Communications Biology* **8**, 233 (2025).

21 Yang, M. *et al.* Contrastive learning enables rapid mapping to multimodal single-cell atlas of multimillion scale. *Nature Machine Intelligence* **4**, 696-709 (2022).

22 Heimberg, G. *et al.* A cell atlas foundation model for scalable search of similar human cells. *Nature*, 1-3 (2024).

23 Heryanto, Y. D., Zhang, Y.-z. & Imoto, S. Predicting cell types with supervised contrastive learning on cells and their types. *Scientific Reports* **14**, 430 (2024).

24 Yang, Z. *et al.* in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 3057-3069.

25 Goodfellow, I. *et al.* Generative adversarial networks. *Communications of the ACM* **63**, 139-144 (2020).

26 Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nature methods* **16**, 715-721 (2019).

27 Ganin, Y. & Lempitsky, V. in *International conference on machine learning.* 1180-1189 (PMLR).

28 Sohn, K., Lee, H. & Yan, X. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems* **28** (2015).

29 Riba, A. *et al.* Cell cycle gene regulation dynamics revealed by RNA velocity and deep-learning. *Nature communications* **13**, 2865 (2022).

30 Wen, Z. & Li, Y. in *International Conference on Machine Learning.* 11112-11122 (PMLR).

31 Jiang, J. *et al.* D-SPIN constructs gene regulatory network models from multiplexed scRNA-seq data revealing organizing principles of cellular perturbation response. *BioRxiv*, 2023.2004. 2019.537364 (2024).

32 Kotliar, D. *et al.* Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *Elife* **8**, e43803 (2019).

33 Alaqeeli, O. A comparison of dropout rate of three commonly used single cell RNA-sequencing protocols. *Biotechnology & Biotechnological Equipment* **38**, 2379837 (2024).

34 Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome biology* **18**, 174 (2017).

35 Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nature methods* **19**, 41-50 (2022).

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

36 Venna, J. & Kaski, S. in *International conference on artificial neural networks.* 485-491 (Springer).

37 McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv. *arXiv preprint arXiv:1802.03426* **10** (2018).

38 Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9** (2008).

39 Tenenbaum, J. B., Silva, V. d. & Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *science* **290**, 2319-2323 (2000).

40 Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. The embryonic cell lineage of the nematode Caenorhabditis elegans. *Developmental biology* **100**, 64-119 (1983).

41 Large, C. R. *et al.* Lineage-resolved analysis of embryonic gene expression evolution in C. elegans and C. briggsae. *bioRxiv*, 2024.2002. 2003.578695 (2024).

42 Packer, J. S. *et al.* A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution. *Science* **365**, eaax1971 (2019).

43 Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A. & Theis, F. J. A test metric for assessing single-cell RNA-seq batch correction. *Nature methods* **16**, 43-49 (2019).

44 Tintori, S. C., Nishimura, E. O., Golden, P., Lieb, J. D. & Goldstein, B. A transcriptional lineage of the early C. elegans embryo. *Developmental cell* **38**, 430-444 (2016).

45 Ortiz, C. O. *et al.* Searching for neuronal left/right asymmetry: genomewide analysis of nematode receptor-type guanylyl cyclases. *Genetics* **173**, 131-149 (2006).

46 Yu, S., Avery, L., Baude, E. & Garbers, D. L. Guanylyl cyclase expression in specific sensory neurons: a new family of chemosensory receptors. *Proceedings of the National Academy of Sciences* **94**, 3384-3387 (1997).

47 Koreth, J. & van den Heuvel, S. Cell-cycle control in Caenorhabditis elegans: how the worm moves from G1 to S. *Oncogene* **24**, 2756-2764 (2005).

48 Huycke, T. R. *et al.* Patterning and folding of intestinal villi by active mesenchymal dewetting. *Cell* **187**, 3072-3089. e3020 (2024).

49 Zwick, R. K. *et al.* Epithelial zonation along the mouse and human small intestine defines five discrete metabolic domains. *Nature Cell Biology* **26**, 250-262 (2024).

50 Morarach, K. *et al.* Diversification of molecularly defined myenteric neuron classes revealed by single-cell RNA sequencing. *Nature neuroscience* **24**, 34-46 (2021).

51 Derecka, M. *et al.* (American Society of Hematology Washington, DC, 2017).

52 Chen, C.-P., Wang, L.-K., Chen, C.-Y., Chen, C.-Y. & Wu, Y.-H. Placental multipotent mesenchymal stromal cell-derived Slit2 may regulate

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

macrophage motility during placental infection. *Molecular Human Reproduction* **27**, gaaa076 (2021).

53 Al-Shboul, O. A. The importance of interstitial cells of cajal in the gastrointestinal tract. *Saudi Journal of Gastroenterology* **19**, 3-15 (2013).

54 Torihashi, S. *et al.* Blockade of kit signaling induces transdifferentiation of interstitial cells of cajal to a smooth muscle phenotype. *Gastroenterology* **117**, 140-148 (1999).

55 Ancona, M., Ceolini, E., Öztireli, C. & Gross, M. Gradient-based attribution methods. *Explainable AI: Interpreting, explaining and visualizing deep learning*, 169-191 (2019).

56 Atlas, H. C. Human cell atlas. *Human cell atlas sequences first 250K developmental cells* (2018).

57 Wang, F. & Liu, H. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 2495-2504.

58 Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems* **29** (2016).

59 Lei Ba, J., Kiros, J. R. & Hinton, G. E. Layer normalization. *ArXiv e-prints*, arXiv: 1607.06450 (2016).

60 Douze, M. *et al.* The faiss library. *arXiv preprint arXiv:2401.08281* (2024).

61 Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

62 Tauzin, G. *et al.* giotto-tda:: A topological data analysis toolkit for machine learning and data exploration. *Journal of Machine Learning Research* **22**, 1-6 (2021).

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Acknowledgement

We thank Dr. Junhyong Kim, Dr. John Murray, and Dr. Honesty Kim for providing valuable feedback on the manuscript. We also thank the authors of Large et al.[41] for sharing the *C. elegans* and *C. briggsae* dataset, with special acknowledgement to Dr. Christopher R. L. Large for facilitating data access. Additionally, we thank members of the Gartner Lab for providing critical discussions and support and help with testing early versions of CONCORD. We thank ChatGPT for assistance with code refinement and annotation. This research was supported by grants from the NIH (U01CA199315, R01GM135462, R01DK126376, U01DK103147, and R33CA247744) and the UCSF Center for Cellular Construction (DBI-1548297), an NSF Science and Technology Center. Q.Z. is supported by a Cancer Research Institute Immuno-Informatics Postdoctoral Fellowship (CRI5054). Z.J.G. is a Chan Zuckerberg BioHub San Francisco Investigator.

Author Contributions

Q.Z. and Z.J.G. conceived the project. Q.Z. designed and implemented the method. Q.Z. conducted benchmarking analyses with assistance from Z.J. M.T. provided critical feedback and methods for topological data analysis. Z.J.G. supervised the project. Q.Z. and Z.J.G. wrote the manuscript. All authors reviewed and edited the manuscript.

Competing Interests

ZJG is an author on patents associated with sample multiplexing and ZJG is an equity holder and advisor to Provenance Bio.

Materials & Correspondence

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Correspondence to Q. Zhu (qin.zhu@ucsf.edu) and Z. J. Gartner (zev.gartner@ucsf.edu).

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figures

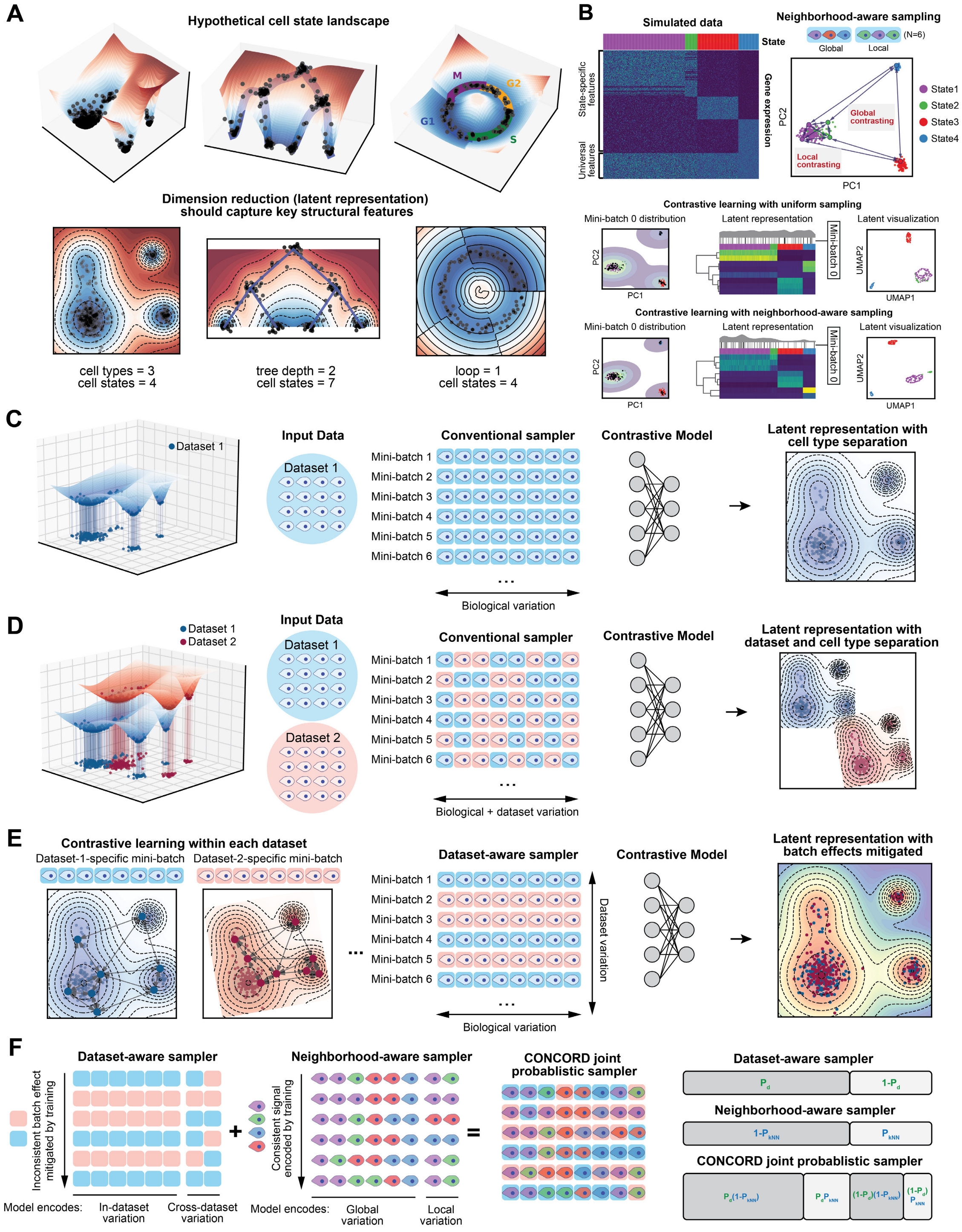
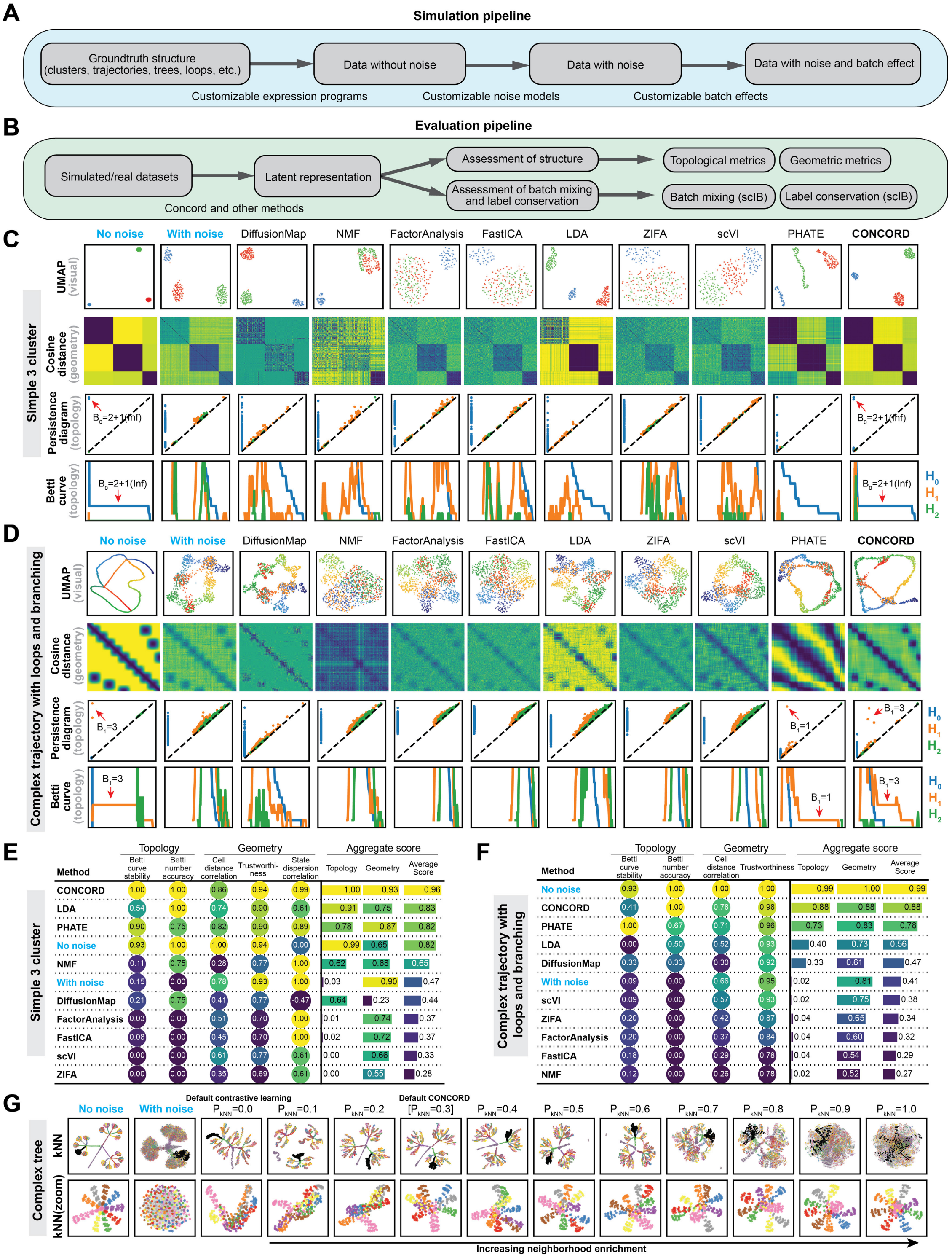


Figure 1. The CONCORD sampler enables contrastive learning to generate a high-resolution, batch-effect-mitigated latent representation of scRNA-seq data.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(A) Illustration of hypothetical cell state landscapes and corresponding dimensionality-reduced representations that capture key structural features of the landscape. (**B**) Comparison of neighborhood-aware and uniform sampling and their impact on contrastive learning in a simulated four-cell-state dataset. The heatmap shows the actual simulated expression. For each sampling scheme, PCA plots are color-coded by cell state, with black points indicating cells selected in a representative mini-batch, accompanied by density curves illustrating their distribution. Latent heatmaps display the representations learned using uniform and neighborhood-aware sampling, with black lines marking cells included in the selected mini-batch and density plots depicting their distribution. The resulting UMAP embeddings computed from the latent representations for each sampling method are also shown. (**C**) Contrastive learning performed in a single batch with the conventional sampler, which draws cells uniformly from the entire dataset to form mini-batches. (**D**) When applying standard contrastive learning to multiple datasets (represented by the blue or pink background), contrasting cells from different datasets within the same mini-batch amplifies dataset-specific biases, which is manifested in the latent embeddings. (**E**) CONCORD mitigates these dataset-specific artifacts by predominantly contrasting cells within each dataset and randomly shuffling mini-batches for each training epoch. (**F**) The CONCORD sampling framework. A "leaky" dataset-aware sampler addresses minimal or absent overlap between datasets and can be combined with the neighborhood-aware sampler to support both data integration and enhanced resolution. This is achieved by a joint probabilistic sampling framework, where the likelihood of selecting a given cell reflects the combined probabilities of dataset-aware (P_d) and neighborhood-aware sampling (P_{kNN}).

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



E Simple 3 cluster

	Topology		Geometry			Aggregate score		
Method	Betti curve stability	Betti number accuracy	Cell distance correlation	Trustworthiness	State dispersion correlation	Topology	Geometry	Average Score
CONCORD	1.00	1.00	0.86	0.94	0.99	1.00	0.93	0.96
LDA	0.54	1.00	0.74	0.90	0.61	0.91	0.75	0.83
PHATE	0.90	0.75	0.82	0.90	0.89	0.78	0.87	0.82
No noise	0.93	1.00	1.00	0.94	0.00	0.99	0.65	0.82
NMF	0.11	0.75	0.28	0.77	1.00	0.62	0.68	0.65
With noise	0.15	0.00	0.78	0.93	1.00	0.03	0.90	0.47
DiffusionMap	0.21	0.75	0.41	0.77	-0.47	0.64	0.23	0.44
FactorAnalysis	0.03	0.00	0.51	0.70	1.00	0.01	0.74	0.37
FastICA	0.08	0.00	0.45	0.70	1.00	0.02	0.72	0.37
scVI	0.00	0.00	0.61	0.77	0.61	0.00	0.66	0.33
ZIFA	0.00	0.00	0.35	0.69	0.61	0.00	0.55	0.28

F Complex trajectory with loops and branching

	Topology		Geometry		Aggregate score		
Method	Betti curve stability	Betti number accuracy	Cell distance correlation	Trustworthiness	Topology	Geometry	Average Score
No noise	0.93	1.00	1.00	1.00	0.99	1.00	0.99
CONCORD	0.41	1.00	0.78	0.98	0.88	0.88	0.88
PHATE	1.00	0.67	0.71	0.96	0.73	0.83	0.78
LDA	0.00	0.50	0.52	0.93	0.40	0.73	0.56
DiffusionMap	0.33	0.33	0.30	0.92	0.33	0.61	0.47
With noise	0.09	0.00	0.66	0.95	0.02	0.81	0.41
scVI	0.09	0.00	0.57	0.93	0.02	0.75	0.38
ZIFA	0.20	0.00	0.42	0.87	0.04	0.65	0.34
FactorAnalysis	0.20	0.00	0.37	0.84	0.04	0.60	0.32
FastICA	0.18	0.00	0.29	0.78	0.04	0.54	0.29
NMF	0.12	0.00	0.26	0.78	0.02	0.52	0.27

Figure 2. Benchmarking CONCORD and other dimensionality reduction methods across diverse structures.

(A) Simulation pipeline for generating data structures. The pipeline first produces a noise-free gene expression matrix based on a user-defined structure, then introduces noise following a specified noise model, and finally

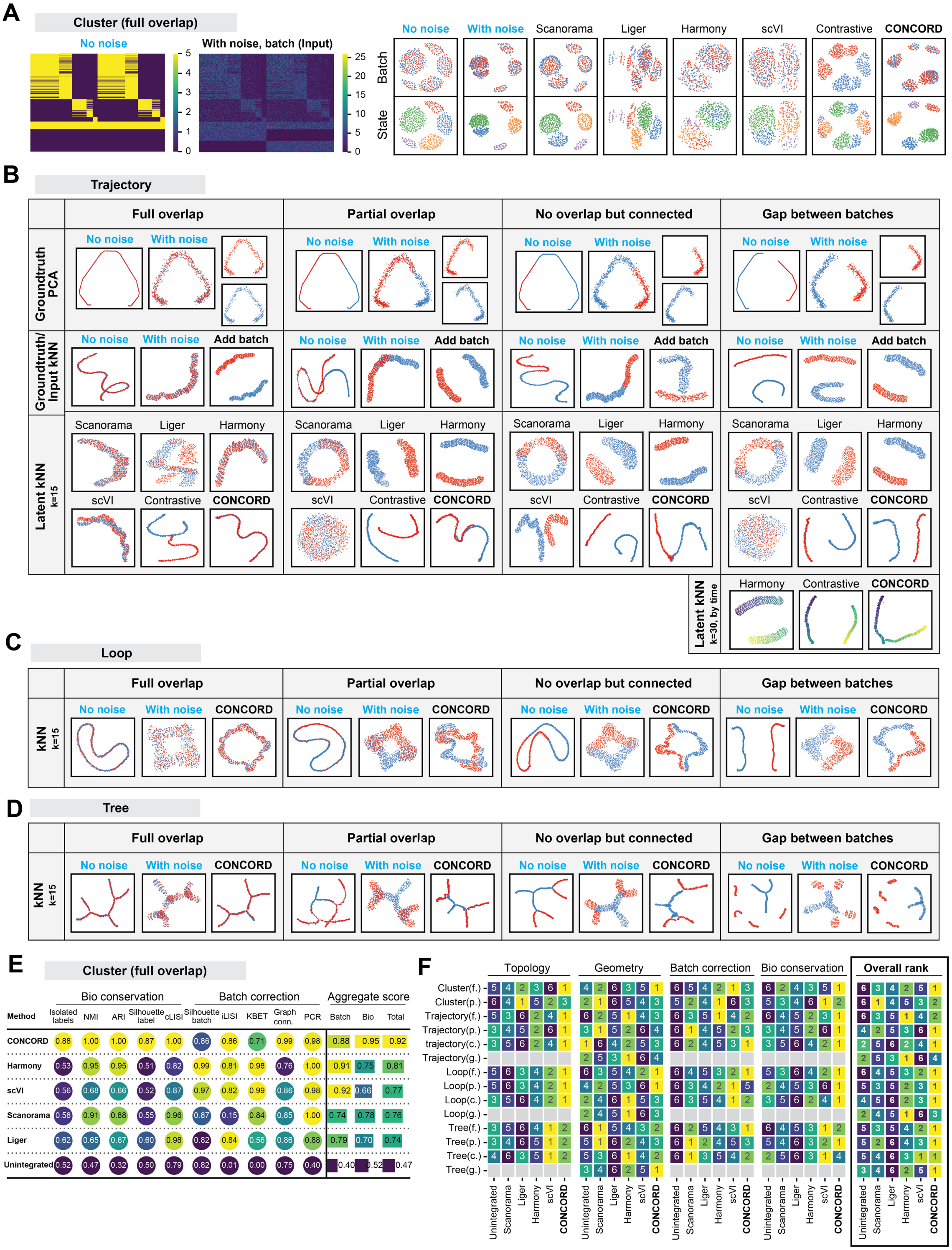
bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

applies batch effects in various forms. (**B**) Evaluation pipeline. Using the simulated datasets, the latent representations produced by each method was compared with the noise-free ground truth to assess how well topological and geometric features are preserved. For cluster simulations, we further evaluate the correlation of cluster-specific variances in the noisy data versus the latent space. Metrics from the scIB[35] package were incorporated for evaluating conservation of biological labels and harmonization of batch effects. (**C**) Performance on simulated clusters, highlighting the resulting UMAP visualization, cosine distance matrices, persistence diagrams, and Betti curves for CONCORD and other methods. In the persistent homology analysis, the H_0 point representing infinity was excluded from the persistence diagram and curve. (**D**) Performance on a complex trajectory with 3 loops, highlighting the same diagnostic plots as C. (**E**) Summary table for the three-cluster simulation, listing key topological and geometric evaluation metrics. (**F**) Table summarizing the methods' performance on the complex trajectory-loop simulation. (**G**) KNN graph visualization of latent embeddings from each method on a complex tree simulation, with zoomed-in views of the darkened region highlighting detail on one of the branches.

1213

1214

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



Method	Bio conservation					Batch correction					Aggregate score		
	Isolated labels	NMI	ARI	Silhouette label	cLISI	Silhouette batch	iLISI	KBET	Graph conn.	PCR	Batch	Bio	Total
CONCORD	0.88	1.00	1.00	0.87	1.00	0.86	0.86	0.71	0.99	0.98	0.88	0.95	0.92
Harmony	0.53	0.95	0.95	0.51	0.82	0.99	0.81	0.98	0.76	1.00	0.91	0.75	0.81
scVI	0.56	0.68	0.66	0.52	0.87	0.97	0.82	0.99	0.86	0.98	0.92	0.66	0.77
Scanorama	0.58	0.91	0.88	0.55	0.96	0.87	0.15	0.84	0.85	1.00	0.74	0.78	0.76
Liger	0.62	0.65	0.67	0.60	0.98	0.82	0.84	0.56	0.86	0.88	0.79	0.70	0.74
Unintegrated	0.52	0.47	0.32	0.50	0.79	0.82	0.01	0.00	0.75	0.40	0.40	0.52	0.47

	Topology						Geometry						Batch correction						Bio conservation						Overall rank					
	Unintegrated	Scanorama	Liger	Harmony	scVI	CONCORD	Unintegrated	Scanorama	Liger	Harmony	scVI	CONCORD	Unintegrated	Scanorama	Liger	Harmony	scVI	CONCORD	Unintegrated	Scanorama	Liger	Harmony	scVI	CONCORD	Unintegrated	Scanorama	Liger	Harmony	scVI	CONCORD
Cluster(f.)	5	4	2	3	6	1	4	2	6	3	5	1	6	5	4	2	1	3	6	2	4	3	5	1	6	3	4	2	5	1
Cluster(p.)	6	4	1	5	2	3	2	1	6	5	4	3	5	2	4	1	6	3	5	3	4	6	1	2	6	1	4	5	3	2
Trajectory(f.)	5	3	6	2	4	1	6	2	5	1	4	3	6	3	5	1	4	2	6	4	5	2	3	1	6	3	5	1	4	2
Trajectory(p.)	3	4	5	2	6	1	3	1	5	2	6	4	6	1	4	5	2	3	3	4	5	2	6	1	4	2	5	3	6	1
trajectory(c.)	3	5	6	2	4	1	1	6	4	2	5	3	6	2	4	5	3	1	3	5	6	4	2	1	2	5	6	2	4	1
Trajectory(g.)							2	5	3	1	6	4													2	5	3	1	6	4
Loop(f.)	5	6	3	4	2	1	6	3	5	4	2	1	6	4	5	3	2	1	5	3	6	4	2	1	6	4	5	3	2	1
Loop(p.)	5	6	3	4	2	1	4	2	5	3	6	1	6	3	4	2	1	5	2	4	5	3	6	1	5	3	5	2	3	1
Loop(c.)	3	5	6	4	2	1	2	4	6	1	5	3	6	3	5	4	1	2	3	5	6	2	4	1	4	5	6	2	3	1
Loop(g.)							2	4	5	1	6	3													2	4	5	1	6	3
Tree(f.)	3	5	6	4	1	2	6	1	5	4	3	2	6	2	5	4	3	1	6	4	5	3	1	2	5	3	5	4	2	1
Tree(p.)	3	4	6	5	1	2	5	1	6	2	4	3	6	1	2	4	5	3	5	4	6	3	2	1	5	2	6	4	3	1
Tree(c.)	4	6	3	5	1	2	5	3	6	1	4	2	6	1	3	5	4	2	2	5	6	3	1	4	5	4	6	3	1	1
Tree(g.)							3	4	6	2	5	1													3	4	6	2	5	1

Figure 3. Benchmarking CONCORD and other data integration methods across diverse structures.

(**A**) Two-batch, five-cluster simulation. For ground truth, we show kNN graphs (k=15 by default, with edges omitted) of both the noise-free and noise-added data (no batch effect). Latent spaces from each integration method are

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

visualized by kNN graphs, colored by batch (top) and cluster (bottom). (**B**) Trajectory simulation with varying batch overlap. The ground truth is shown with PCA and a kNN graph. For each method, the resulting latent space is depicted with a kNN graph (k = 15) to assess how well cells are integrated across batches along the trajectory. In the gap simulation, an additional kNN graph (k = 30), colored by simulated time, demonstrates that CONCORD accurately captures the correct orientation of the trajectories along time despite the gap. (**C**) Loop simulation with varying batch overlap. Shown here are kNN graphs of the ground truth (with edges omitted) and the CONCORD latent space. Full results for other methods are provided in Supplementary Figure 3D. (**D**) Tree simulation with varying batch overlap. kNN graphs for the ground truth and the CONCORD latent space are shown. Full results for other methods are provided in Supplementary Figure 3E. (**E**) scIB benchmarking on the two-batch, five-cluster simulation. Integration performance was evaluated using metrics from the scIB-metrics package[35]. (**F**) Ranking of integration methods. Each method's performance is scored across topological, geometric, and scIB metrics. The overall rank is based on the average ranking across all metrics.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

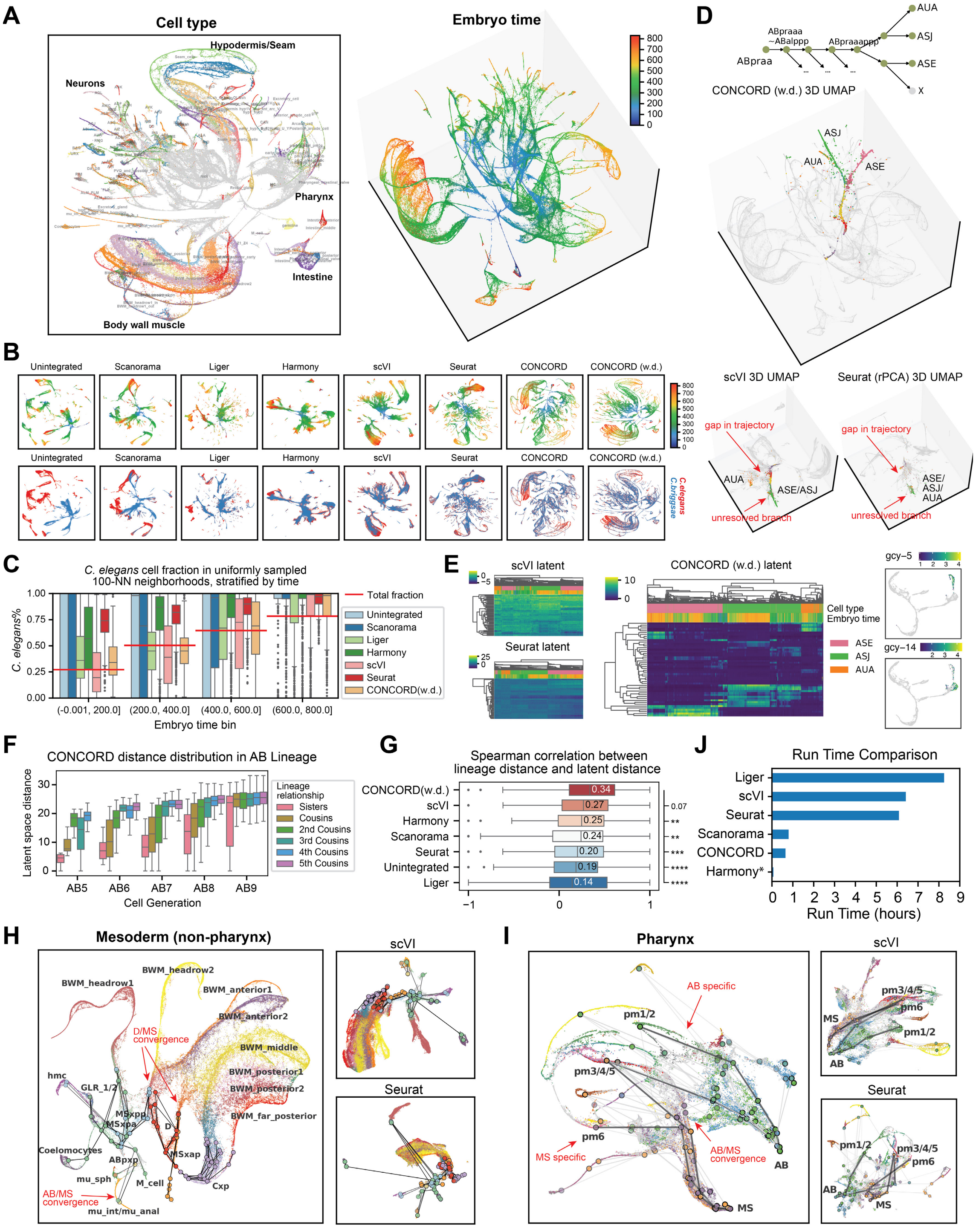


Figure 4. Benchmarking CONCORD on *C. elegans/C.briggsae* embryogenesis atlas.

(A) Global 2D and 3D UMAPs of CONCORD (with decoder) colored by cell type and estimated embryo time. (**B**) UMAP of CONCORD and other integration methods colored by estimated embryo time and species. (**C**) Boxplots show the fraction of *C. elegans* cells within randomly sampled 100-

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

nearest-neighbor (100-NN) neighborhoods, stratified by embryo time bins. The red horizontal line represents the expected species fraction based on the global composition of each time bin. Well-integrated datasets should show species fractions closely matching this expected value with minimal variation. (**D**) Global 3D UMAPs of CONCORD (with decoder), Seurat and scVI, highlighting cells mapped to the lineage sub-tree that give rise to ASE, ASJ and AUA neurons. (**E**) Heatmap showing the top 50 most variable latent dimensions in the ASE, ASJ, and AUA neuron subset for scVI, Seurat, and CONCORD (with decoder). Expression of gcy-5 and gcy-14 were plotted on the CONCORD (with decoder) UMAP. (**F**) Latent space distance between medoids of ectodermal cells (AB lineage), stratified by cell generation and lineage relationship. AB5 refers to cells derived after five successive divisions of the AB founder cell, with AB6 to AB9 representing progressively later generations. (**G**) Spearman correlation between lineage distance and latent space distance across integration methods for AB lineages from generations 5 to 9. Statistical significance of differences in correlations was assessed using a two-sided Mann-Whitney U test, with asterisks indicating significance levels (**p < 0.01, ***p < 0.001, ****p < 0.0001). (**H**) Zoom-in UMAPs for mesoderm cells excluding pharynx. Major input lineages and cell types were highlighted. Each lineage was represented by its cluster medoid on the UMAP, and lines connect each parental lineage to its daughter lineages following the lineage tree. (**I**) Zoom-in UMAPs for pharynx, annotated with cell types and broad input lineages. Selected lineage paths that give rise to pm1/2, pm3-5, and pm6 are highlighted. (**J**) Run time comparison of different integration methods. *Harmony was run using a 300-dimensional PCA input, whereas all other methods were applied to the gene expression matrix containing 10,000 variably expressed genes.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

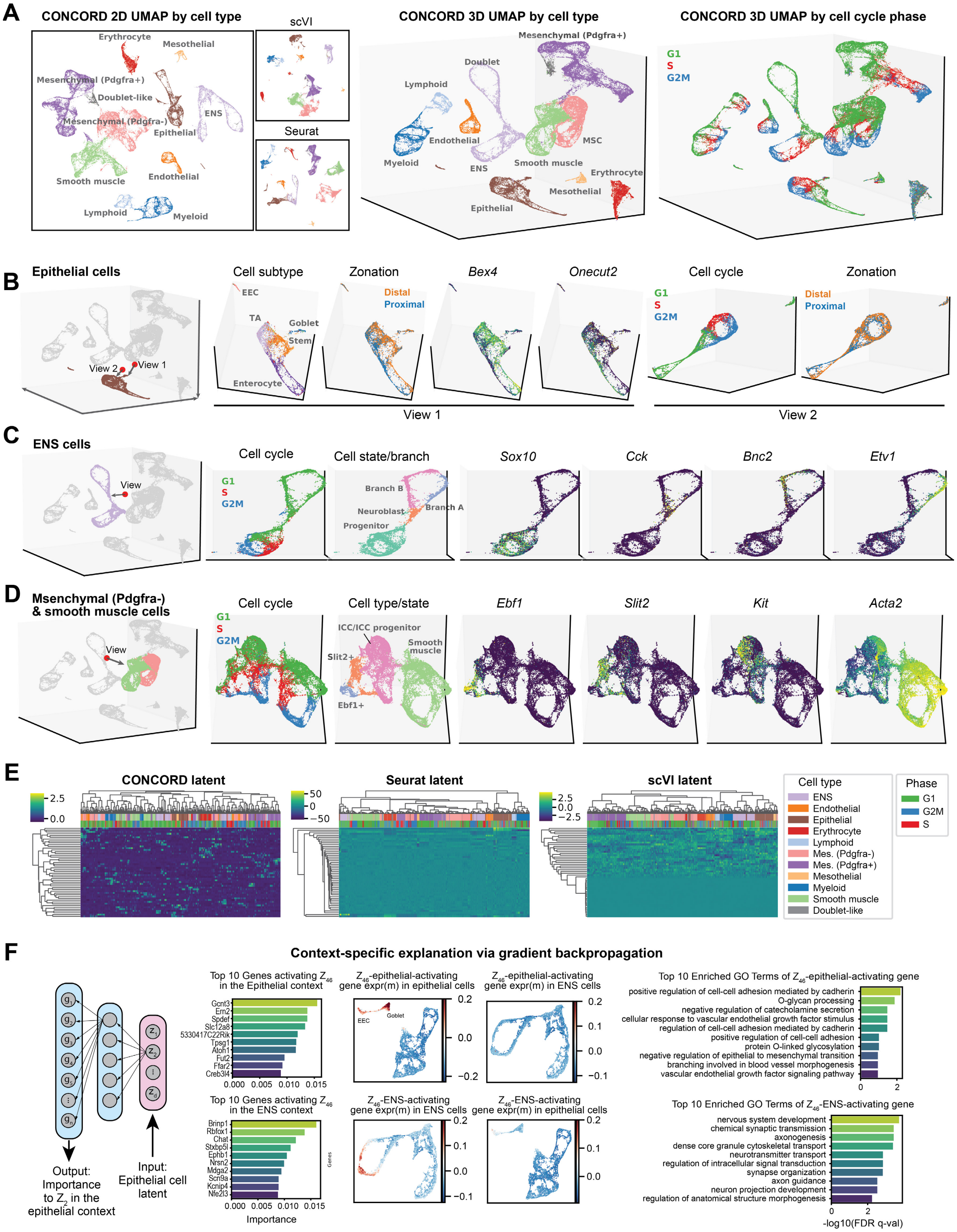


Figure 5. Benchmarking CONCORD on mammalian intestine development.

(**A**) 2D and 3D UMAP visualizations of CONCORD latent space, colored by cell type and cell cycle phase, with cell-type-colored UMAPs from scVI and Seurat shown for comparison. (**B**) Zoom-in views of epithelial cells in the 3D global

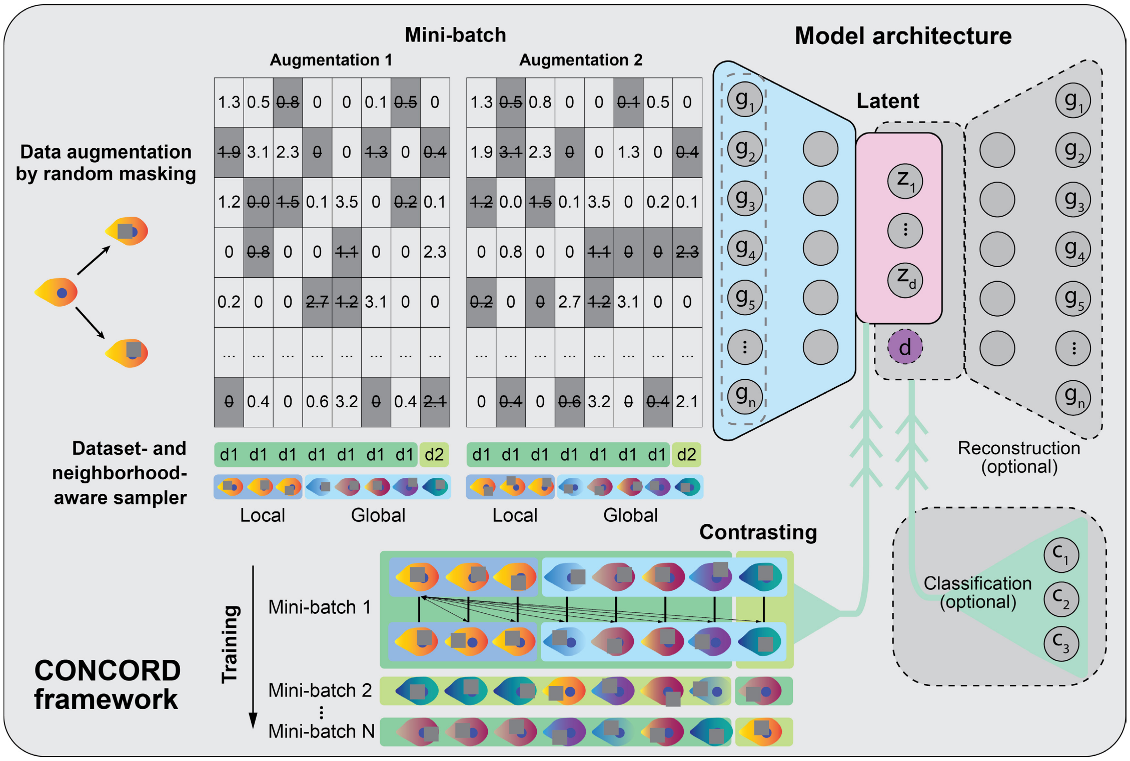
bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

UMAP, colored by cell subtype, zonation, and expression of zonation-specific markers (Bex1, Onecut2). A red marker and arrow indicate the viewing angle within the 3D global UMAP. (**C**) Zoom-in view of enteric nervous system (ENS) cells, colored by cell cycle phase and cell state/branch annotations, based on Morarach et al[50], along with state-specific gene expression. A red marker and arrow indicate the viewing angle. (**D**) Zoom-in view of Pdgfra- mesenchymal cells and smooth muscle cells, colored by cell cycle phase, subtype annotation, and selected subtype-specific markers. A red marker and arrow indicate the viewing angle. (**E**) Heatmap of all latent encodings generated by CONCORD, Seurat, and scVI. (**F**) Interpretation of CONCORD latent space using gradient-based attribution techniques. Activation of Neuron 46 (Z46) in epithelial and ENS cells is attributed to the co-expression of epithelial- and neuron-specific gene sets in their respective contexts. GO enrichment analysis of these gene sets is shown.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Supplementary Information

Supplementary Figures



Augmentation 1

1.3	0.5	~~0.8~~	0	0	0.1	~~0.5~~	0
~~1.9~~	3.1	2.3	~~0~~	0	~~1.3~~	0	~~0.4~~
1.2	~~0.0~~	~~1.5~~	0.1	3.5	0	~~0.2~~	0.1
0	~~0.8~~	0	0	~~1.1~~	0	0	2.3
0.2	0	0	~~2.7~~	~~1.2~~	3.1	0	0
...	...	...	...	...	...	...	...
~~0~~	0.4	0	0.6	3.2	~~0~~	0.4	~~2.1~~

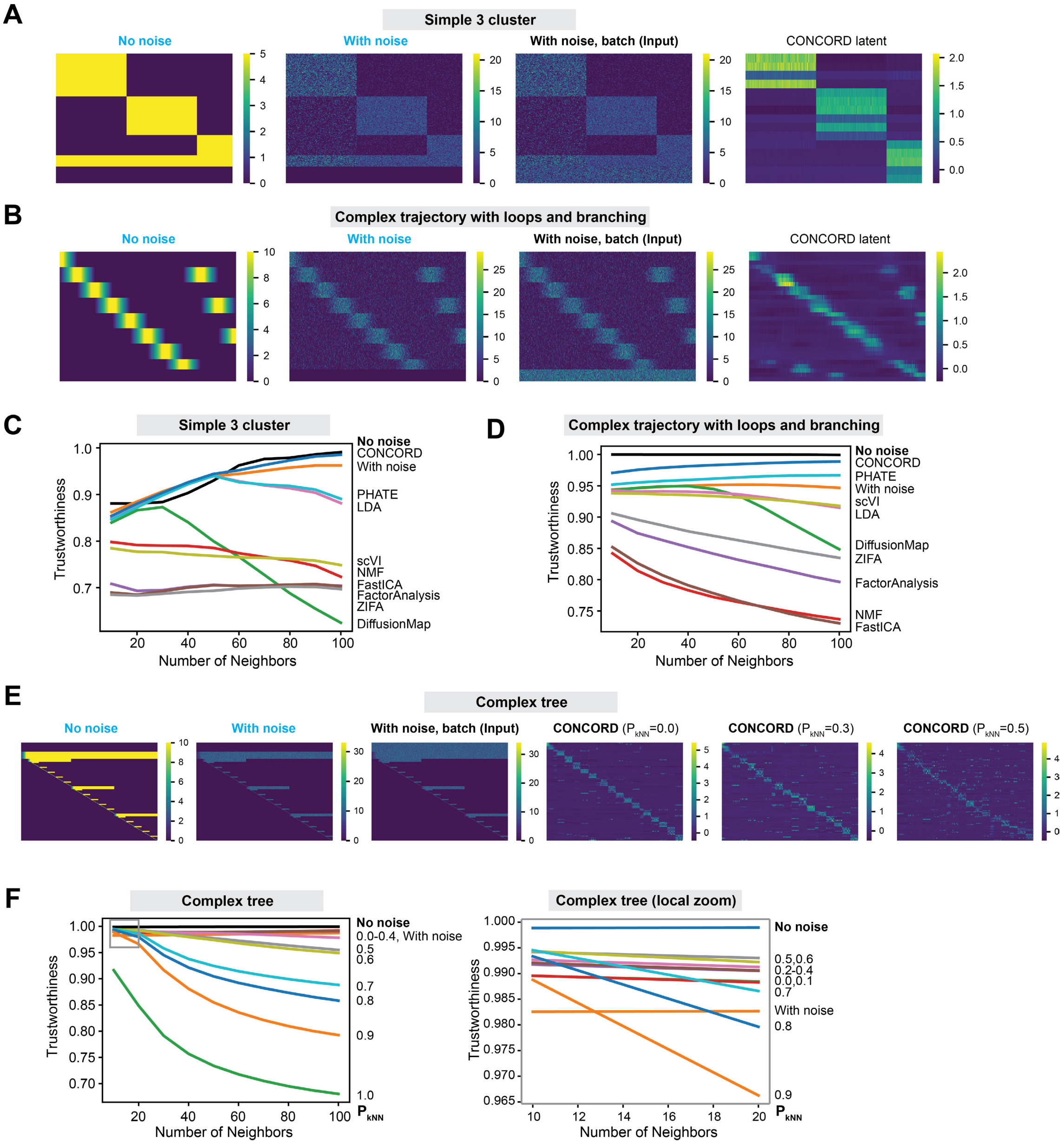
Augmentation 2

1.3	~~0.5~~	0.8	0	0	~~0.1~~	0.5	0
1.9	~~3.1~~	2.3	~~0~~	0	1.3	0	~~0.4~~
~~1.2~~	0.0	~~1.5~~	0.1	3.5	0	0.2	0.1
0	0.8	0	0	~~1.1~~	~~0~~	~~0~~	~~2.3~~
~~0.2~~	0	~~0~~	2.7	~~1.2~~	3.1	0	0
...	...	...	...	...	...	...	...
0	~~0.4~~	0	~~0.6~~	3.2	~~0~~	~~0.4~~	2.1

Supplemental Figure 1. The CONCORD framework.

Each cell undergoes random masking twice, and contrastive loss is computed on the latent encodings of mini-batches sampled using the dataset- and neighborhood-aware sampler. An optional decoder, incorporating a learnable dataset covariate, enables batch-free gene expression inference, while an optional classifier can be included for cell type classification or annotation-guided learning.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



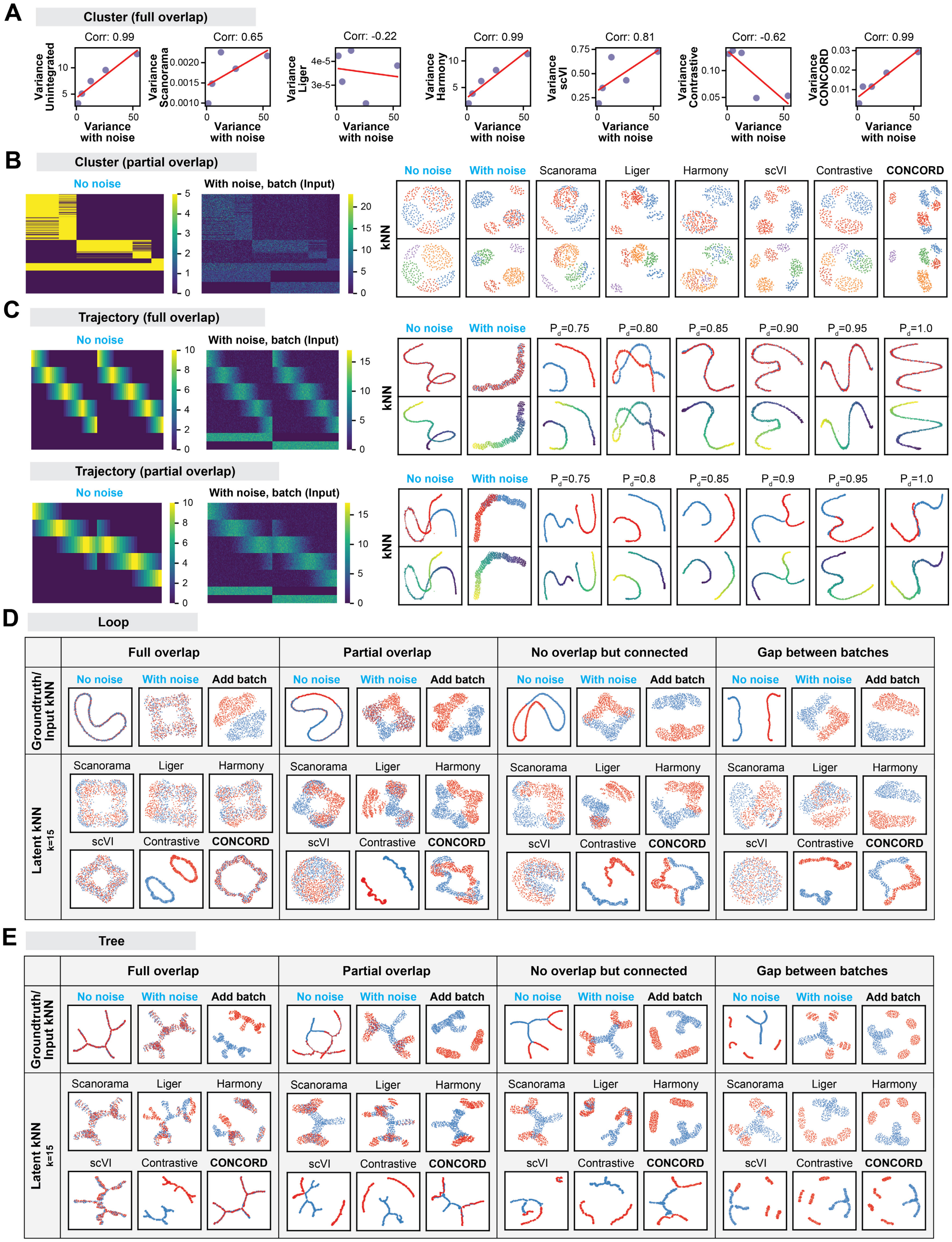
Supplemental Figure 2. Benchmarking CONCORD and other dimensionality reduction methods across diverse structures.

(A) Heatmaps of the simulated expression for the 3-cluster structure and the corresponding CONCORD latent encoding. (**B**) Heatmaps of the simulated expression for the trajectory-loop structure and the corresponding CONCORD latent encoding. (**C**) Trustworthiness measured across varying neighborhood sizes in the 3-cluster simulation. Note that in the noise-free setting, within-cluster neighbors are assigned randomly, so trustworthiness does not equal 1 for the no-noise data. (**D**) Trustworthiness measured across varying neighborhood sizes in the complex trajectory-loop simulation. (**E**) Heatmaps of the simulated expression for the complex tree structure shown in Figure 2G and the CONCORD latent encoding with different intra-kNN enrichments. (**F**)

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Trustworthiness measured across varying neighborhood sizes in the complex tree simulation, evaluated under different intra-kNN enrichment conditions. A zoomed-in view of the region where k < 20 and trustworthiness > 0.96 highlights improvements in capturing local neighborhood relationships with moderate intra-kNN enrichments.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

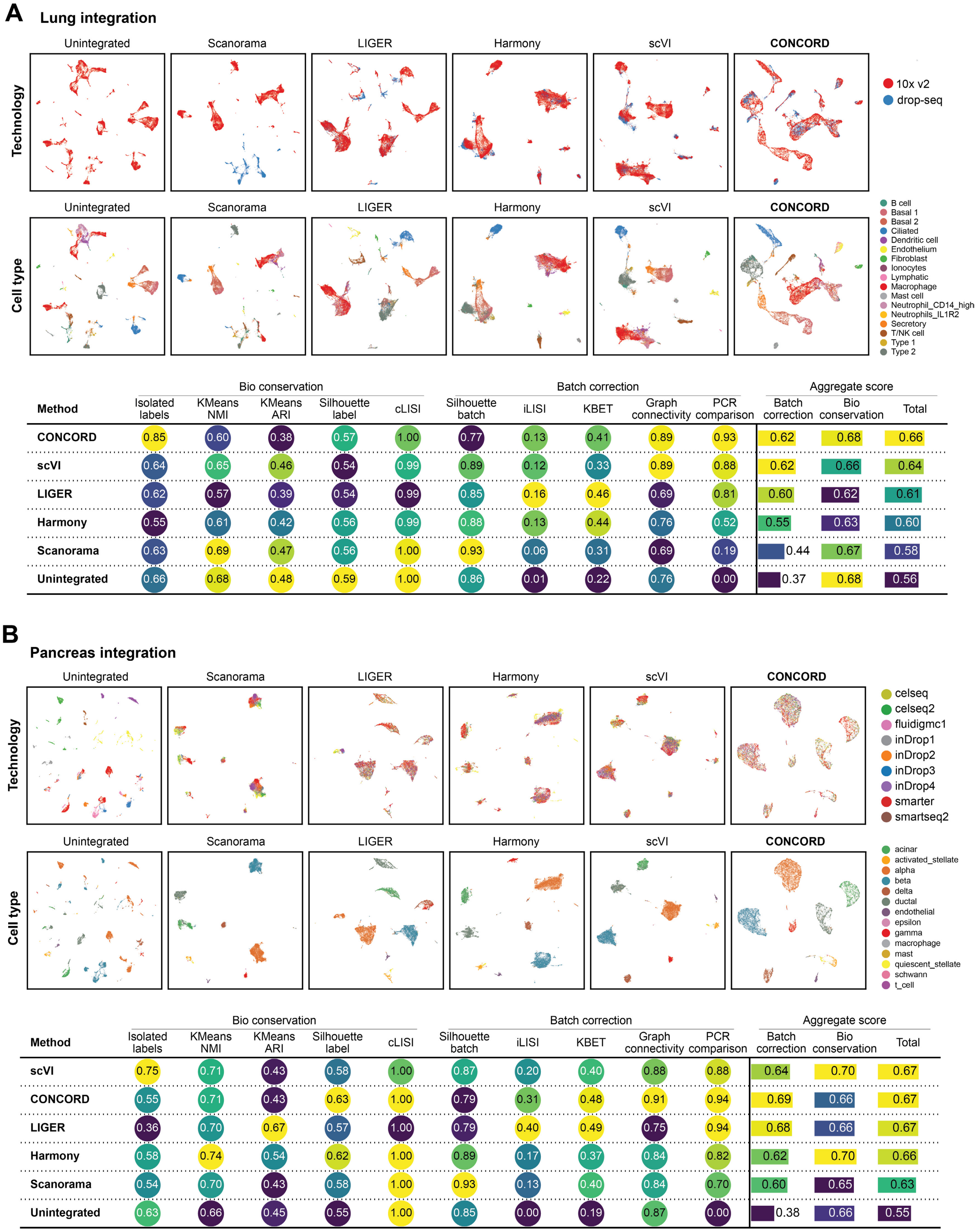


Supplemental Figure 3. Benchmarking CONCORD and other data integration methods across diverse structures. (**A**) Correlation of cluster-specific variances between noise-added ground truth and latent from each integration methods. (**B**) Two-batch, five-cluster simulation with minimal overlap. The two batches share only one cluster. kNN graphs (edges omitted) from each integration method are shown. (**C**) Trajectory simulation with full and

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

partial overlap. kNN graphs (edges omitted) of the ground truth and CONCORD latent with varying intra-dataset enrichment (P_d) are shown, colored by batch (top) and time (bottom). (**D**) Loop simulation with varying batch overlap. kNN graphs (edges omitted) of the ground truth and each integration method are shown. (**E**) Tree simulation with varying batch overlap. kNN graphs (edges omitted) of the ground truth and each integration method are shown.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



	Bio conservation					Batch correction					Aggregate score		
Method	Isolated labels	KMeans NMI	KMeans ARI	Silhouette label	cLISI	Silhouette batch	iLISI	KBET	Graph connectivity	PCR comparison	Batch correction	Bio conservation	Total
CONCORD	0.85	0.60	0.38	0.57	1.00	0.77	0.13	0.41	0.89	0.93	0.62	0.68	0.66
scVI	0.64	0.65	0.46	0.54	0.99	0.89	0.12	0.33	0.89	0.88	0.62	0.66	0.64
LIGER	0.62	0.57	0.39	0.54	0.99	0.85	0.16	0.46	0.69	0.81	0.60	0.62	0.61
Harmony	0.55	0.61	0.42	0.56	0.99	0.88	0.13	0.44	0.76	0.52	0.55	0.63	0.60
Scanorama	0.63	0.69	0.47	0.56	1.00	0.93	0.06	0.31	0.69	0.19	0.44	0.67	0.58
Unintegrated	0.66	0.68	0.48	0.59	1.00	0.86	0.01	0.22	0.76	0.00	0.37	0.68	0.56

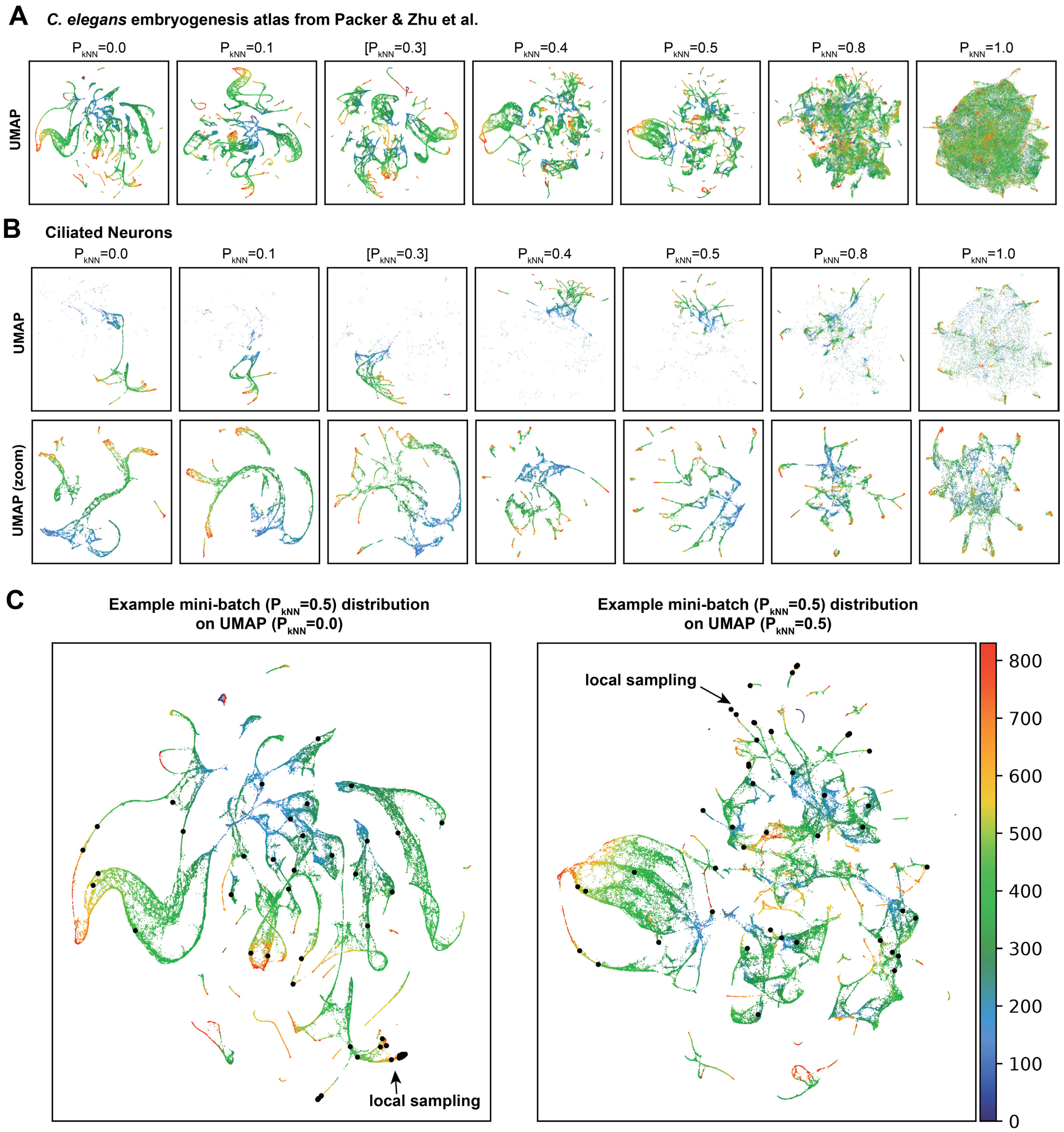
	Bio conservation					Batch correction					Aggregate score		
Method	Isolated labels	KMeans NMI	KMeans ARI	Silhouette label	cLISI	Silhouette batch	iLISI	KBET	Graph connectivity	PCR comparison	Batch correction	Bio conservation	Total
scVI	0.75	0.71	0.43	0.58	1.00	0.87	0.20	0.40	0.88	0.88	0.64	0.70	0.67
CONCORD	0.55	0.71	0.43	0.63	1.00	0.79	0.31	0.48	0.91	0.94	0.69	0.66	0.67
LIGER	0.36	0.70	0.67	0.57	1.00	0.79	0.40	0.49	0.75	0.94	0.68	0.66	0.67
Harmony	0.58	0.74	0.54	0.62	1.00	0.89	0.17	0.37	0.84	0.82	0.62	0.70	0.66
Scanorama	0.54	0.70	0.43	0.58	1.00	0.93	0.13	0.40	0.84	0.70	0.60	0.65	0.63
Unintegrated	0.63	0.66	0.45	0.55	1.00	0.85	0.00	0.19	0.87	0.00	0.38	0.66	0.55

Supplemental Figure 4. Performance of CONCORD on lung and pancreas datasets. (**A**) UMAPs were generated for both unintegrated lung data[35] and the latent space from each integration method, colored by technology and cell type. Benchmarking statistics from the scIB-metrics package[35] were shown. (**B**) Benchmarking of CONCORD on the integration of pancreas dataset[35]. UMAPs were generated for both unintegrated data and the latent space from each

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

integration method, colored by technology and cell type. Benchmarking statistics from the scIB-metrics package[35] were shown.

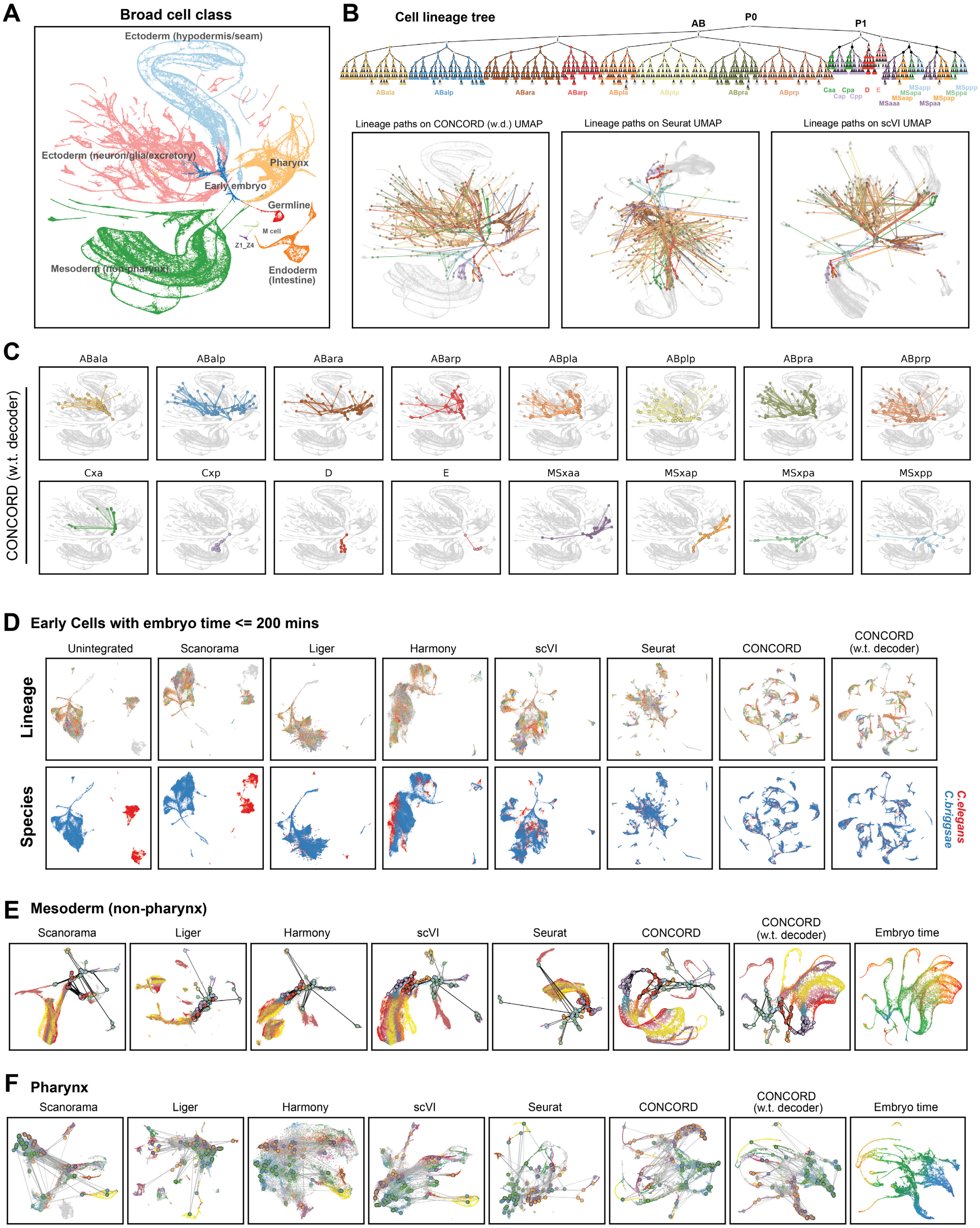
bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



Supplemental Figure 5. Effect of neighborhood-aware sampling on *C. elegans* embryogenesis atlas. (**A**) Global UMAP of CONCORD latent space with varying levels of intra-kNN enrichment, colored by embryonic time. (**B**) Global and zoomed-in UMAPs highlighting ciliated neurons, colored by embryonic time, generated from CONCORD latent representations across different levels of intra-kNN enrichment. (**C**) Visualization of cells sampled within a single mini-batch using an intra-kNN probability of 0.5. These cells are highlighted in the global UMAPs of CONCORD without neighborhood enrichment and with 0.5 neighborhood enrichment, demonstrating the effect of local sampling on structural resolution.

1230

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

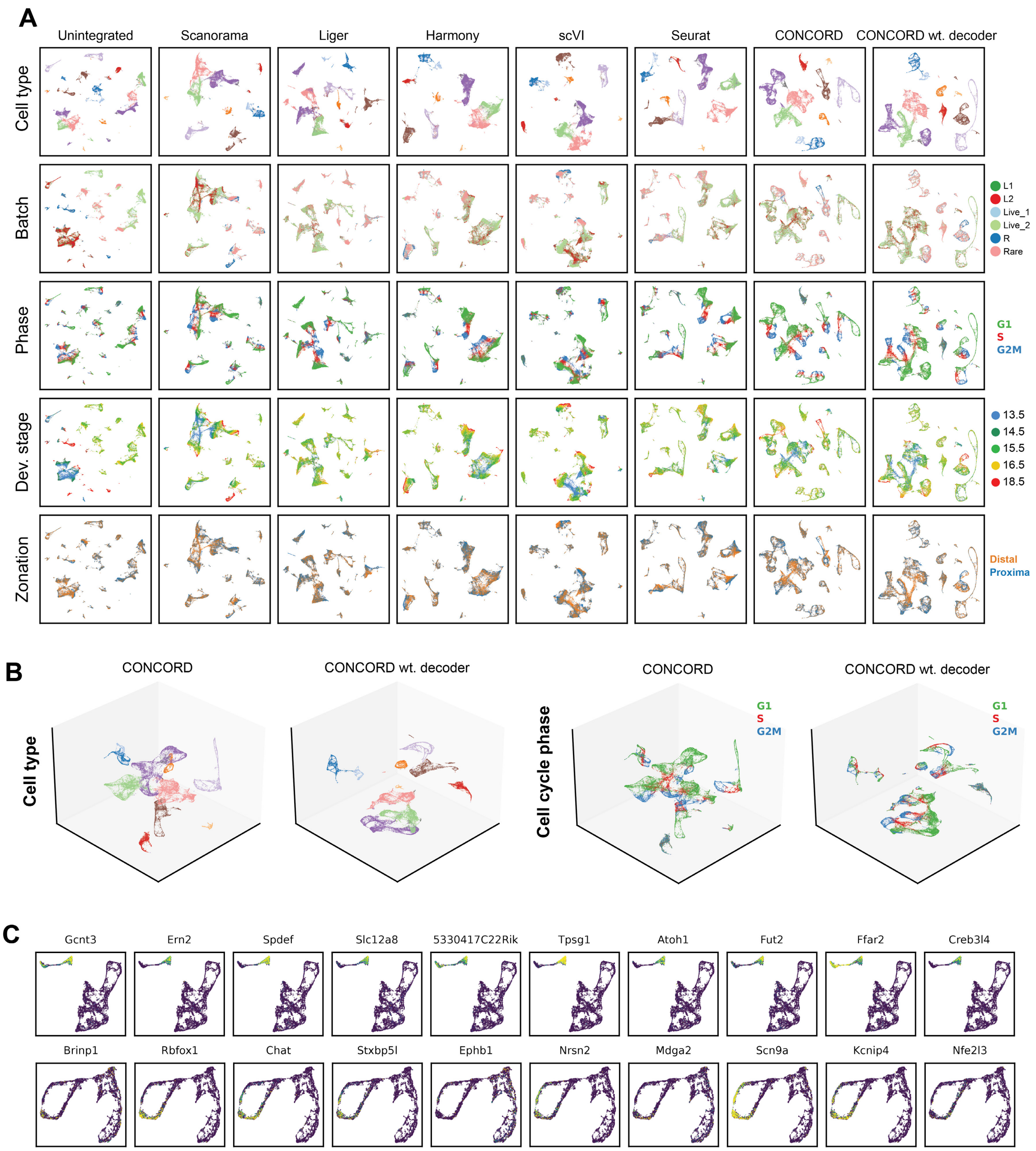


Supplemental Figure 6. Cell lineage tree and zoom-in analysis of C*. elegans/C.briggsae* embryogenesis atlas. (**A**) Global UMAP of CONCORD (with decoder) colored by broad cell class. (**B**) The lineage annotations by Large et al. were mapped to the *C. elegans* lineage tree, with some ambiguous mappings due to symmetry. Each lineage was represented by its cluster medoid on the UMAP, and lines connect each parental lineage to its daughter lineages following the lineage tree. (**C**) For CONCORD (with decoder), lineage

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

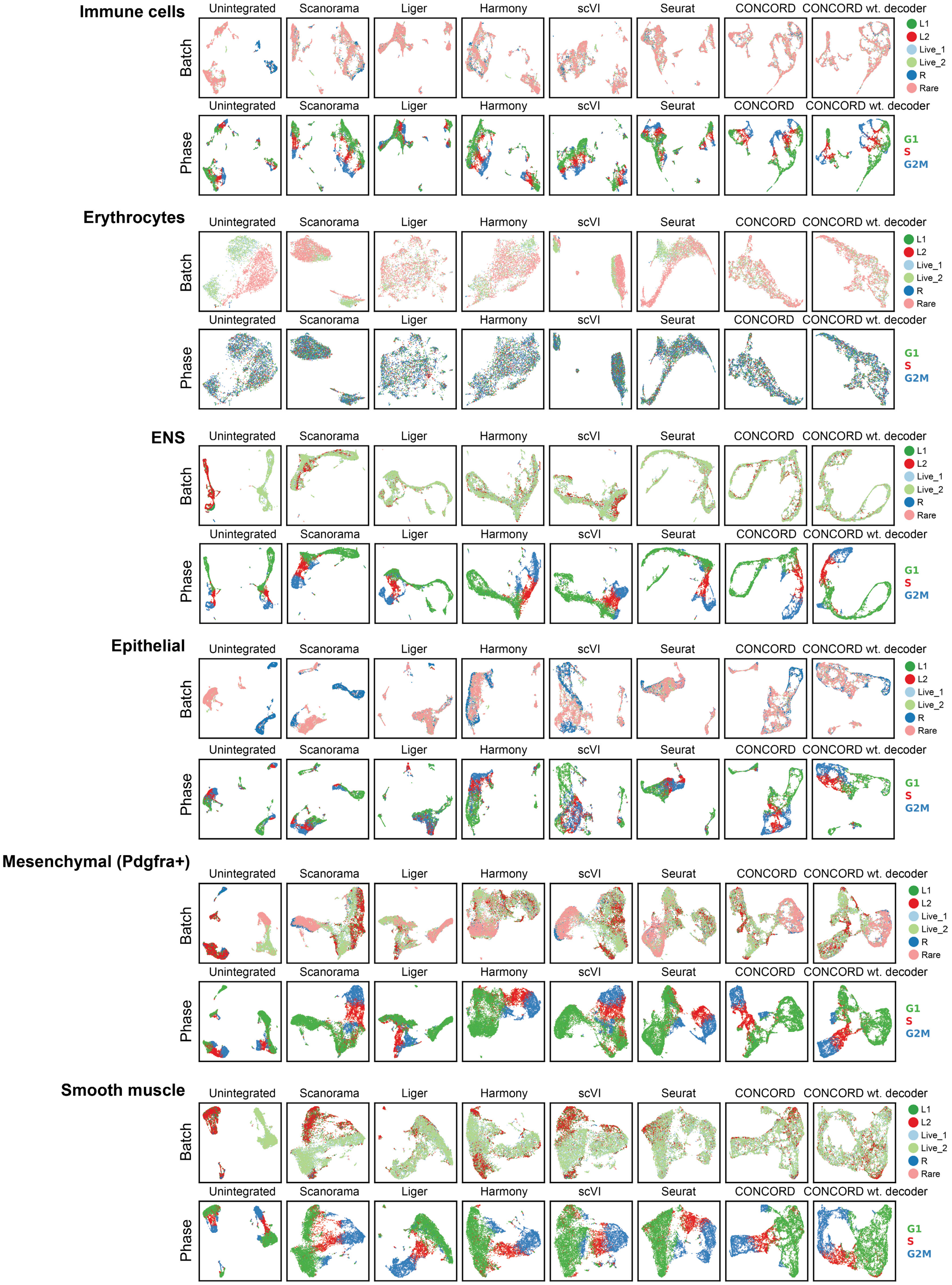
sub-trees of major lineage groups were separately highlighted. (**D**) UMAPs of early embryo cells with embryo time before or equal to 200 minutes, colored by lineage and species. (**E**) Zoom-in UMAPs for mesoderm cells excluding pharynx. Each input lineage was represented by its cluster medoid on the UMAP, and lines connect each parental lineage to its daughter lineages following the lineage tree. Estimated embryo time was also plotted on the UMAP generated with CONCORD (with decoder) latent. (**F**) Zoom-in UMAPs for pharynx, colored by cell types and input lineages. Lineage paths were plotted following the lineage tree. Estimated embryo time was also plotted on the UMAP generated with CONCORD (with decoder) latent.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



Supplemental Figure 7. Benchmarking CONCORD on mammalian intestine development. (**A**) UMAPs of the mouse intestinal developmental atlas[48] generated from CONCORD latent space, other integration methods, and unintegrated data, colored by broad cell type, batch, cell cycle phase, developmental stage, and zonation. (**B**) 3D UMAPs of CONCORD with and without the decoder, colored by cell type and cell cycle phase. (**C**) Expression patterns of the top genes contributing to Neuron 46 activation in the epithelial context (top) and ENS context (bottom).

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



Supplemental Figure 8. Zoom-in UMAPs of major cell types in the mammalian intestinal developmental atlas. Zoom-in UMAPs for each major cell type computed from CONCORD latent space, other integration methods, and unintegrated data, colored by batch (top) and cell cycle phase (bottom).

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.13.643146; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Supplementary Tables

Supplemental Table 1. Benchmarking data integration methods across diverse structures with varying degrees of overlap. For each structure type (cluster, trajectory, loop, and tree), we simulated varying degrees of overlap (full overlap, partial overlap, connected without overlap, and gap between batches). The table displays the score of each method using topological, geometric, and scIB metrics[35].