CIPHER: An end-to-end framework for designing optimized aggregated spatial transcriptomics experiments

Zachery Hemminger[1], Haley De Ocampo[1], Fangming Xie[2], Zhiqian Zhai[3], Jingyi Jessica Li[3,4], Roy Wollman[1,4,5*]

1 Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, California, United States of America
2 Department of Chemical Biology, David Geffen School of Medicine at UCLA, Los Angeles, California, United States of America
3 Department of Statistics and Data Science, University of California, Los Angeles, Los Angeles, California, United States of America
4 Institute of Quantitative Biosciences, University of California, Los Angeles, Los Angeles, California, United States of America
5 Department of Integrative Biology and Physiology, University of California, Los Angeles, Los Angeles, California, United States of America

* rwollman@ucla.edu

Abstract

Motivation Most spatial transcriptomics methods measure individual genes, which limits scalability and typically requires integration with scRNA-seq to recover full cellular states. Recent approaches such as CISI, FISHnCHIPs, and ATLAS address this limitation by measuring aggregate transcriptional signatures, where multiple genes are pooled into each channel to increase throughput. While aggregate measurements improve scalability, they shift the problem from gene selection to feature design. For effective integration with scRNA-seq, these signatures must be not only discriminative in transcriptional space but also straightforward to measure, with balanced signal, sufficient dynamic range, and robustness to experimental noise. By optimizing decoding accuracy in isolation, existing methods leave substantial performance on the table.

Results We present CIPHER (Cell Identity Projection using Hybridization Encoding Rules), a neural-network framework that jointly optimizes the experimental encoding matrix and the downstream cell-type embedding. CIPHER incorporates physical limits of imaging assays directly into its loss function, shaping the latent space to maximize discriminability while preserving robustness to measurement noise and signal constraints. Using a large-scale mouse brain scRNA-seq reference, we show that CIPHER-designed encodings produce latent spaces with improved cell-type separability, more uniform signal utilization, and increased resilience to hybridization variability, yielding higher decoding accuracy in both simulated and experimental datasets.

Conclusion CIPHER formulates aggregate signature design as a joint optimization problem over decoding accuracy and experimental measurability. This enables systematic, scRNA-seq-aligned feature design for scalable spatial transcriptomics based on aggregate measurements.

Availability Code and documentation are available at `https://github.com/wollmanlab/Design/`.

Author summary

Spatial transcriptomics reveals how cells are organized within tissues by mapping where genes are expressed. To achieve both scale and resolution, many approaches now combine spatial imaging with single-cell RNA-seq references

to reconstruct complete transcriptomes *in situ*. New methods such as CISI, FISHnCHIPs, and ATLAS accelerate this process by measuring combinations of co-expressed genes rather than each gene individually. These aggregate measurements simplify experiments but introduce a new challenge: deciding which genes to combine so that the resulting features are both experimentally reliable and computationally informative for integration with scRNA-seq data. We developed CIPHER, a computational framework that learns how to design and decode these aggregate measurements optimally. By integrating experimental constraints and decoding accuracy into a unified neural-network model, CIPHER provides a principled approach for designing signature-based spatial transcriptomics experiments that enable efficient and accurate reconstruction of cellular transcriptomes.

Introduction

Spatial transcriptomics aims to connect gene expression with cellular organization in intact tissues, but existing platforms face fundamental trade-offs between throughput, sensitivity, and spatial resolution [1, 2]. Imaging-based approaches that directly detect individual RNA molecules can achieve high spatial precision but scale poorly. High-magnification, spot-based methods require long acquisition times and are limited to small fields of view [3–5], while lower-magnification imaging platforms trade molecular resolution for speed and coverage [6, 7]. Sequencing-based spatial assays avoid optical crowding altogether but capture only a small fraction of the transcriptome per spatial unit, with detection biased toward abundant transcripts [**?**, 2, 8]. As a result, no current technology can simultaneously achieve high sensitivity, large spatial coverage, and comprehensive transcriptome profiling. Most spatial transcriptomics studies therefore rely on integration with single-cell RNA sequencing data to recover cell identities and transcriptional states, placing increasing emphasis on how spatial measurements are designed to support downstream integration rather than direct gene-by-gene readout.

To overcome the limitations of gene-by-gene spatial measurements, a new class of spatial transcriptomics methods adopts a different objective: designing measurements that enable easy integration with single-cell RNA sequencing references rather than attempting to directly recover partial transcriptomes *in situ*. Methods such as CISI, FISHnCHIPs, and ATLAS replace individual gene measurements with aggregate transcriptional signatures, in which multiple genes are combined into a smaller number of composite features [9–11]. By leveraging prior knowledge of transcriptional structure learned from single-cell data, these approaches shift the burden of resolution from the imaging assay to the computational decoding step. This reframing enables higher throughput, reduced imaging demands, and more efficient use of experimental signal while preserving the ability to assign cells to reference-defined identities and states.

Although CISI, FISHnCHIPs, and ATLAS share the goal of integration-first spatial measurement, they differ fundamentally in how aggregate encodings are constructed and interpreted. CISI introduced the use of composite measurements by applying compressed sensing to randomly generated gene mixtures, followed by sparse recovery to reconstruct individual gene expression profiles [9]. FISHnCHIPs instead defines aggregate features by grouping co-expressed genes into modules derived from single-cell data, prioritizing signal amplification and sensitivity through coordinated hybridization [10]. ATLAS generalizes aggregate measurement design by representing each feature as a weighted combination of genes and explicitly learning a linear mapping from gene expression space into a low-dimensional latent space aligned with known cell types [11]. This formulation enables supervised optimization of encodings for cell-type separability and scalability, allowing ATLAS to profile tens of millions of cells across hundreds of tissue sections while maintaining alignment with single-cell references [11]. Across all three approaches, aggregate measurement design follows a common workflow of computational encoding, *in situ* measurement, and computational decoding, but differs in how the latent space is defined and optimized.

Despite their success, existing aggregate spatial transcriptomics methods do not explicitly model the experimental factors that determine whether an encoding can be measured reliably. Encoding design is typically performed entirely in silico, without accounting for finite signal dynamic range, variability in hybridization efficiency across probes, background structure, or uneven signal allocation across channels. These effects directly shape signal separability and decoding performance in real experiments, yet they are absent from the encoding optimization itself [3, 4, 10]. In practice, this leads to encodings that may be highly discriminative in theory but are difficult to read out robustly, or that waste measurement capacity on features that are poorly constrained by the assay. Addressing this gap requires

design frameworks that treat physical and biochemical measurement limits as first-class components of the encoding problem.

Here we present CIPHER (Cell Identity Projection using Hybridization Encoding Rules), a computational framework for designing aggregate spatial transcriptomics encodings that explicitly account for experimental measurability. CIPHER jointly optimizes encoding and decoding accuracy under physical constraints on signal dynamic range, probe usage, and measurement robustness. We first introduce the conceptual framework underlying measurability-aware encoding design (Figure **??**), then quantify the trade-offs imposed by experimental constraints and their impact on encoding performance (Figures **??**–**??**). We compare CIPHER-designed encodings to existing strategies, including DPNMF, the method used in the ATLAS paper [**?**](Figure **??**), and evaluate performance using large-scale simulations (Figure **??**) and experimental validation in mouse brain tissue (Figure **??**). Finally, we demonstrate how CIPHER can be applied to additional organs beyond the mouse brain (Figure **??**). Together, these results establish CIPHER as a principled framework for information-optimal and experimentally feasible design in signature-based spatial transcriptomics.

Methods

Terminology

We use the term *bit* to denote a single dimension of the encoded measurement space. Unlike classical MERFISH-style bits, which represent binary presence or absence of signal, CIPHER bits correspond to continuous-valued aggregate transcriptional features formed by pooling multiple genes into a single readout channel.

Datasets and Preprocessing

CIPHER was evaluated across three biological contexts: adult mouse brain, developing mouse, and mouse lymph node. For all datasets, raw scRNA-seq count matrices were processed uniformly. Genes were retained if probe designs were available in the PaintSHOP probe database; no additional filtering based on expression level, transcript length, or sequence complexity was applied. Mitochondrial and ribosomal genes were retained if probe designs were available. Gene expression matrices were normalized to counts per 100,000 (CP100k) by dividing each cell by its total library size and multiplying by 10^5. No cells were excluded based on library size or sparsity. Each dataset was split into training (80%) and test (20%) sets using stratified sampling to preserve cell-type proportions.

Adult mouse brain. We used the Allen Brain Cell Atlas whole mouse brain 10X scRNA-seq dataset (WMB-10X) [12]. Cells with valid metadata and genes shared across 10X v2 and v3 chemistries were retained. To ensure balanced representation during training, the dataset was subsampled to 100,000 cells using stratified sampling, with oversampling of rare cell types. Cell-type labels correspond to the subclass annotation level ($K = 336$ categories).

Developing mouse. We used a large-scale single-cell atlas of mouse development spanning embryonic and postnatal stages [13]. Gene names were mapped to Ensembl identifiers for compatibility with probe design resources. Stratified sampling was used to construct a balanced dataset of 500,000 cells based on the author-provided cell-type annotations ($K = 190$ categories).

Mouse lymph node. We constructed a composite lymph node reference by merging multiple published mouse lymph node scRNA-seq datasets (listed at [**?**]). Cell-type labels were harmonized across datasets, and doublets were removed using the original study annotations. The final dataset was stratified to 100,000 cells across 43 harmonized cell types.

CIPHER Architecture

CIPHER is an auto-encoding neural network consisting of a biologically constrained linear encoder and a nonlinear decoder.

Encoder (experimental design layer). The encoder represents the physical hybridization of oligonucleotide probes to mRNA molecules *in situ*. It learns a nonnegative encoding matrix $E \in \mathbb{R}^{G \times B}$, where G is the number of genes and B is the number of bits. Each entry E_{ij} specifies the number of probes targeting gene i that contribute to bit j.

To ensure physical realizability, encoder weights are constrained using a bounded activation:

$$E_{ij} = \tanh(W_{ij}) \cdot T_i$$

where W_{ij} are unconstrained trainable parameters, $\tanh(\cdot)$ is scaled to map to the interval $[0, 1]$, and T_i is the total number of probes available for gene i across all bits. T_i is fixed across datasets and depends only on gene properties such as transcript length and probe design availability, as determined by PaintSHOP [14]. Continuous values of E_{ij} are discretized by rounding to the nearest integer for probe synthesis.

The latent projection is defined as:

$$P = XE$$

where X is the normalized gene expression matrix and P is the continuous-valued latent signal matrix.

Decoder (cell-type prediction). The decoder maps latent projections P to cell-type probabilities. It consists of a multilayer perceptron with h hidden layers, each of width $3B$. Hidden layers use batch normalization followed by GELU activation. The output layer is linear and followed by a softmax transformation to produce class probabilities. Unless otherwise specified, we use a single hidden layer ($h = 1$). Dropout is not used during training or in any experiments reported in this paper.

Loss Functions and Optimization

CIPHER is trained end-to-end using a weighted sum of loss terms:

$$L_{\text{total}} = w_A L_A + w_H L_H + w_M L_M + w_{\text{reg}} L_{\text{reg}}$$

Loss terms are not normalized by batch size, gene count, or number of bits, as absolute loss magnitudes are not compared across designs.

Accuracy loss (L_A). Cell-type discrimination is enforced using categorical cross-entropy with label smoothing ($\epsilon = 0.1$).

Hybridization constraints (L_H). Two constraints enforce probe feasibility: a global probe budget limiting $\sum_{i,j} E_{ij}$ to a target value, and a per-gene constraint ensuring that $\sum_j E_{ij} \leq T_i$.

Measurability losses (L_M). Latent signals are shaped to ensure experimental robustness using three criteria: minimum brightness per bit, sufficient dynamic range (defined using the 5–15 and 85–95 percentiles), and separation between cell-type centroids in latent space. All denominators include $\epsilon = 10^{-8}$ to prevent numerical instability.

Regularization (L_{reg}). Additional penalties discourage reliance on single dominant genes per bit and penalize correlations between bits to promote independent information channels.

Training schedule and parameters. Unless otherwise stated, models were trained with $B = 18$ bits and a single decoder hidden layer. Training ran for 100,000 iterations with batch size 500 using Adam optimization (learning rate 0.01). Loss weights and target values are listed in the Default Parameters section. Key targets were linearly ramped over the first 75% of training iterations.

Noise Injection and Robustness

During training, only additive measurement noise was applied to stabilize optimization and account for nonzero background fluorescence. This noise was sampled uniformly between 0 and 10^4 and added to latent projections. During evaluation, learned encodings were subjected to additional transcriptional and hybridization noise as described in the Results.

Evaluation Metrics

Performance was evaluated on held-out test data using accuracy, separation, and Dynamic Fold. Cell-type centroids were computed as means over test cells. Separation and Dynamic Fold metrics report the 10th percentile across all cell-type pairs and bits, respectively, ensuring robustness to worst-case behavior.

Probe Sequence Design and Assembly

Encoding matrices were converted into oligonucleotide libraries using PaintSHOP-designed 30-mer gene-specific probes [14]. For each gene–bit pair, integer probe counts were allocated according to the discretized encoding matrix. Readout sequences were drawn from a prevalidated orthogonal set and assembled into final probe constructs using rolling circle amplification (RCA). Full experimental synthesis details will be provided elsewhere.

Experimental Imaging and Decoding

ATLAS experiments were performed on adult mouse brain sections using CIPHER-designed probe pools and automated cyclic imaging. Measured signals were decoded using the SCALE framework, which harmonizes experimental data with scRNA-seq references using spatial priors and hierarchical classification. Full transcriptomes were imputed using nearest-neighbor averaging in latent space.

Results

CIPHER (Cell Identity Projection using Hybridization Encoding Rules) is a neural-network framework for the joint design and decoding of transcriptional signature measurements (Figure **??**). The model links single-cell RNA sequencing (scRNA-seq) expression profiles (X) to cell-type identities (C) through an intermediate latent representation (P) that defines a low-dimensional projection of transcriptional space. Each dimension of P corresponds to an aggregate transcriptional signature, implemented as a weighted combination of genes that can, in principle, be realized *in situ* using pools of DNA oligonucleotide probes, hereafter referred to as encoding probes. Unlike prior aggregate encoding approaches that focus primarily on representing reference information accurately, CIPHER explicitly incorporates the experimental realizability of these signatures. Its objective is to design transcriptional signatures that remain maximally informative under realistic molecular and optical constraints imposed by *in situ* hybridization within intact tissue.

CIPHER adopts a neural-network architecture that accommodates multiple design objectives through distinct loss functions (L_H, L_M, and L_A; Figure **??**). The encoder constitutes the first linear layer of the network and represents the molecular design of the experiment, specifying how many probes hybridize to each gene and how their combined signals contribute to each readout channel. Because probe hybridization is additive and implemented through physical oligonucleotides, encoder weights are constrained to be nonnegative and correspond to the number of encoding probes assigned per gene–channel pair. This layer performs a linear projection from gene space to latent space, directly mirroring the physical encoding of gene combinations *in situ*.

The decoder introduces nonlinearity through hidden layers that map transcriptional signatures to predicted cell types. Unlike the encoder, the decoder has no molecular counterpart and exists purely *in silico*. Together, these components capture the linear physics of measurement and the nonlinear decision boundaries of cell-type classification within a unified framework. Training is performed end-to-end using scRNA-seq data, allowing the encoder and decoder to co-optimize such that the learned encoding matrix reflects both biological structure and experimental constraints.

CIPHER incorporates three complementary loss terms, each targeting a distinct design objective. The hybridization loss L_H penalizes excessive probe usage per gene to promote sparse, experimentally feasible encodings. A weight w_{ij} mapping gene i to readout channel j corresponds to the number of bivalent encoding probes hybridizing to gene i and contributing to channel j. Because the number of accessible binding sites on an mRNA is finite and depends on transcript length and sequence composition, the sum of w_{ij} across channels is physically constrained.

The accuracy loss L_A enforces biological fidelity by measuring categorical cross-entropy between predicted and known cell-type labels derived from the scRNA-seq reference. This term captures computational decoding performance independent of experimental considerations. The measurability loss L_M promotes robustness under realistic experimental conditions by encouraging broad separation of cell-type signatures in signal space and sufficient dynamic range of encoded features. This term avoids degenerate solutions in which cell types are separable in principle but differ only weakly in absolute signal, rendering them difficult to distinguish experimentally. Together, these losses balance biological informativeness, computational accuracy, and experimental feasibility within a unified optimization framework.

We next examined how encoding dimensionality and loss-function composition influence CIPHER performance and latent-space structure (Figure **??**). All analyses used the Allen Mouse Brain Cell Type Atlas [12], which was also employed for ATLAS signature design. We trained CIPHER across increasing numbers of latent dimensions while systematically varying which loss terms were included.

Decoding accuracy increased rapidly with dimensionality before plateauing, indicating that most biologically relevant variation can be captured within a compact encoding space (Figure **??**). Models trained using only L_A achieved high accuracy with as few as 10–15 latent features. Adding the sparsity constraint L_H caused a modest reduction in accuracy, while inclusion of L_M produced an additional but small decrease. Using either component of L_M alone—dynamic range ($L_M^{(D)}$) or separability ($L_M^{(S)}$)—had a similar impact on accuracy as using both, indicating largely overlapping effects. These results show that incorporating experimental constraints does not substantially compromise decoding accuracy and that accuracy alone saturates early.

We quantified how each optimization objective affects cell-type separation in latent space (Figure **??**). Separation reflects robustness to measurement perturbations rather than idealized decoding accuracy. Models trained using only L_A, with or without L_H, achieved high accuracy but low separation, revealing a mismatch between *in silico* optimization and *in situ* measurement constraints. Inclusion of L_M directly addressed this gap: $L_M^{(D)}$ expanded dynamic range, $L_M^{(S)}$ increased inter–cell-type distances, and combining both yielded the largest gains, particularly at higher dimensionality. Dynamic fold exhibited the same behavior, remaining narrow under L_A and L_H alone and expanding substantially only when L_M was included. These gains occurred with only minor reductions in accuracy, demonstrating that accuracy alone is insufficient as a design criterion.

To assess the nonlinear capacity required for decoding, we varied the depth of the decoder f_D while holding the encoder and loss configuration fixed (Figure **??**). A purely linear decoder ($h = 0$) underperformed at low dimensionality, particularly with 3–6 features, indicating that a strictly linear mapping from P to C is insufficient. Introducing a single hidden layer largely eliminated this gap, while deeper decoders provided no consistent benefit (Figure **??**). Accordingly, all subsequent analyses use a decoder with one hidden layer ($h = 1$).

We next examined how signal-brightness constraints influence encoding robustness (Figure **??**). Relaxing brightness constraints caused a sharp drop in separation, whereas accuracy and dynamic fold declined only modestly. This indicates that brighter encodings often collapse inter–cell-type distances by recruiting weakly discriminative genes, reducing robustness despite high decoding accuracy. In contrast, varying the total probe budget produced gradual and parallel effects across all metrics (Figure **??**), with performance saturating at approximately 50,000 probes, a regime compatible with current synthesis capabilities.

We compared CIPHER to the unsupervised aggregate-signature method DPNMF using matched dimensionality and the same scRNA-seq reference (Figure **??**). DPNMF concentrates separability into coarse divisions, primarily neurons versus non-neurons, whereas CIPHER distributes separability across subclasses (Figure **??**a). Correspondingly, DPNMF exhibited a long tail of poorly decoded subclasses (Figure **??**b) and highly uneven brightness distributions dominated by a few saturated features (Figure **??**c). CIPHER avoided these pathologies and maintained robustness

under simulated noise perturbations (Figure **??**d).

We validated CIPHER *in situ* in mouse cortex, decoding cell types with SCALE and observing expected laminar and hippocampal organization (Figure **??**). Greedy spatial matching to MERFISH data showed high concordance, providing an upper bound on achievable agreement given imperfect registration. Finally, we applied CIPHER to the developing mouse brain and lymph node (Figure **??**), observing tissue-specific saturation of dimensionality at $\sim$24 and $\sim$9 features, respectively. These results demonstrate that CIPHER identifies compact, context-dependent encodings and generalizes across diverse biological systems.

Discussion

Aggregate spatial transcriptomics methods redefine the goal of *in situ* measurement: rather than recovering complete transcriptomes gene by gene, they aim to acquire just enough structured information to reliably align spatial observations with rich single-cell RNA sequencing references. This reframing has enabled dramatic gains in scale and throughput, but it also introduces a new and largely unresolved design problem. Once individual genes are no longer the fundamental measurement unit, the central question becomes how to construct aggregate features that are not only informative in transcriptional space but also robust to the physical constraints and noise inherent to *in situ* hybridization and imaging. Our results show that existing design strategies, which optimize decoding accuracy or reconstruction fidelity in isolation, leave substantial performance on the table. CIPHER addresses this gap by treating aggregate encoding as a joint optimization problem over biological informativeness, signal robustness, and experimental feasibility.

A central insight emerging from our analyses is that classification accuracy alone is an inadequate objective for aggregate encoding design. Across tissues and model configurations, decoding accuracy saturates rapidly with relatively few encoded features, often well before the encoding becomes robust to measurement variability. This behavior reflects a fundamental mismatch between in silico optimization and *in situ* data collection real effects. When encodings are evaluated against the same scRNA-seq reference used for training, small margins between cell types are sufficient for high apparent accuracy. However, in physical experiments, hybridization efficiency varies across probes and genes, background fluorescence is spatially structured, and signal dynamic range is finite [15–17]. Under these conditions, narrow margins collapse quickly. By explicitly optimizing for latent-space separation and usable dynamic range, CIPHER exposes trade-offs that are invisible to accuracy-based optimization and demonstrates that modest sacrifices in nominal accuracy yield disproportionate gains in robustness.

These findings place CIPHER in a distinct position relative to existing aggregate spatial transcriptomics frameworks. CISI introduced the idea of composite measurements using compressed sensing, focusing on efficient reconstruction of individual gene expression profiles from random mixtures [9]. While powerful, this approach is agnostic to how discriminative power and signal strength are distributed across features, and does not explicitly constrain dynamic range or hybridization feasibility. FISHnCHIPs takes a different approach by grouping co-expressed genes into modules to amplify signal and improve sensitivity [10]. However, module selection remains largely heuristic and gene-centric, and the optimization is not performed end to end with the decoding task. ATLAS and related supervised factorization approaches explicitly align aggregate features with reference-defined cell types, enabling scalable atlas construction [11]. Yet, as our comparison to DPNMF, the encoding schemed used in ATLAS, illustrates, supervised alignment alone does not guarantee robust encodings: without constraints on brightness balance, separability, and dynamic range, discriminative power can concentrate into a small number of saturated features, leaving fine-grained distinctions fragile.

By contrast, CIPHER integrates these considerations into a single framework. The encoder represents the physical design of the experiment, constrained by probe budgets and brightness limits, while the decoder captures the nonlinear decision boundaries required for cell-type assignment. Importantly, we find that this decoder need not be deep or complex. A single nonlinear hidden layer is sufficient to exploit the structure organized by the encoder, and deeper networks provide no consistent benefit. This observation underscores that the primary challenge in aggregate spatial transcriptomics lies not in classifier expressivity but in how information is projected into a measurable space. Once that projection is well designed, decoding becomes straightforward.

Our analysis of brightness and probe-budget constraints clarifies a critical asymmetry in aggregate encoding design. Aggressively increasing signal brightness produces encodings that remain accurate under idealized conditions but exhibit a sharp collapse in latent-space separability, rendering them highly sensitive to experimental noise. In contrast, constraining the total number of probes imposes a more uniform and predictable cost across accuracy, separation, and dynamic range, with performance plateauing at probe counts that are already experimentally accessible. These results identify brightness, rather than probe availability, as the dominant bottleneck in robust aggregate measurements. Importantly, they also suggest a practical path forward: combining CIPHER-designed encodings with post-hybridization signal amplification strategies, such as hybridization chain reaction (HCR), branched DNA amplification, or rolling-circle–based methods, should substantially improve separability and robustness without sacrificing encoding structure [**?**, 18–20]. Because these amplification approaches increase signal uniformly across encoded features rather than redistributing gene weights, they are well matched to CIPHER's emphasis on balanced signal allocation and margin preservation. This combination is expected to outperform naive brightness-driven panel design strategies by removing brightness as a design constraint altogether. With amplified readout, signal detectability no longer competes with separability, allowing aggregate encodings to achieve markedly higher and more uniformly distributed cell-type separation that remains robust across tissues, scales, and experimental conditions.

The comparison to DPNMF, the encoding used in ATLAS [11], illustrates the consequences of ignoring these constraints. Although DPNMF captures broad transcriptional structure, it allocates most separability to coarse divisions, such as neurons versus non-neurons, while leaving closely related subclasses weakly separated. This behavior is not incidental but follows directly from optimizing global reconstruction without regard to how separability and signal are distributed. As a result, DPNMF encodings exhibit highly uneven brightness profiles and degrade rapidly under simulated measurement noise. CIPHER, by explicitly enforcing balanced signal utilization and separation margins, produces encodings that maintain subclass-level resolution and degrade gracefully as noise increases. Notably, in relatively forgiving experimental contexts such as adult mouse cortex, both designs achieve acceptable decoding accuracy *in situ*. This observation is important: CIPHER does not claim universal superiority in all regimes. Rather, its advantages become most pronounced as measurement conditions become more challenging or as finer-grained distinctions are required.

Our *in situ* validation confirms that CIPHER-designed encodings can be implemented at scale and decoded reliably using existing pipelines [11]. The agreement with MERFISH-derived references, assessed through spatially local matching, demonstrates that CIPHER measurements capture local transcriptional structure with high fidelity. At the same time, this comparison highlights a broader issue shared by all spatial transcriptomics technologies: precise one-to-one cell correspondence across modalities is rarely achievable, and validation necessarily relies on neighborhood-level or distributional agreement [21, 22]. CIPHER does not resolve this limitation, but it ensures that the information captured by aggregate measurements is structured in a way that maximizes what can be reliably recovered given these constraints. Extending CIPHER to mouse development dataset and the lymph node reveals an additional and conceptually important result: the dimensionality required for robust encoding is strongly tissue dependent. Developmental systems with continuous trajectories and overlapping transcriptional programs require substantially more encoded features to stabilize separation and dynamic range than lineage-structured immune tissues. This observation suggests that the number of required aggregate features reflects an intrinsic property of the biological system and the question being asked, rather than a fixed technological parameter. In this sense, CIPHER provides a quantitative way to estimate the effective dimensionality of spatial variation that must be measured to achieve robust integration with single-cell references.

Several limitations of the current framework warrant discussion. First, CIPHER depends on the availability and quality of scRNA-seq references and their associated taxonomies. If rare or novel cell states are absent from the reference, or if the taxonomy itself is incorrect, the encoding will faithfully optimize for the wrong distinctions. Second, the measurement model incorporated through the LM loss captures coarse constraints on dynamic range and separability but does not account for sequence-specific probe effects, channel cross-talk, or spatially structured background. Incorporating richer, modality-specific noise models is an important direction for future work. Third, the framework is optimized for discrete cell-type classification and does not explicitly preserve continuous gradients or trajectories, which may be central in developmental or disease contexts [13]. Extending the loss functions to protect manifold structure rather than class margins alone would broaden the applicability of the approach. Finally, although we demonstrate compatibility with SCALE decoding, the encoder is trained with a specific decoder architecture, and

formal guarantees of transfer across decoding algorithms remain an open question.

Despite these limitations, CIPHER establishes a principled foundation for measurability-aware design in aggregate spatial transcriptomics. By exposing the trade-offs between accuracy, robustness, and experimental feasibility, it shifts panel design from heuristic marker selection toward quantitative optimization grounded in physical constraints. More broadly, the framework suggests a path toward information-theoretic approaches to spatial measurement, in which the capacity of an assay is matched to the intrinsic structure of the biological system being studied. As spatial transcriptomics continues to scale across tissues, conditions, and organisms [1, 2], such principled design strategies will be essential for turning raw throughput into reliable biological insight.

Acknowledgments

This project was supported by NIH grant R01-HG012925.

References

1. Marx V. Method of the Year: spatially resolved transcriptomics. Nat Methods. 2021;18(1):9–14.
2. Moses L, Pachter L. Museum of spatial transcriptomics. Nat Methods. 2022;.
3. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. Science. 2015;348(6233):aaa6090.
4. Eng CHL, Lawson M, Zhu Q, Dries R, Koulena N, Takei Y, et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. Nature. 2019;568(7751):235–239.
5. Codeluppi S, Borm LE, Zeisel A, La Manno G, van Lunteren JA, Svensson CI, et al. Spatial organization of the somatosensory cortex revealed by osmFISH. Nat Methods. 2018;15(11):932–935.
6. He S, Bhatt R, Brown C, Brown EA, Buhr DL, Chantranuvatana K, et al. High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. Nat Biotechnol. 2022;40(12):1794–1806.
7. Plummer JT, Dezem FS, Cook DP, Park J, Zhang L, Liu Y, et al. Standardized metrics for assessment and reproducibility of imaging-based spatial transcriptomics datasets. Nat Biotechnol. 2025; p. 1–13.
8. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016;353(6294):78–82.
9. Cleary B, Simonton B, Bezney J, Murray E, Alam S, Sinha A, et al. Compressed sensing for highly efficient imaging transcriptomics. Nat Biotechnol. 2021;39(8):936–942.
10. Zhou X, Seow WY, Ha N, Cheng TH, Jiang L, Boonruangkan J, et al. Highly sensitive spatial transcriptomics using FISHnCHIPs of multiple co-expressed genes. Nat Commun. 2024;15(1):2342.
11. Hemminger Z, Sanchez-Tam G, Ocampo HD, Wang A, Underwood T, Xie F, et al. Spatial single-cell mapping of transcriptional differences across genetic backgrounds in mouse brains. bioRxivorg. 2024;.
12. Yao Z, van Velthoven CTJ, Kunst M, Zhang M, McMillen D, Lee C, et al. A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. Nature. 2023;624(7991):317–332.
13. Qiu C, Martin BK, Welsh IC, Daza RM, Le TM, Huang X, et al. A single-cell time-lapse of mouse prenatal development from gastrula to birth. Nature. 2024;626(8001):1084–1093.
14. Hershberg EA, Camplisson CK, Close JL, Attar S, Chern R, Liu Y, et al. PaintSHOP enables the interactive design of transcriptome- and genome-scale oligonucleotide FISH experiments. Nat Methods. 2021;18(8):937–944.

15. Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. Imaging individual mRNA molecules using multiple singly labeled probes. Nat Methods. 2008;5(10):877–879.

16. Young AP, Jackson DJ, Wyeth RC. A technical review and guide to RNA fluorescence in situ hybridization. PeerJ. 2020;8:e8806.

17. Yilmaz LS, Noguera DR. Mechanistic approach to the problem of hybridization efficiency in fluorescent in situ hybridization. Appl Environ Microbiol. 2004;70(12):7126–7139.

18. Choi HMT, Schwarzkopf M, Fornace M, Acharya A, Artavanis G, Stegmaier J, et al. Third-generation in situ hybridization chain reaction: multiplexed, quantitative, sensitive, versatile, robust. Development. 2018;145(12):dev165753.

19. Ke R, Mignardi M, Pacureanu A, Svedlund J, Botling J, Wählby C, et al. In situ sequencing for RNA analysis in preserved tissue and cells. Nat Methods. 2013;10(9):857–860.

20. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science. 2018;361(6400).

21. Biancalani T, Scalia G, Buffoni L, Avasthi R, Lu Z, Sanger A, et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. Nat Methods. 2021;18(11):1352–1362.

22. Kleshchevnikov V, Shmatko A, Dann E, Aivazidis A, King HW, Li T, et al. Cell2location maps fine-grained cell types in spatial transcriptomics. Nat Biotechnol. 2022;40(5):661–671.

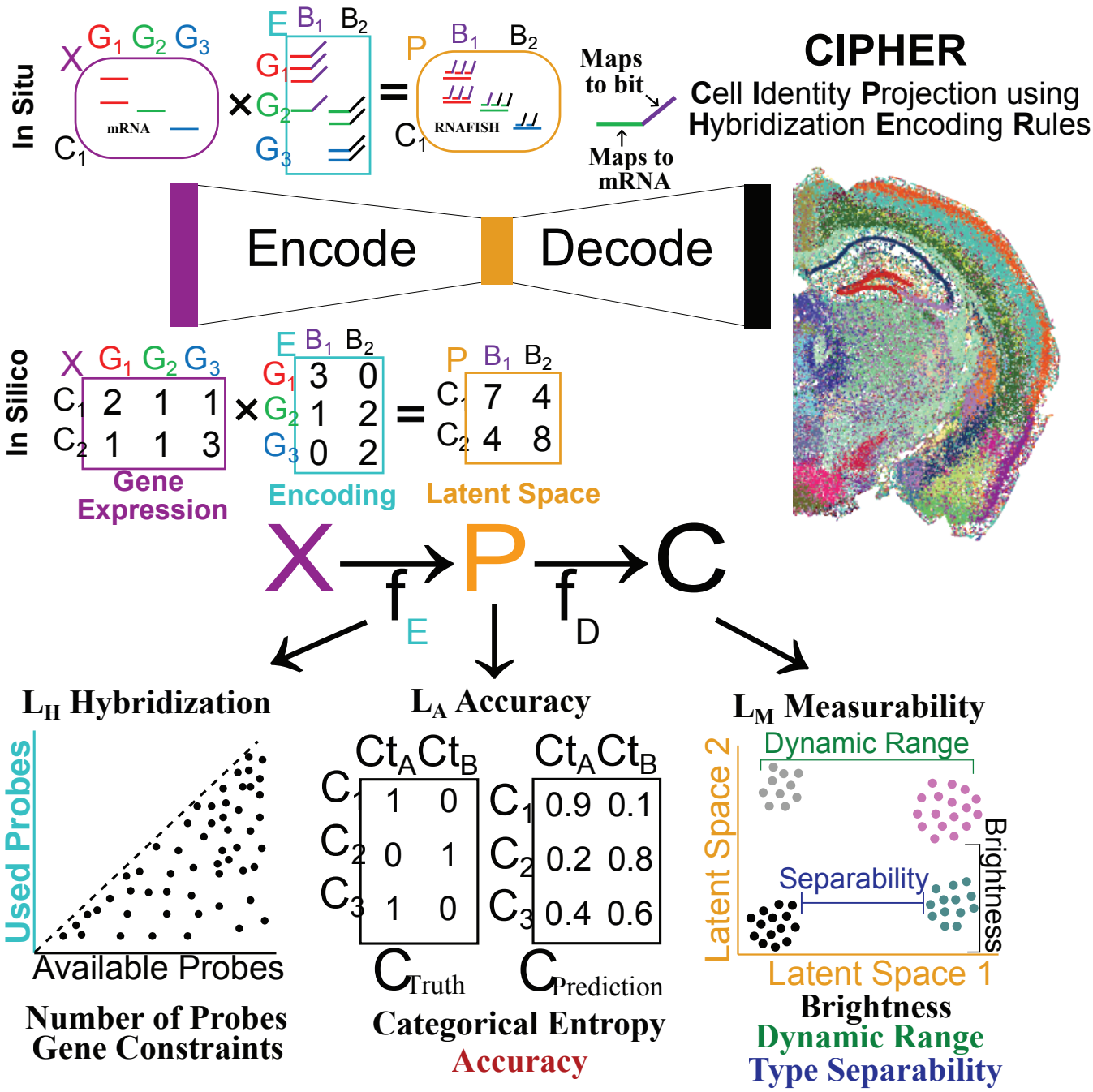

Fig 1. Figure 1. The CIPHER Framework. Schematic of the end-to-end neural network for aggregate spatial transcriptomics design. The top panel illustrates the principle of aggregate FISH, demonstrating how molecular probes physically implement a linear weighted sum projection of gene expression *in situ* prior to measurement. The bottom panel details the CIPHER framework, which employs a machine learning approach to fit the optimal encoder matrix (E) using single-cell RNA-seq reference inputs (X). This optimization maximizes decoding accuracy (L_A) while strictly enforcing physical hybridization constraints (L_H) and ensuring experimental measurability (L_M), yielding a latent space (P) that is both informative and physically realizable. The decoder (D) subsequently reconstructs cell identity (C) from these optimized signals.

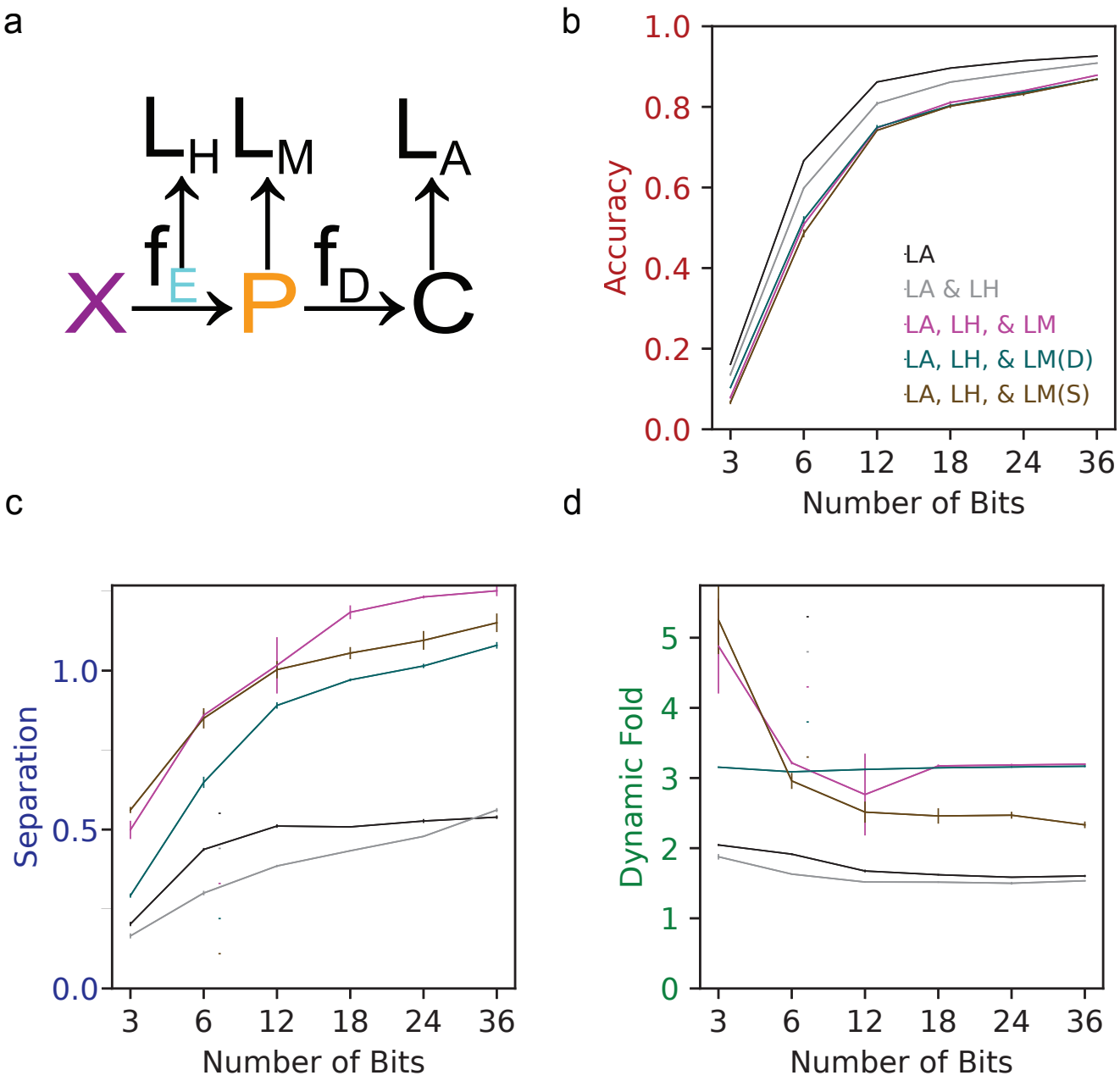

Fig 2. Figure 2. Impact of Loss Functions on Encoding Performance. (A) Schematic of the composite loss function components: L_H (Hybridization Rules) enforces physical probe constraints such as global probe budgets and per-gene limits; L_M (Measurability) enforces sufficient dynamic fold and type separation as well as overall brightness; and L_A (Accuracy) maximizes cell type classification. (B–D) Performance metrics plotted against the number of bits. All three panels display the same five loss function configurations to isolate the specific impact of each term: Accuracy only (L_A); Accuracy + Hybridization (L_A, L_H); and the full model (L_A, L_H, L_M), including variants with only Dynamic Range or Separability terms enabled. (B) Classification accuracy shows a minor decrease with the addition of measurability terms. However, this trade-off yields significant gains in (C) latent space separability and (D) dynamic fold change, showing how L_M ensures sufficient signal dynamic range for robust readout compared to unconstrained encoders.

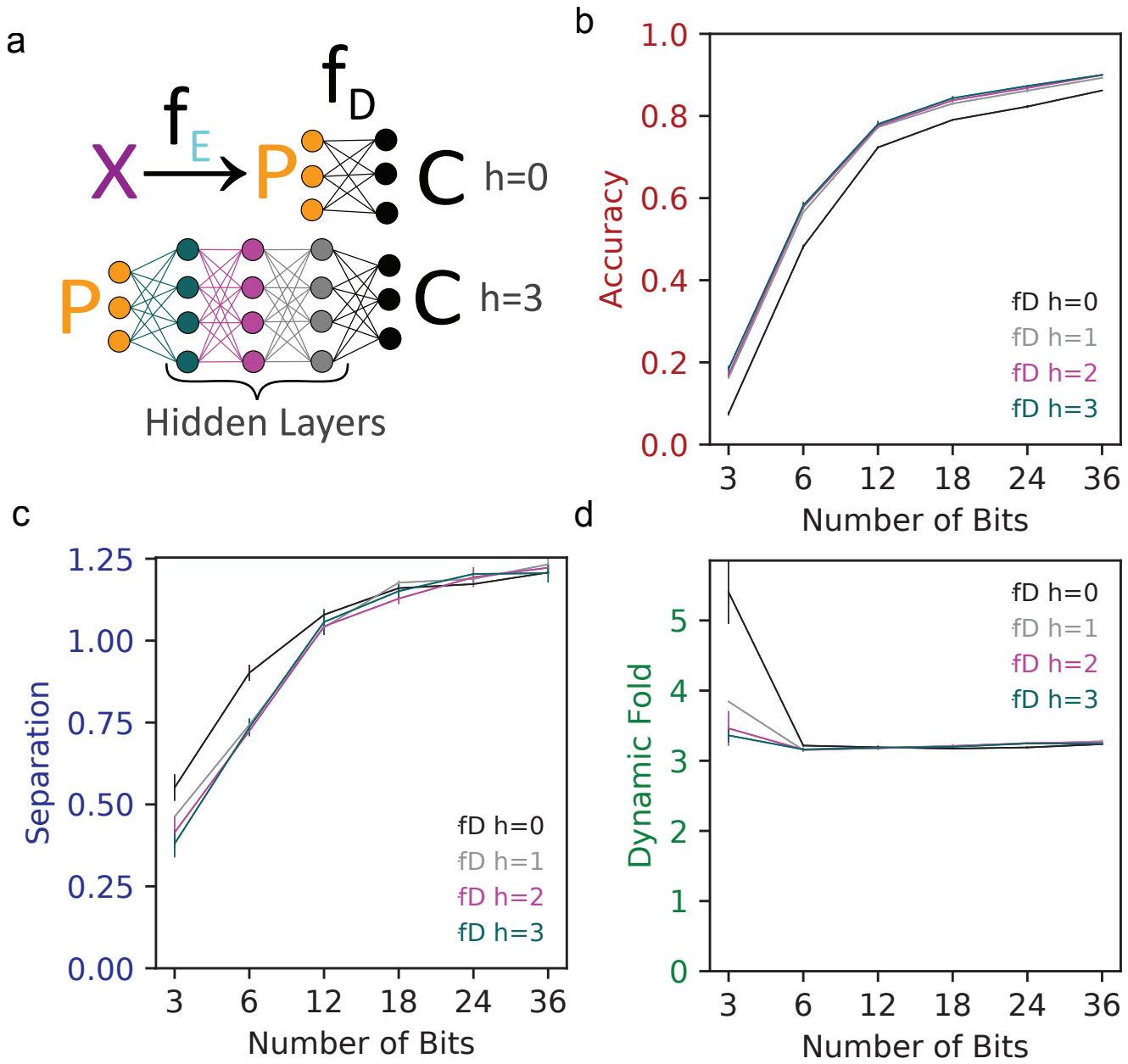

Fig 3. Figure 3. Linear vs Non Linear Decoder Complexity. (A) Schematic comparing decoder architectures with varying depths (hidden layers, h). A linear decoder ($h = 0$) represents simple weighted summation, while non-linear decoders ($h = 3$) capture combinatorial patterns. (B–D) Performance metrics plotted against the number of bits (imaging rounds) for different decoder complexities. (B) Classification accuracy improves with bit number but plateaus differently based on complexity. While there is a slight performance increase from a linear ($h = 0$) to a non-linear decoder ($h = 1$), increasing complexity further ($h > 1$) yields virtually no additional gain. (C) Latent space separability metrics. (D) The relationship between available bits and the required dynamic fold change, illustrating the trade-off between decoder complexity and signal constraints.

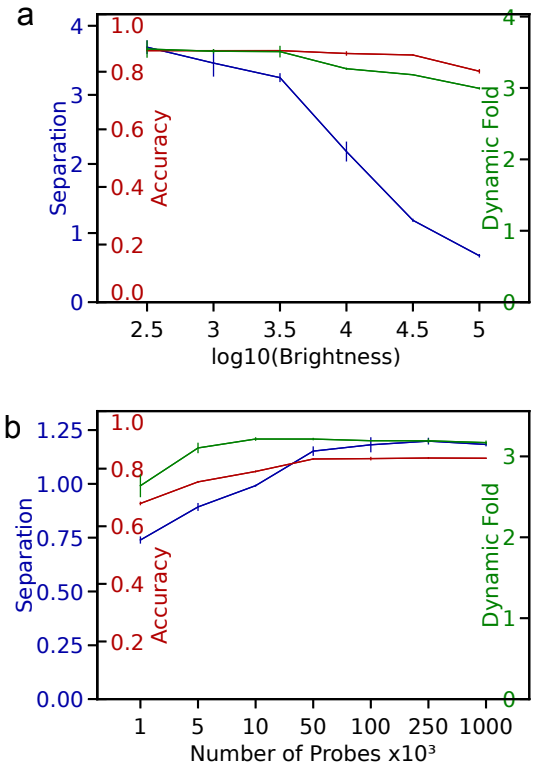

Fig 4. Figure 4. Experimental Constraints and Probe Design. (A) Analysis of encoding performance as a function of target signal brightness. The x-axis represents the target median brightness (log10). The left y-axis displays latent space separability (blue) and classification accuracy (red), while the right y-axis displays the dynamic fold change (green). Results indicate a significant advantage to lower brightness requirements, which are determined by microscope sensitivity and hybridization signal efficiency. (B) Analysis of encoding performance as a function of the total probe budget. The x-axis represents the maximum number of encoding probes ($\times 10^3$). Axes for separability (blue, left), accuracy (red, left), and dynamic fold (green, right) correspond to those in (A). The analysis demonstrates that performance metrics plateau at approximately 50,000 probes, indicating a diminishing return for larger probe libraries.

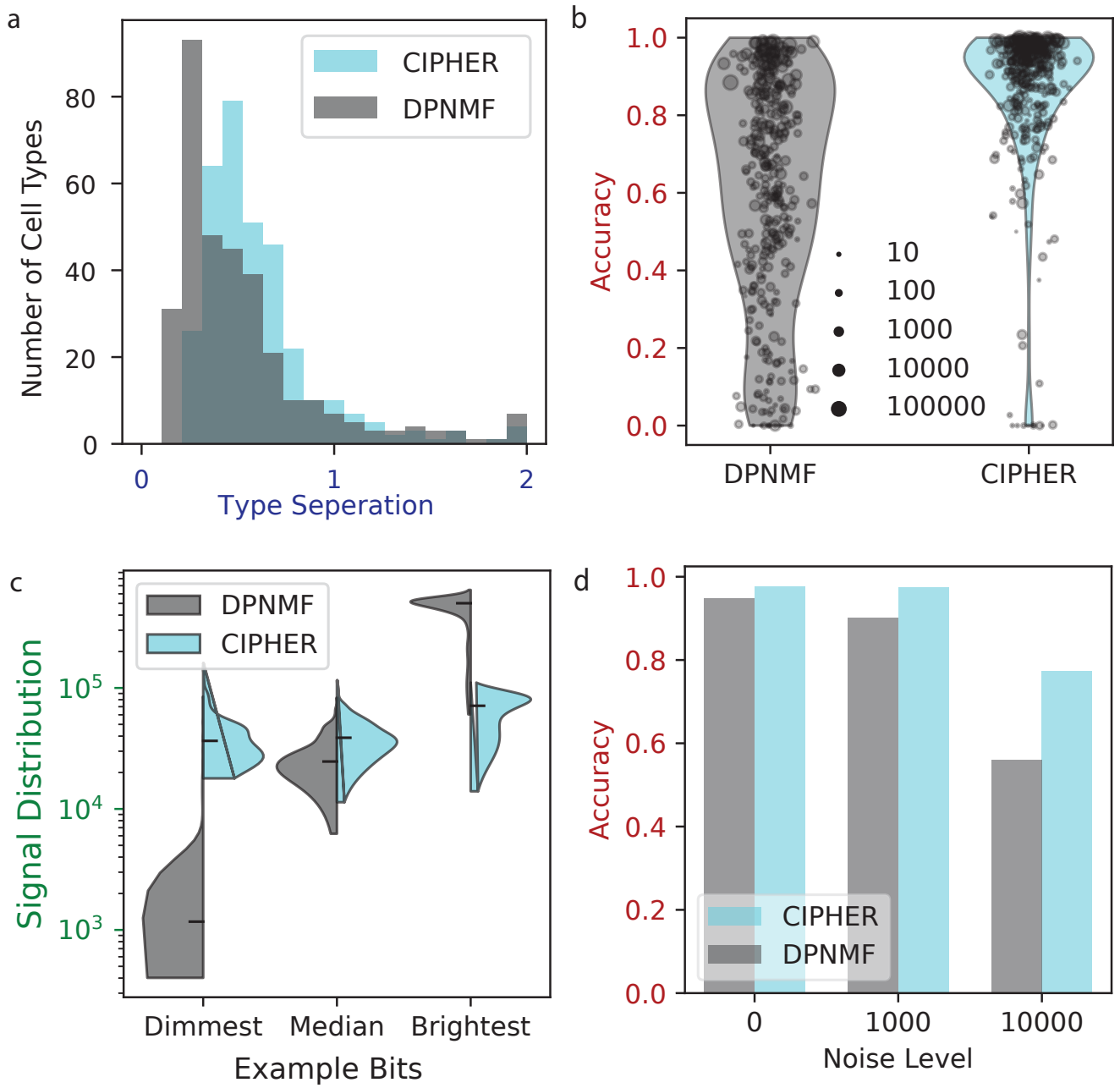

Fig 5. Figure 5. Comparison of CIPHER and DPNMF Performance. (A) Histogram of pairwise distances between cell type centroids in the learned latent space. DPNMF designs exhibit an increased number of cell types with separations near zero compared to CIPHER. (B) Classification accuracy broken down by individual cell types in simulation, demonstrating that a significantly larger proportion of cell types achieve higher accuracy in CIPHER compared to DPNMF. (C) Analysis of individual bit brightness. CIPHER encodings maintain consistent signal brightness across bits, whereas DPNMF encodings display high variance with significant disparities between bits. (D) Simulation performed using imputed MERFISH data showing expected accuracy under varying levels of additive noise. Results demonstrate that CIPHER yields higher accuracy and increased robustness to noise compared to DPNMF.

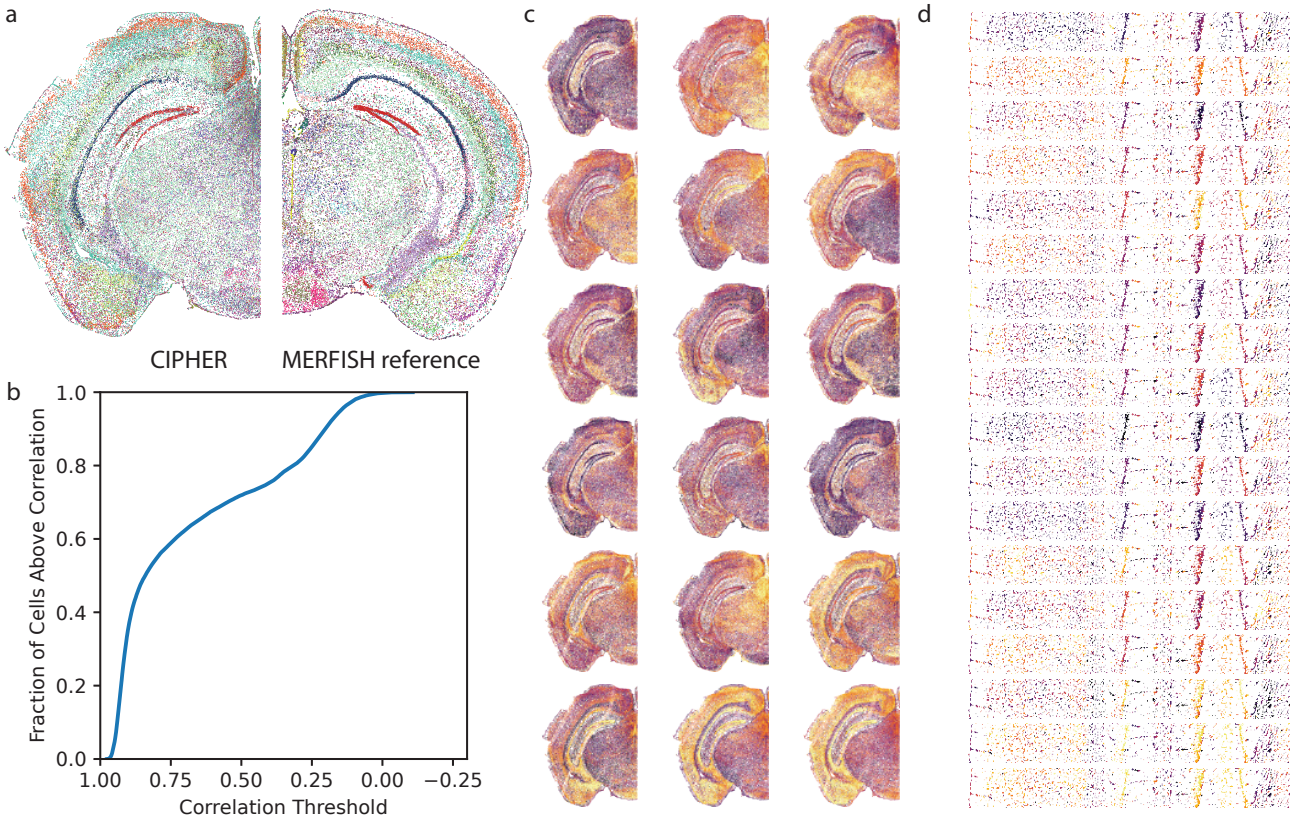

Fig 6. Figure 6. Experimental Validation in Mouse Cortex. (A) Left: Experimental decoding results for a representative half-coronal section, where each dot represents a single cell colored by its assigned cell type identity. Right: Corresponding reference map generated from MERFISH data for the same anatomical region, serving as a ground-truth comparison. (B) Agreement metric showing the concordance between CIPHER measurements and reference data. Cells were greedily paired across datasets based on spatial proximity, and the correlation of their gene expression profiles was calculated to quantify alignment quality. (C) Visualization of the measured latent space signals across the tissue section. The heatmap displays intensity values for all bits, with dimmer cells shown in purple and brighter cells in yellow, illustrating the combinatorial encoding. (D) High-resolution zoom of panel (D) highlighting the cortex and hippocampus. The inset reveals individual cells and the distinct combinatorial encoding patterns that distinguish cortical layers and hippocampal subfields.

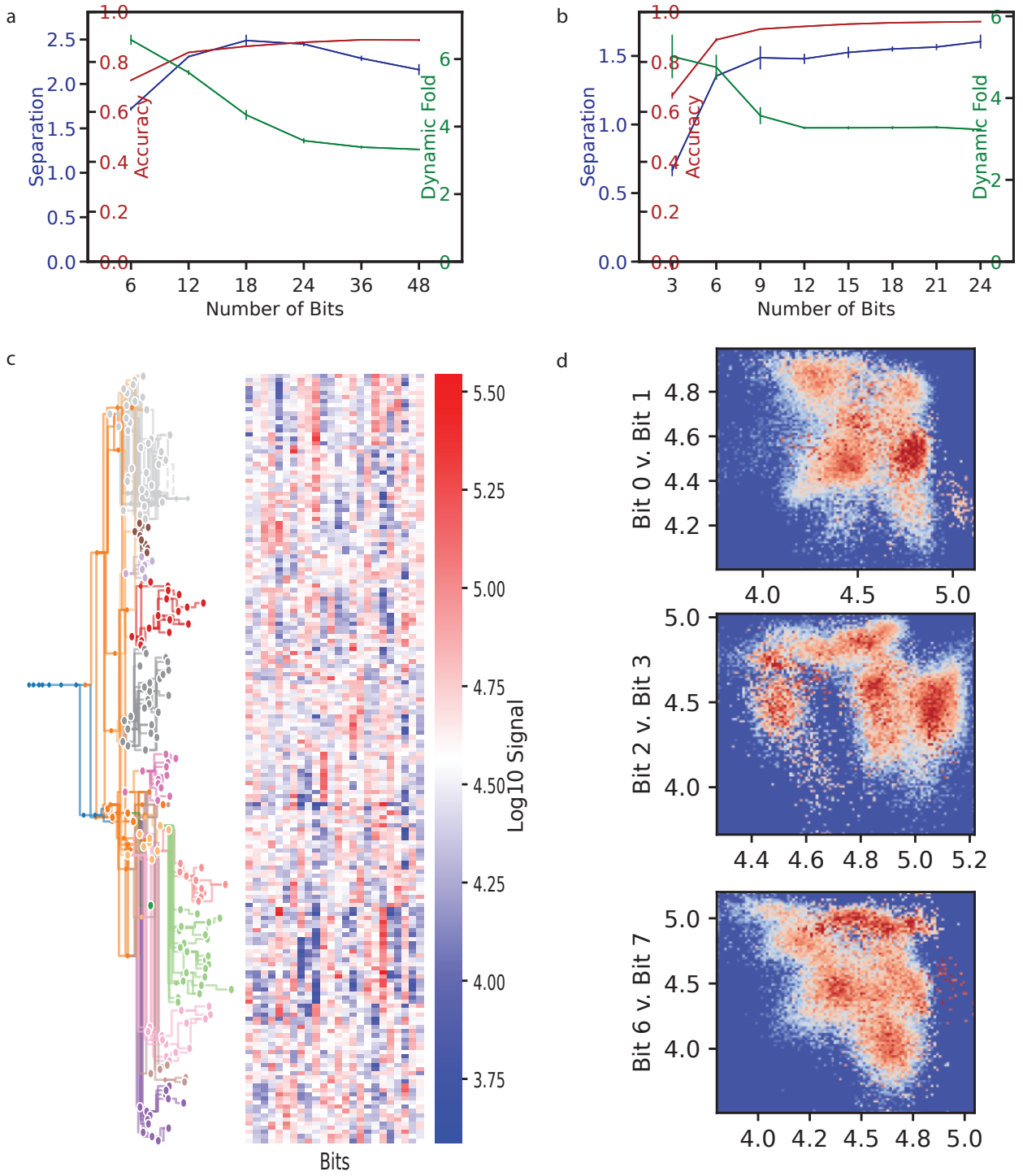

Fig 7. Figure 7. Generalization to Alternative Tissues. Demonstration of the CIPHER framework's applicability to diverse biological contexts. (A) Performance metrics for the developing mouse brain dataset plotted against the number of bits (x-axis). Similar to Figure 4, axes display Separability (blue, left), Accuracy (red, left), and Dynamic Fold (green, right). Results indicate that performance saturates at approximately 24 bits. (B) Heatmap of the learned latent vectors (P) organized by the developmental tree, visualizing the structured assignment of developmental progenitors to specific bits. (C) Performance metrics for the lymph node dataset plotted against the number of bits, showing that performance saturates at approximately 9 bits. (D) Density visualization of the projection space for the lymph node data, demonstrating significant separation between distinct cell populations across the latent bits.