PROPOSE: Predictive and robust probe selection for spatial transcriptomics

Ian Covert[1], Rohan Gala[2], Tim Wang[3], Karel Svoboda[3,4], Uygar Sümbül[2,*], and Su-In Lee[1,*]

[1]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA
[2]Allen Institute for Brain Science, Seattle, WA, USA
[3]HHMI Janelia Research Campus, Ashburn, VA, USA
[4]Allen Institute for Neural Dynamics, Seattle, WA, USA
[*]corresponding authors: uygars@alleninstitute.org (U.S.), suinlee@cs.washington.edu (S.-I.L.)

Abstract

Fluorescence *in situ* hybridization (FISH) is a widely used method for visualizing gene expression in cells and tissues. A key challenge is determining a small panel of genes (typically less than 1% of the genome) to probe in a FISH experiment that are most informative about gene expression, either in general or for a specific experimental objective. We introduce *predictive and robust probe selection* (PROPOSE), a method that uses deep learning to identify informative marker genes using data from single-cell RNA sequencing (scRNA-seq). Using datasets spanning different brain regions, species, and scRNA-seq technologies, we show that our method reliably identifies gene panels that provide more accurate prediction of the genome-wide expression profile, thereby capturing more information while using fewer probes. Furthermore, PROPOSE can be readily adapted to meet specific experimental goals, such as classifying cell types or discerning neuronal electrical properties from scRNA-seq data. Finally, we demonstrate using a recent MERFISH dataset that PROPOSE's binarization of gene expression levels enables models trained on scRNA-seq data to generalize with input data obtained via FISH, despite the complex domain shift between these technologies.

Introduction

Cell type classification has been revolutionized by recent advances in single-cell genomics. It is now possible to profile the entire mRNA repertoire (i.e., the transcriptome) of tens, or even hundreds, of thousands of individual cells (scRNA-seq) in a single experiment. Large-scale scRNA-seq studies have provided high-resolution taxonomies of the transcriptomic cell types in many tissues across several species. Leveraging data from these scRNA-seq studies, spatial transcriptomic methods, in particular fluorescence *in situ* hybridization (FISH), can be applied to examine molecularly defined cells in their native context [1, 2, 3, 4, 5, 6]. FISH reveals the spatial organization of cells within tissues, thereby helping uncover the fundamental principles of brain organization [7]. In conjunction with other experimental modalities (e.g., morphological, connectivity, or electrophysiological studies), FISH can also link transcriptomic identity with other data to provide a better understanding of the functional role of individual cell types. [6, 8, 9, 10, 11].

Whereas scRNA-seq detects genes in a largely unbiased manner, FISH assays a pre-defined list of genes. Notably, only a tiny fraction of the transcriptome is probed in an experiment: the largest studies survey <5% of transcribed genes, with more typical studies examining <0.2%; this is in part because the complexity and duration of FISH experiments increases sharply with the number genes probed. Thus, judicious selection of a small number of highly informative marker genes to probe (i.e., a gene panel) is key for successful FISH experiments. Experimentalists often rely on *ad hoc* approaches to gene selection, most commonly choosing markers based on prior knowledge or very high expression in a limited subset of cells. Such methods are suboptimal as they tend to overlook genes with more complex expression patterns and rarely account for correlated expression between selected genes, which yields redundant information. Here, we frame the identification of markers as a feature selection problem and seek to address it in a principled manner.

Feature selection problems arise in many domains [12, 13, 14, 15]. However, FISH experiments present unique challenges and thus demand a specialized solution. Importantly, because reference datasets consisting of FISH-based measurements for thousands of genes are unavailable, scRNA-seq datasets are used instead to guide the selection process. The use of a surrogate dataset presents new obstacles: the expression counts observed by scRNA-seq and FISH may differ significantly, with a relationship that is nonlinear and noisy [16, 17, 18]. Hence, a gene panel selected without considering the difference between the datasets is unlikely to perform as well as intended, as we demonstrate with several existing methods.

Moreover, the optimal gene panel should account for specifics of the target experiment, which may demand tuning of the selection criterion. For instance, linking of spatial characterization of gene expression through FISH and the electrical properties of neurons [9, 10] may require a gene panel that relates to membrane excitability. Another scenario could involve investigating a specific subclass of cells. For example, exploring neurons expressing a specific marker gene in a particular brain region demands a gene panel based on reference data from this molecularly and spatially constrained population. Furthermore, certain FISH methods might require probes that have either relatively high or low expression levels.

We address these challenges by introducing *predictive and robust probe selection* (PROPOSE), an unsupervised algorithm for selecting genes for FISH studies using scRNA-seq data as a reference. PROPOSE uses deep learning to find a small number of highly informative genes whose expression can predict the genome-wide expression profile. PROPOSE trains a reconstruction model with a loss function that accounts for noisy gene dropouts[1] in scRNA-seq; incorporates expert knowledge by pre-selecting or pre-filtering genes; scales to very large datasets using minibatched training; and quantizes gene expression levels to account for the domain shift between scRNA-seq and FISH. Furthermore, our deep learning-based selection mechanism is flexible: by changing the prediction target and loss function, PROPOSE can also operate in a supervised manner to address specific experimental aims, such as cell type classification or electrophysiological characterization.

We validate our approach using reference datasets from different technologies (plate-based SmartSeq and droplet-based 10X), multiple brain regions (V1, ALM, MOp) and different species (mouse, human) on classification and reconstruction tasks. We then highlight PROPOSE's flexibility and show how to incorporate a different data modality (electrophysiology) with FISH experiments using a large Patch-seq dataset [9]. Finally, we devise a novel evaluation procedure to showcase the effectiveness of our robust inference approach based on gene quantization using a recent MERFISH dataset [6].

[1] Gene 'dropout' refers to the failure to detect mRNA transcripts due to inefficiencies in cDNA library preparation; dropouts are more prevalent when using lower resolution (and typically higher throughput) platforms such as 10X [19, 20, 21].

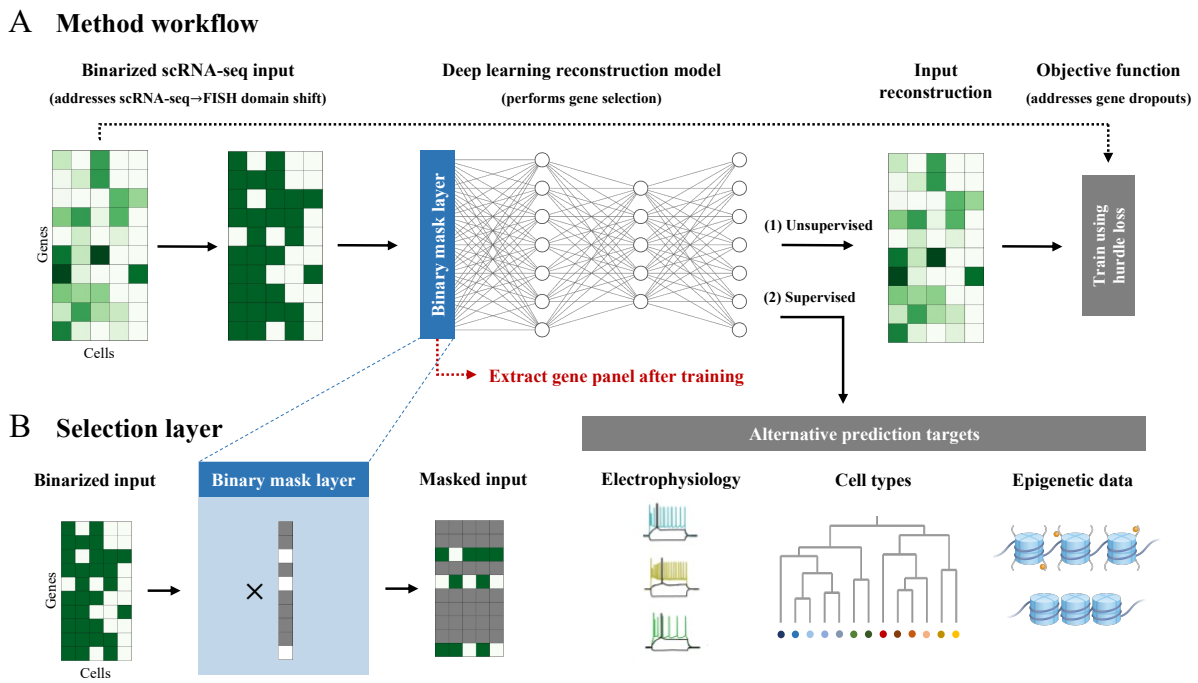

Figure 1: **Overview of the *predictive and robust probe selection* (PROPOSE) method.** PROPOSE uses deep learning to select highly informative genes that can approximately reconstruct the genome-wide expression profile. **A**, The model input is scRNA-seq data, which is binarized to address the domain shift between scRNA-seq and FISH data. By default, the model's output is a reconstruction of the original scRNA-seq gene expression levels, and the objective function for training is a hurdle loss designed to account for noisy gene dropouts. Alternatively, a supervised prediction target can be used to address specific experimental aims such as cell type classification. After training, the selected gene panel is extracted from the model's binary mask layer. **B**, The binary mask layer is responsible for selecting genes, and it does so by multiplying the input by a learned binary mask that controls which input features are used by the model.

Results

Selecting genes using deep learning

Given an scRNA-seq dataset of a cell population to be profiled using FISH, PROPOSE selects a small panel of genes that can optimally reconstruct the entire scRNA-seq expression profile. Intuitively, such gene panels should be useful for a variety of downstream tasks because they sacrifice minimal information. Our approach is inspired by classical dimension-reduction techniques like principal components analysis (PCA) [22], but PROPOSE selects a small number of genes rather than using linear combinations. Additionally, it reconstructs the original data using a non-linear model and with a quality measure more appropriate for scRNA-seq data[2]: we implement a *hurdle loss function* to account for noisy gene dropouts [25]. Our loss function involves separately predicting each gene's expression level and whether it is actually expressed, which lets the model explicitly represent dropout noise in its predictions (see Methods).

PROPOSE uses a deep learning model with a custom layer designed to pinpoint a small number of useful input features (Fig. 1). This approach is inspired by recent work on *differentiable feature selection*[3] [14, 26]

[2]PCA measures reconstruction quality using a mean squared error (MSE) loss, which recent work has found to be ill-suited for scRNA-seq [23, 24].

[3]Differentiable feature selection refers to the use of gradient-based optimization; it is unrelated to differential expression analysis.

and we train the model using gradient descent [27, 28], which enables the method to scale to very large datasets through the use of minibatched training. The selection layer applies a learned binary mask that sparsifies over the course of the optimization process; information initially flows from all genes, but the relevant inputs are gradually reduced down to a user-specified number (Fig. 1B). In practice, training the model with a very large number of candidate genes can consume a large amount of computer memory, but PROPOSE can optionally use a selection layer with fewer learnable parameters to reduce the set of candidates before selecting the final panel (see Methods).

By default, PROPOSE operates in an unsupervised manner by reconstructing the full scRNA-seq dataset without using any labels or manual annotation. However, PROPOSE can also operate in a supervised manner by incorporating cell-level annotations as the model's prediction target, such as cell type labels or complementary epigenetic data like chromatin accessibility and methylation (Fig. 1A). As we show in our experiments, this allows PROPOSE to select genes that are tailored to meet specific experimental objectives.

We evaluate the PROPOSE gene panels by simulating their use in a FISH study. Such an evaluation is difficult because scRNA-seq and FISH have very different noise sources and detection issues; for the same cell, the number of transcript counts observed by each technology can differ substantially. In order to mitigate the domain shift, we binarized gene expression levels, recording only whether a gene was detected in each cell. This pre-processing step is crucial to evaluate the utility of the gene panel in a FISH study, and we implement it by binarizing expression levels at the input layer of the PROPOSE deep learning model (Fig. 1A). After selecting a gene panel and collecting data using FISH, the newly collected data can be binarized on a gene-by-gene basis by matching the quantiles of the scRNA-seq thresholds; this allows models to transfer across technologies despite the domain shift, as we demonstrate in an experiment with MERFISH data.

Our experiments compare PROPOSE to several state-of-the-art gene selection methods. We tested the Seurat [29] and Cell Ranger [30] gene selection procedures, which are based on per-gene variance and dispersion levels and are implemented in the popular ScanPy package [31]. Next, we tested the recently proposed geneBasis method [32] that selects genes using a greedy algorithm to preserve the data manifold. Finally, we considered two methods that aim to differentiate cell types: a method that maximizes the information about cell type labels[4] (MutInfo) [6, 34], and the scGeneFit method [35] that uses linear programming. These methods span a range of selection criteria, but PROPOSE is the first method that offers the flexibility to be used in a supervised or unsupervised manner, and which is designed explicitly to enable transferability to FISH studies.

PROPOSE enables more accurate scRNA-seq expression profile reconstruction

We first tested PROPOSE on two scRNA-seq datasets: a SmartSeq v4 [36] dataset consisting of 22,160 neurons from the mouse primary visual (V1) and anterior lateral motor (ALM) cortices [37] (hereafter referred to as **SSv4**), and a 10X [30] dataset consisting of 72,629 neurons from the human motor (M1) cortex [38] (hereafter referred to as **10X**). These datasets profile different brain regions and species and were collected using two library preparation platforms that yield different levels of sparsity. Working with an initial set of 10,000 high-variance genes, we used PROPOSE and the other gene selection methods to identify panels of 8-256 marker genes, a range that spans the vast majority of FISH studies.

As a benchmark metric for our comparisons, we calculated the portion of variance in the full scRNA-seq expression profile that could be explained using each selected gene panel. For gene panels of all sizes on

[4]This is measured via *mutual information* [33].

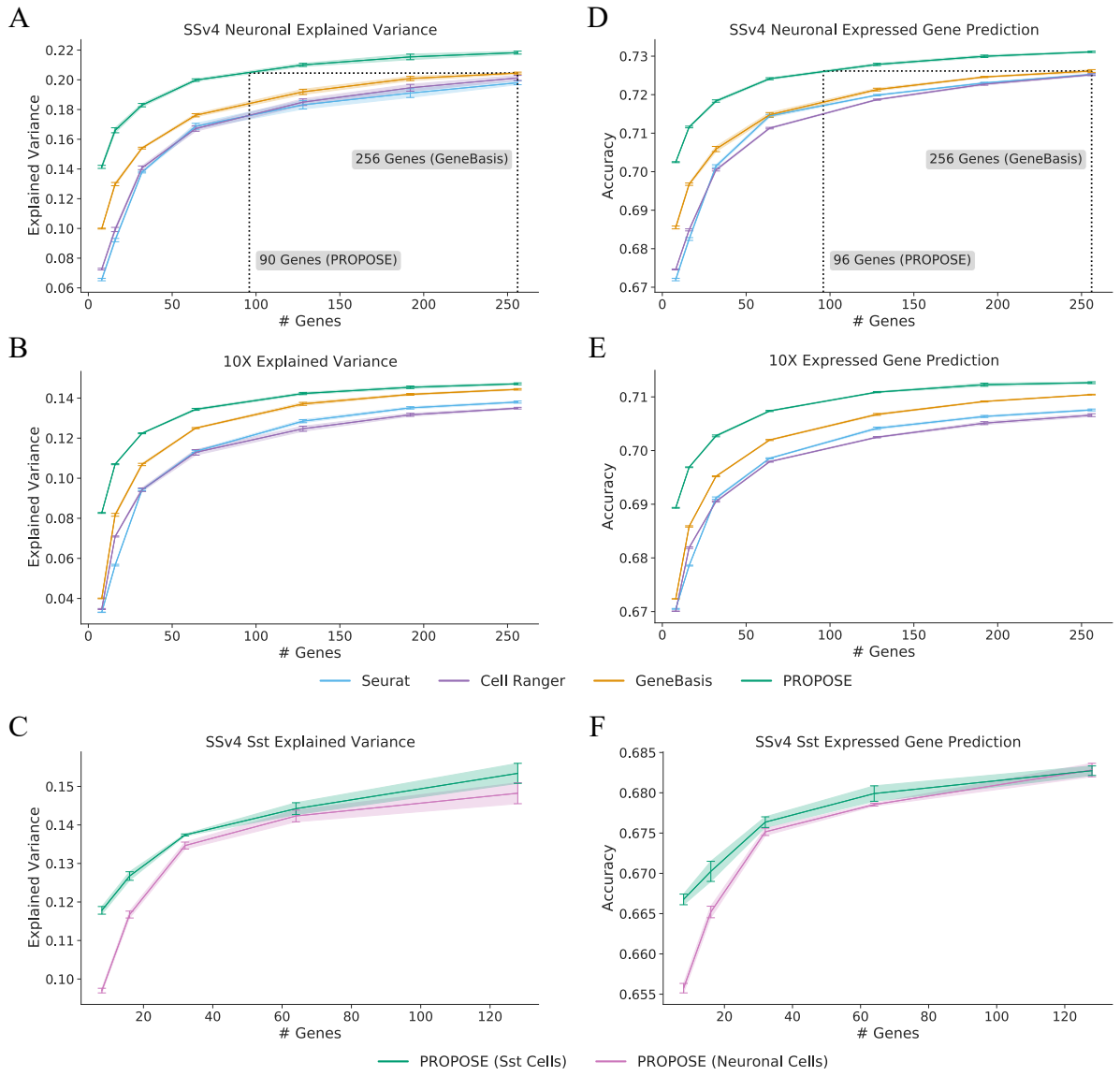

Figure 2: **Evaluating gene panels based on their ability to reconstruct scRNA-seq expression profiles.** The selected genes are used to predict either log-normalized expression counts for the remaining genes (**A**-**C**), or whether each gene is expressed (**D**-**F**). Expression levels are binarized for the selected genes to ensure transferability to FISH experiments. The panels selected by PROPOSE perform better than those selected by previous methods. **A**, Explained variance for gene panels identified by PROPOSE and other methods with the SSv4 V1/ALM neuronal cells (higher is better). **B**, Explained variance for gene panels with the 10X M1 cells. **C**, Explained variance for gene panels with the SSv4 V1/ALM Sst cells. **D**, Expressed gene prediction accuracy with the SSv4 V1/ALM neuronal cells (higher is better). **E**, Expressed gene prediction accuracy for panels with the 10X M1 cells. **F**, Expressed gene prediction accuracy for panels with the SSv4 V1/ALM Sst cells.

both datasets, PROPOSE outperformed Seurat, Cell Ranger and geneBasis (Fig. 2A-C). The geneBasis approach is most competitive with PROPOSE, but it explained considerably less variance for smaller gene panels, which are most commonly used experimentally. There is diminishing improvement as the panel size increases, and even large panels, such as those with 256 genes, fail to explain all the variance; this is due not only to the many factors of variation in the full expression profiles, but to high noise levels in the data.

Importantly, PROPOSE binarizes gene expression levels during training whereas Seurat, Cell Ranger and

geneBasis all use either raw or logarithmized expression counts. We asked whether this pre-processing step could account for differences in performance; a modified version of geneBasis becomes more competitive with PROPOSE, whereas Seurat and Cell Ranger perform worse (Supp. Fig. 11). For MutInfo and scGeneFit, which leverage cell type labels to select genes, we find that PROPOSE outperforms both methods independent of binarization (Supp. Fig. 10). Overall, these results show that PROPOSE's binarization step can be incorporated into several of the baselines to enable better transferability to FISH studies.

Recent work has demonstrated that the digital "dropout pattern" of a cell (i.e., the set of genes not detected by scRNA-seq) is nearly as informative as quantitative expression levels for identifying cell types [20, 39]. We thus wondered whether the selected gene panels could predict the set of genes with non-zero transcript counts in the rest of the scRNA-seq dataset (Fig. 2D-F). In this task, PROPOSE again performs more accurately than all other methods on both datasets for panels of all sizes; this is due in part to our hurdle loss function, which involves predicting whether each gene is detected. For this analysis, we excluded genes that were rarely expressed and housekeeping genes that are ubiquitously expressed, focusing on those expressed in 20-80% of cells (n=4,972 genes), but similar results were found for varying cutoffs (Supp. Fig. 18).

Finally, from an experimental standpoint, it would be most practical to design FISH probes targeting a single 'general-purpose' gene panel selected using the entire dataset rather than generating probes for each particular subtype of interest. To assess the feasibility of this strategy, we focused on somatostatin-expressing (Sst) interneurons (2,701 cells) in the SSv4 dataset and compared gene sets selected by PROPOSE when trained on the entire scRNA-seq dataset versus just the Sst subpopulation. Performance was better when only data from the specific cell type of interest was included during gene selection. However, the improvement diminished when 32 or more genes were selected, which suggests that general purpose gene panels may be appropriate for methods that assay large numbers of genes (e.g., MERFISH).

PROPOSE enables accurate cell type classification

As another evaluation metric, we tested how accurately the gene panels selected by each method can classify cell types, which is a common goal of FISH studies [2, 4, 6, 40]. For all gene selection approaches, larger panels enabled increasingly accurate cell type classification (Fig. 3). As expected, supervised methods that use cell type annotations during their selection procedure (e.g., scGeneFit, MutInfo) perform better than unsupervised methods that use only unlabeled scRNA-seq data (e.g., Seurat, Cell Ranger). To emphasize this distinction between methods, we present results separately for unsupervised (Figs. 3A, C) and supervised methods (Figs. 3B, D).

Among the unsupervised approaches, PROPOSE outperforms Seurat, Cell Ranger and geneBasis for panels of all sizes. For example, using 64 genes, PROPOSE reaches 74% accuracy with the SSv4 data and 78% with the 10X data considering a total of 113 and 117 cell types, respectively. geneBasis is the most competitive unsupervised baseline with the 10X data, and either geneBasis or Seurat are most competitive with the SSv4 data. The gap in performance is largest for small panels, and the various methods roughly converge in accuracy for panels of 128 genes.

PROPOSE is an unsupervised gene selection method by default, but we can also modify it for cell type classification by explicitly using cell type labels as the prediction target. This supervised version of our approach, termed PROPOSE-Classification, matches or exceeds the performance of all other supervised approaches. For example, it reaches 81% accuracy in the SSv4 dataset and 82% accuracy in the 10X dataset using panels with 64 genes—a significant improvement over the unsupervised version. This illustrates how our

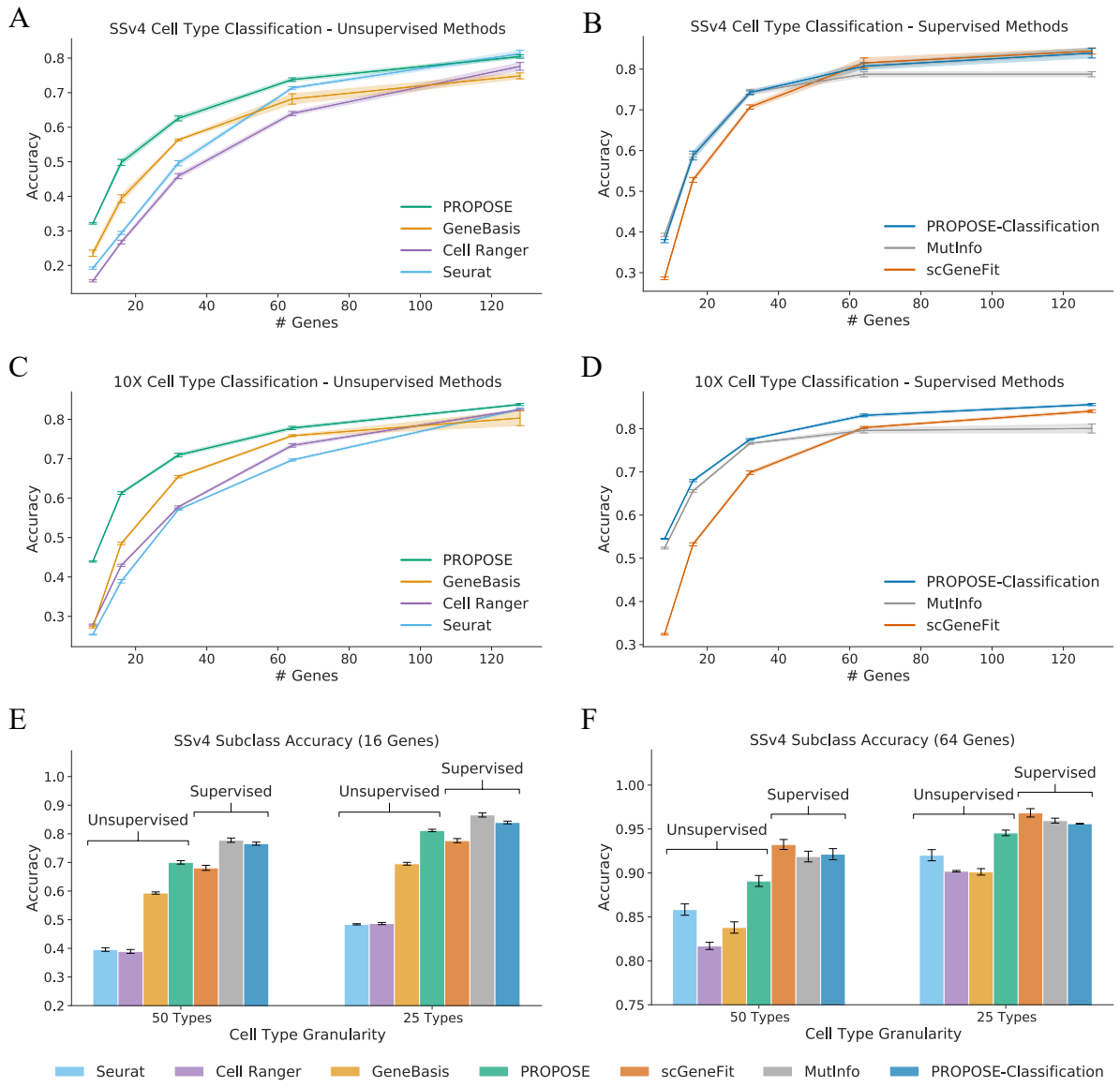

Figure 3: **Evaluating gene panels based on their ability to classify cell types.** The selected genes are used to classify scRNA-seq cell types, with binarized expression counts used to ensure transferability to FISH experiments. PROPOSE achieves strong performance despite not using cell type labels during the selection process, and PROPOSE-Classification, a version modified to distinguish cell types, matches the best existing supervised approaches (MutInfo, scGeneFit). **A**, Cell type classification accuracy for gene panels selected by unsupervised methods with the SSv4 dataset (higher is better). **B**, Accuracy for gene panels selected by supervised methods with the SSv4 dataset. **C**, Accuracy for unsupervised gene panels with the 10X dataset. **D**, Accuracy for supervised gene panels with the 10X dataset. **E**, Accuracy with SSv4 cell subtypes for panels of 16 genes. **F**, Accuracy with SSv4 cell subtypes for panels of 64 genes.

deep learning-based selection approach is flexible, in that PROPOSE can be adapted to specific experimental objectives by adjusting its prediction target and loss function. It also suggests that gene panels of these sizes ($\leq$ 128 probes) may be incapable of perfectly distinguishing fine-grained cell types, and therefore that FISH studies may benefit from analyzing more coarse clusterings.

Prior work suggests that the highly similar terminal nodes of the classification hierarchy may not all cor-

respond to distinct cell types, but may instead reflect cell states or spatial gradients in gene expression [41]. We therefore divided neurons into broader subclasses, which leads to greater classification accuracy because the groupings are more distinct and, trivially, there are fewer of them. For the SSv4 dataset, if we classify neurons into 25 subclasses rather than the full 113 cell types, PROPOSE-Classification reaches 84% accuracy using 16 genes (vs. 59% for 113 types) and 96% accuracy using 64 genes (vs. 81% for 113 types). In comparison, the unsupervised version of PROPOSE provides accuracy just 3% and 2% worse, respectively (Fig. 3E-F). The results are similar but slightly less accurate when we classify into 50 subclasses. With a reduced number of cell type subclasses, Seurat, Cell Ranger and geneBasis are still not competitive with PROPOSE, and PROPOSE-Classification remains on par with the supervised MutInfo and scGeneFit procedures.

Although PROPOSE does not match PROPOSE-Classification in terms of classification accuracy, it is notable that PROPOSE remains competitive despite not having access to cell type labels. This indicates that PROPOSE successfully captures cell type information in an unsupervised manner. We attribute the superior performance over the other unsupervised methods to our deep learning-based selection mechanism, which identifies non-redundant genes that help reconstruct the full expression profile, and to our use of gene expression binarization. In an ablation study, we also find that PROPOSE's hurdle loss function is an important design choice, because it leads to better cell type classification accuracy than training with mean squared error loss (Supp. Fig. 9). These results are promising because a consensus definition of cell types, and their continuous versus discrete nature, are far from settled [10, 37, 42]. Moreover, reference label information is currently available for only a handful of mouse and human tissues, and PROPOSE can be used in an unsupervised manner in settings that lack established cell type hierarchies.

PROPOSE can be adapted to predict electrophysiological properties

As a further demonstration of our method's flexibility, we developed a specialized variant of PROPOSE to identify marker genes that predict a cell's electrophysiological properties. For this purpose, we used a multimodal Patch-seq dataset [9, 43] containing transcriptomic and electrophysiological information from 3,411 GABAergic neurons across 53 cell types in the mouse visual cortex. Specifically, the transcriptomic profile consists of the scRNA-seq counts for 1,252 curated genes, and the electrophysiological profile consists of a set of 44 sparse principal components (sPCs) summarizing different portions of the measurement protocol, as well as 24 biologically relevant features [43]. The sPC values were scaled to have unit variance, and the remaining features were individually normalized to have zero mean and unit norm.

To select gene panels using the Patch-seq dataset, we first used the baseline methods that require only unlabeled expression data – Seurat, Cell Ranger and geneBasis. Next, we ran PROPOSE in an unsupervised manner, selecting genes that can reconstruct the full expression profile. Finally, we also ran PROPOSE in a supervised manner by using the vector of electrophysiological features as the prediction target, yielding a variant of our approach termed PROPOSE-Ephys. We then investigated how well each gene panel represents a cell's electrophysiological profile.

As an evaluation metric, we attempted to predict the electrophysiological features of each neuron using the expression levels of the genes in each panel. Similar to previous experiments, we binarize gene expression levels to simulate applicability in a subsequent FISH study. We find that PROPOSE-Ephys achieves the highest predictive accuracy, reaching a higher portion of explained variance with panels of all sizes (Fig. 4A). The unsupervised version of PROPOSE is the second most accurate method, achieving comparable explained variance for larger panels, and Cell Ranger performs similarly in this case.

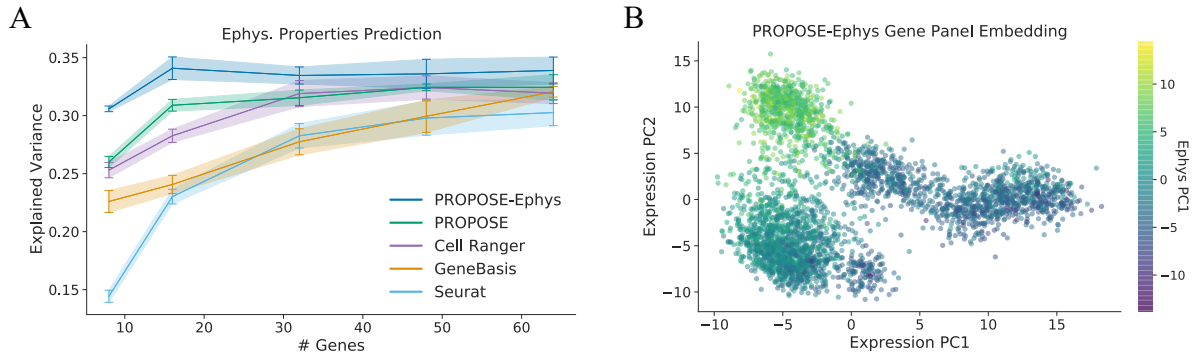

Figure 4: **Using PROPOSE to predict electrophysiological properties.** PROPOSE is adapted to this specific experimental goal by using electrophsyiological properties as its prediction target with a mean squared error loss. The resulting version of our approach, PROPOSE-Ephys, yields gene panels can predict cells' electrophysiological properties more accurately than other methods. **A**, Explained variance in the vector of electrophysiological properties for gene panels selected by each method. **B**, Low-dimensional embedding for the PROPOSE-Ephys panel containing 64 genes, colorized according to the cells' electrophysiological properties (using the first principal component); the cells are effectively clustered according to their properties, with similar cells appearing nearby in the embedding space.

The strong performance of PROPOSE-Ephys is due to it being tailored to electrophysiological characterization. Whereas PROPOSE-Ephys is designed to address this specific predictive task, the other methods rely on selection criteria that are not explicitly related to electrophysiology. As a result, these methods picked distinct genes (Supp Fig. 15) and PROPOSE-Ephys is roughly as accurate with a panel of 8 genes as the other methods are with 64 genes. None of the gene panels we tested exceed 35% explained variance, but notably, we find that even the full set of 1,252 genes does not exceed this level of accuracy; this suggests that the unexplained variance represents either noise in the experimental results, or factors that are not captured by gene expression.

As an exploratory analysis, we examined a low-dimensional embedding of the PROPOSE-Ephys gene panel to understand the relationship between cells that are nearby in the embedding space. Our plot displays individual cells using the first two principal components of the panel containing 64 genes (Fig. 4B). To assess whether nearby cells have similar electrophysiological profiles, we colorized the cells by a scalar summary of the profile – the first principal component of the electrophysiological feature vector. The resulting plot reveals naturally occurring clusters of similar cells, which may be expected because PROPOSE-Ephys selects genes whose expression is maximally indicative of the neuron's electrophysiological profile.

Binarization enables gene expression prediction with MERFISH data

Using scRNA-seq data, PROPOSE can identify informative marker genes for a number of supervised and unsupervised applications. However, the performance of these candidate genes in actual FISH experiments is unclear. It is difficult to ascertain whether a gene set selected by PROPOSE or another method is capable of accurately predicting cell-level properties given FISH data, and furthermore, it is important to determine whether binarization is sufficient to enable models trained using scRNA-seq data to generalize to FISH data. To investigate these questions, we devised an multi-step *in silico* experiment using the SSv4 scRNA-seq dataset in combination with data from a recent, large-scale MERFISH study [6].

In the MERFISH dataset, 258 genes were probed across 280,327 cells from the mouse primary motor cortex (MOp). Because ground truth cell type labels are not available, we instead chose to evaluate perfor-

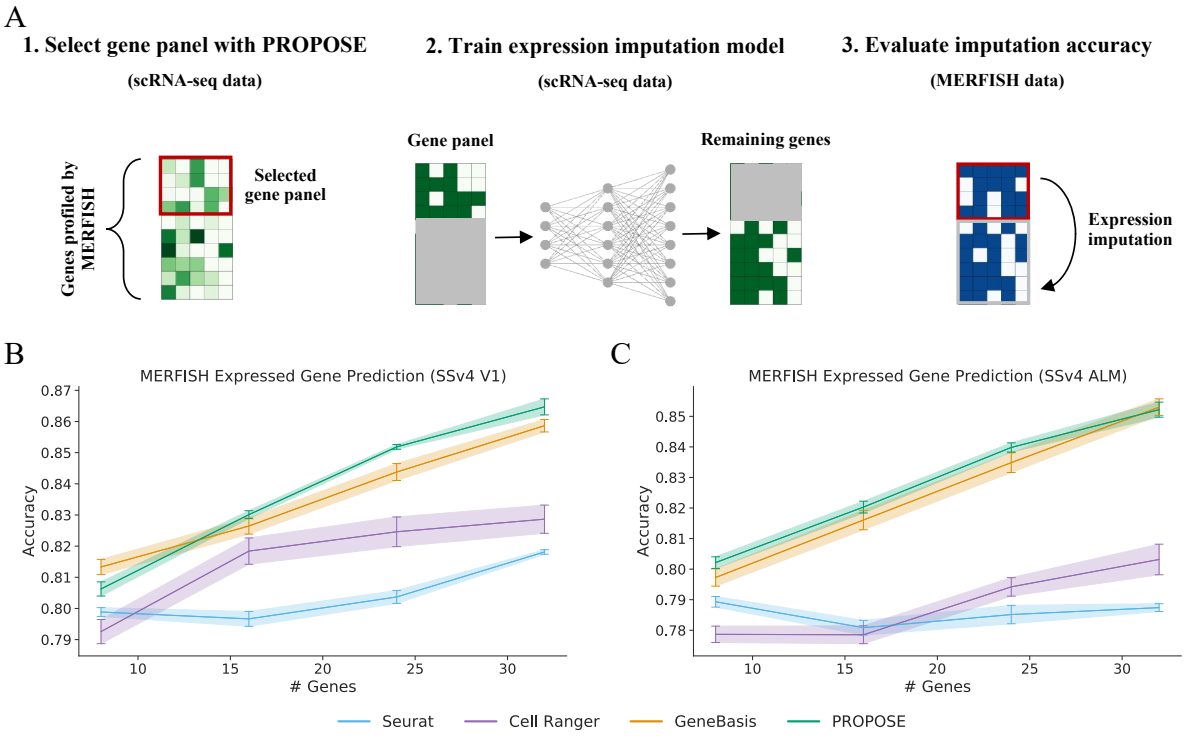

Figure 5: **Predicting MERFISH gene expression based on scRNA-seq training.** To test whether models trained on scRNA-seq can transfer to FISH despite the domain shift between technologies, we design an experiment combining the SSv4 and MERFISH datasets. We find that binarizing gene expression levels enables accurate predictions in MERFISH cells, and that PROPOSE outperforms prior gene selection approaches. **A**, In this multi-step experiment, first, a gene panel is selected using scRNA-seq restricted to the genes profiled in the MERFISH data. Next, an imputation model is trained to predict whether the remaining genes are expressed, again using scRNA-seq data. Finally, the imputation model's accuracy is tested on MERFISH data, using the small gene panel to predict the remaining genes. The imputation model's high accuracy demonstrates that binarization enables successful transfer from scRNA-seq to FISH. **B**, Expressed gene prediction accuracy for panels selected by each method when using the V1 SSv4 data. **C**, Expressed gene prediction accuracy for panels selected by each method when using the ALM SSv4 data.

mance in the expressed gene prediction task (see Fig. 2D-F). First, using the V1 and ALM SSv4 scRNA-seq datasets, we selected panels of 8-32 markers among the MERFISH gene set using PROPOSE and other methods. Then, an imputation model was trained—using only the scRNA-seq data—to predict which of the remaining genes are detected. Finally, using the resulting model, we predicted the set of detected genes in the MERFISH dataset (Fig. 5A). Importantly, to account for the domain shift, we binarized both the scRNA-seq and MERFISH gene expression levels.

The scRNA-seq-trained models can predict expressed genes in the MERFISH data with high accuracy (Fig. 5B-C). The prediction accuracy tends to improve with larger panels; for example, PROPOSE reached 86.5% with a 32-gene panel when trained on the V1 dataset. In most cases, PROPOSE achieves the highest accuracy with geneBasis being the most competitive baseline. We also find that the prediction accuracy with the V1 cells is slightly higher than with the ALM cells. PROPOSE not only outperforms other unsupervised methods that require only unlabeled scRNA-seq data (Fig. 5), but also the MutInfo and scGeneFit approaches that leverage cell type labels for gene selection (Supp. Fig. 13).

In these experiments, we carefully determined the thresholds for binarizing the MERFISH expression counts by matching the quantile of the zero-threshold used with the scRNA-seq data; this procedure enables

us to account for the zero-inflation present in scRNA-seq datasets. If we instead binarize the MERFISH data using a threshold value of zero, the prediction accuracy decreases for all gene panels (Supp. Fig. 13). Finally, because we have access to cells profiled using MERFISH with all genes in this case, we are able to determine that training the imputation model on MERFISH rather than scRNA-seq data results in an accuracy improvement of roughly 4-6% (Supp. Fig. 13). The gap is non-negligible, but it may be due, in part, to differences in the brain regions profiled in the reference dataset.

Our evaluation approach does not represent a practical workflow for a real experiment; but importantly, the fact that the model transferred successfully suggests that the binarization step can allow models for other predictive tasks to transfer from scRNA-seq to FISH data, including models for cell type classification or characterizing electrophysiological properties. This represents the first evidence, to our knowledge, that a predictive model trained using only scRNA-seq data can be applied successfully in a FISH study.

Variability in gene panel selections across algorithms

Because the gene selection algorithms tested here rely on diverse selection criteria, they produce different gene panels given the same scRNA-seq data. Here, we examine the overlap in gene panels selected by each algorithm by calculating the proportion of common genes between 32-gene panels. Expectedly, no two methods select the exact same gene set (Figs. 6A-B). The strongest similarity is between Seurat and Cell Ranger with the 10X dataset, at 53% overlap (17/32 genes); their overlap is lower with the SSv4 dataset, at just 22% (7/32 genes), and their similarity is perhaps due to the fact that both methods are based on per-gene variance levels. Another pair of similar methods is scGeneFit and geneBasis, which have an overlap of 41% and 47% with the SSv4 and 10X datasets, respectively.

We find that PROPOSE has minimal overlap with all other methods. The highest overlap portion for PROPOSE is just 19% with PROPOSE-Classification on the 10X cells, but their overlap is only 9% for the SSv4 dataset. Meanwhile, PROPOSE-Classification and MutInfo achieve moderate levels of overlap, at 31% for the SSv4 dataset and 38% for the 10X dataset; this may be because both methods aim to select genes whose expression can optimally distinguish cell types. For panels containing either 16 or 128 genes, PROPOSE's selections remain distinct from other methods, but they are most similar to those from PROPOSE-Classification and MutInfo (Supp. Fig. 14). Overall, these results reflect that the various gene selection methods select distinct gene panels, and that PROPOSE and PROPOSE-Classification's improved performance across various metrics is enabled by the selection of substantially different panels.

Finally, we examined a unique characteristic of PROPOSE: its stochasticity. Because we implemented a deep learning model that is randomly initialized and trained using stochastic gradient descent, the results from PROPOSE and its supervised variants (PROPOSE-Classification, PROPOSE-Ephys) can differ across trials. This variability is somewhat unusual for a gene selection method, but this property is shared by other state-of-the-art feature selection techniques [14] and by the UMAP embedding method [44]. To examine the variability in the individual genes selected and the performance of the selected gene panels on our evaluation metrics, we ran 32 independent trials of PROPOSE with the SSv4 dataset.

For panels containing 16 genes, four genes were selected in at least half of the trials, with a single gene, *Bmp3*, being selected in 30 of 32 trials (Fig. 6C). The remaining ones were selected less consistently; the sixteenth most frequently selected gene was chosen in just over a third of trials, and 38 genes were selected just once. Reassuringly, differences in the composition between the gene panels had little impact on cell type classification accuracy (Supp. Fig. 16). The stability in performance despite changes in the composition across gene panels is not surprising—many genes have highly correlated expression patterns and thus can

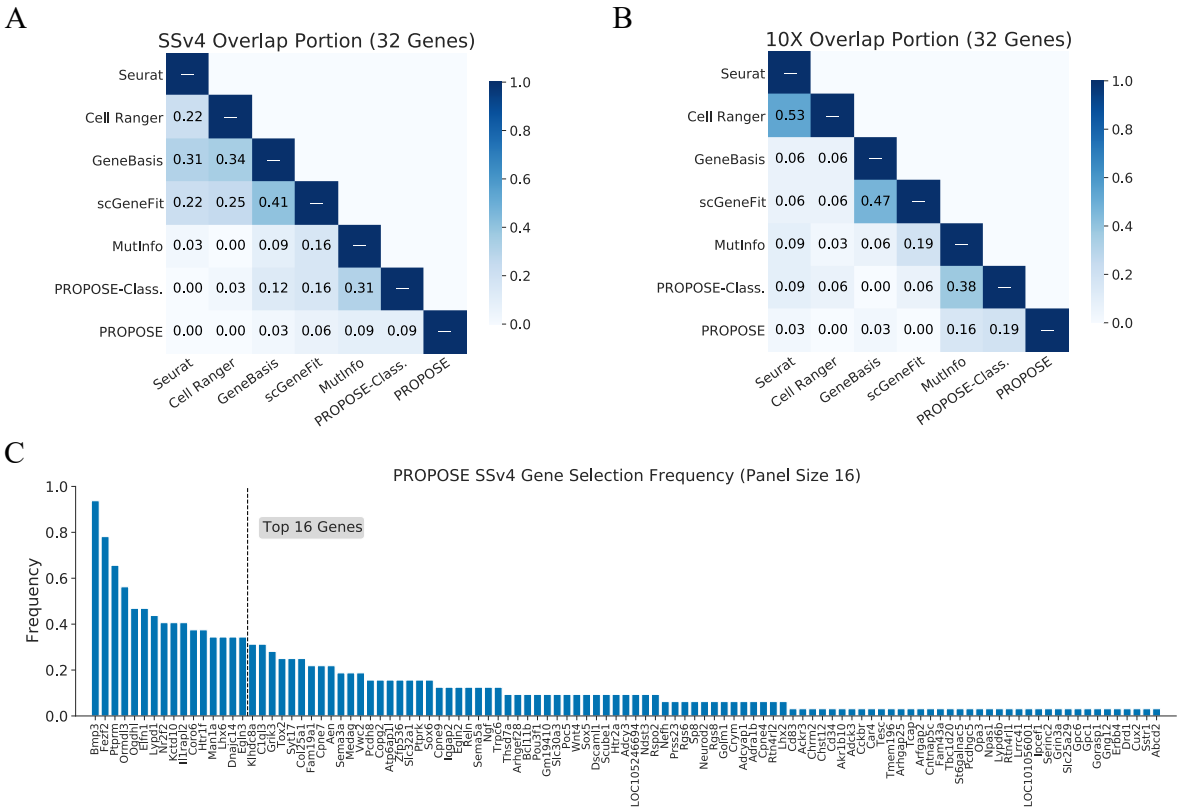

Figure 6: **Diversity in gene panels selected by each method.** Using the SSv4 and 10X datasets, we find that PROPOSE and PROPOSE-Classification select distinct gene panels from prior methods. Additionally, because PROPOSE is non-deterministic across trials, we examine the frequency of its gene selections across multiple trials. **A**, Portion of overlapping genes between panels of 32 genes for the SSv4 cells. **B**, Portion of overlapping genes between panels of 32 genes for the 10X cells. **C**, Frequency of gene selections by PROPOSE for panels containing 16 genes, aggregated across 32 independent trials.

be readily substituted. In practice, we suggest running a small number of trials and selecting the best trial from among them based on the validation loss achieved by the PROPOSE deep learning model.

Discussion

By exploiting statistical relationships in scRNA-seq expression profiles, PROPOSE identifies genes that are highly informative and that can serve as valuable targets in FISH studies. Our experiments with several datasets show that our deep learning-based approach selects more informative targets than existing algorithms, generally providing better predictive accuracy and/or enabling the use of fewer genes. A key contribution of our work is PROPOSE's robust inference ability, achieved by using binarized gene expression levels: this helps mitigate the complex relationship between scRNA-seq and FISH measurements, and we find that it allows models to transfer to FISH studies despite the challenging domain shift.

PROPOSE is by default an unsupervised method that aims to reconstruct the genome-wide expression profile; running it in an unsupervised fashion yields genes that are informative in general, but some information is sacrificed because reconstructing the full expression profile using a small number of genes remains challenging. Instead, PROPOSE can be tailored to arbitrary experimental goals by simply modifying the

prediction target, thus enabling FISH studies to use gene expression as a bridge to other data modalities. Because alternative prediction tasks are often simpler than reconstructing many thousands of genes, using PROPOSE in this fashion can enable the use of small gene panels that sacrifice minimal accuracy. Our experiments show examples with transcriptomic cell type classification and electrophysiological profile prediction, but other focused prediction tasks may include identifying disease properties or bridging with epigenetic data like chromatin accessibility and methylation. PROPOSE represents a powerful general-purpose solution for marker gene selection, and our design choices make it a uniquely effective tool for selecting small gene panels that can be used for FISH studies.

Methods

Predictive and robust probe selection (PROPOSE)

To capture as much information as possible in a small gene panel, PROPOSE aims to select genes that can predict the full scRNA-seq expression profile. Our primary tool for doing so is a deep learning model that reconstructs all the input features while relying on only a subset of the inputs, like an autoencoder but with a learned sparsity pattern. During the model's training, the selected input features are determined by a custom network layer that is trained using stochastic gradient descent.

The main part of PROPOSE is selecting a user-specified number of genes from the set of candidates. This is done using a *binary mask layer*, which is described below. However, if the number of candidate genes is large, selecting a moderate number of probes can consume a significant amount of computer memory, making PROPOSE difficult to run on a graphics processing unit (GPU). We therefore provide an optional preliminary step to eliminate a large number of candidate genes, which makes the final selection step more memory-efficient; this preliminary step is performed using the *binary gates layer*, also described below. In addition to reducing memory usage, we find that a two-stage approach yields slightly better results (Supp. Fig. 9).

Hurdle loss function

Gene dropouts represent a problematic source of noise in scRNA-seq, and because PROPOSE relies on scRNA-seq as a surrogate for FISH data, we must model gene dropouts appropriately when reconstructing the genome-wide expression profile. Recent computational tools for scRNA-seq data have shifted away from using mean squared error loss [45, 46] and towards zero-inflated models to account for this key source of noise [19, 23, 24, 47, 48]. We propose a loss function that is applicable with the continuous-valued measurements that arise from common normalization approaches, such as counts per million (CPM) normalization [49]. Alternatively, a zero-inflated negative binomial (ZINB) loss can be used if CPM normalization is not applied, but variability in the total UMI counts per cell can make integer transcript counts difficult to predict when using binarized gene expression as an input.

When reconstructing the full scRNA-seq expression profile, PROPOSE trains a model that outputs predictions for each target gene $i = 1, \ldots, d$. The predictions consist of a point prediction $\hat{y}_i$ as well as a probability $\hat{p}_i$ of the gene having non-zero expression. Given a fixed weighting parameter $\gamma > 0$, the hurdle loss $\ell_\gamma^{(i)}$ for gene i with observed expression level y_i is defined as

$$\ell_\gamma^{(i)}\big(\hat{p}_i, \hat{y}_i, y_i\big) = \begin{cases} \frac{1}{\gamma}(\hat{y}_i - y_i)^2 - \log \hat{p}_i & y_i > 0 \\ -\log(1 - \hat{p}_i) & y_i = 0. \end{cases} \tag{1}$$

The loss has a cross entropy component for predicting whether the gene is expressed or not, as well as a mean squared error component that is incorporated only when the gene is expressed ($y_i > 0$). Because we use log-normalized gene counts y_i, the loss $\ell_\gamma^{(i)}$ can be understood as the negative log-likelihood for a log-normal hurdle distribution [25, 50], where we implicitly simplify the log-normal component by assuming fixed standard deviations for each gene, and the weighting parameter $\gamma > 0$ controls the trade-off between predicting whether a gene is expressed and predicting its expression level. The total loss ℓ_γ for the full expression profile is the following, where $\hat{\mathbf{y}}, \hat{\mathbf{p}}$ and $\mathbf{y}$ represent vectors of predictions and expression levels for all genes:

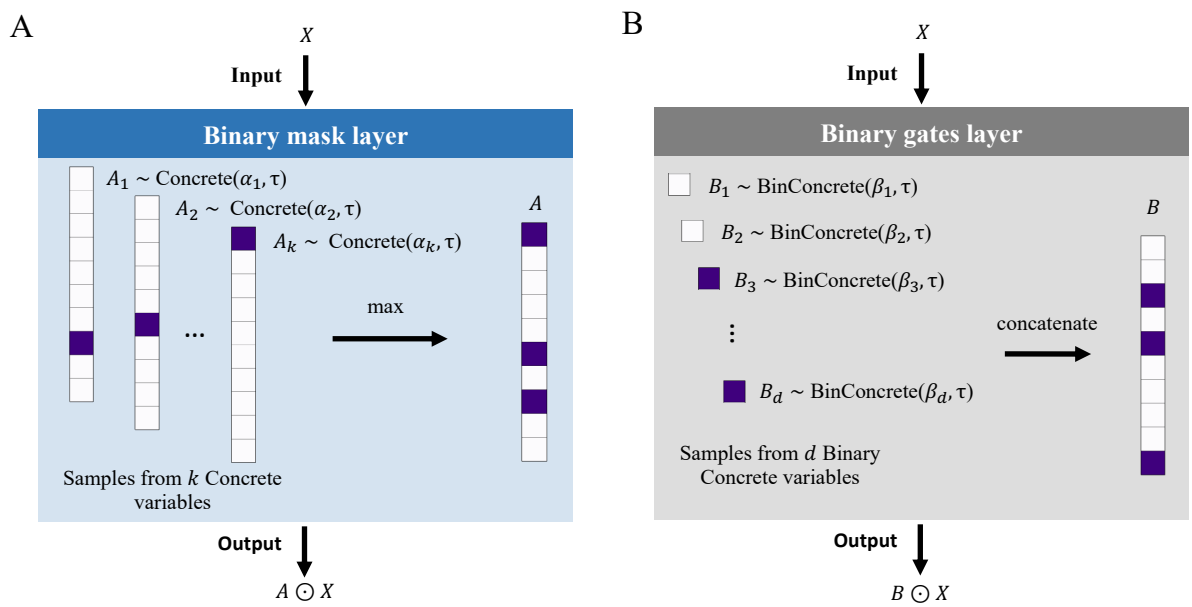

Figure 7: **Diagrams for differentiable feature selection layers. A**, The binary mask layer multiplies its input with a learned, k-hot binary mask whose entries are the element-wise maximum of samples from k Concrete random variables. **B**, The binary gates layer multiplies the input with a learned binary mask whose entries are samples from Binary Concrete random variables.

$$\ell_\gamma(\hat{\mathbf{p}}, \hat{\mathbf{y}}, \mathbf{y}) = \sum_{i=1}^{m} \ell_\gamma^{(i)}(\hat{p}_i, \hat{y}_i, y_i). \tag{2}$$

In practice, we set $\gamma = 10$ so the mean squared error and cross entropy components of the loss have similar scale.

Differentiable feature selection layers

We introduce the *binary mask layer* as a tool to select a user-specified number of inputs from a deep learning model. Our approach is based on the Concrete distribution [27] (also known as the Gumbel-Softmax [28]), which lets us optimize discrete probability distributions using stochastic gradient descent. To select exactly k genes from d candidates, we multiply the model input by a binary mask generated using the element-wise maximum of k Concrete random variables, which are denoted as $A_i \sim \text{Concrete}(\alpha_i, \tau)$ for $i = 1, \ldots, k$ (Supp. Fig. 7A). Each Concrete distribution is parameterized by unnormalized probabilities $\alpha_i \in \mathbb{R}_+^d$ and a temperature value $\tau > 0$, and each Concrete distribution converges to a multinomial distribution with probabilities given by $\frac{\alpha_i}{\mathbf{1}^\top \alpha_i}$ as $\tau \to 0$ [27].

The binary mask layer's input is a vector of binarized gene expression levels $\mathbf{x} \in \mathbb{R}^d$, and its output is given by the element-wise product $\mathbf{a} \odot \mathbf{x}$, where $\mathbf{a} = \max_i a_i \in \mathbb{R}^d$ is the element-wise max of samples a_i from each Concrete random variable $A_i \in \mathbb{R}^d$. The layer is followed by a neural network f_θ that predicts the label $\mathbf{y}$ given $\mathbf{a} \odot \mathbf{x}$, and we train the model by optimizing the following objective:

$$\min_{\theta, \alpha} \mathbb{E}_{\mathbf{x}, \mathbf{y}, \mathbf{a}} \Big[\ell_\gamma \big(f_\theta(\mathbf{a} \odot \mathbf{x}), \mathbf{y} \big) \Big]. \tag{3}$$

To select a precise number of genes, we need only ensure that the temperature parameter τ is sufficiently low at the end of training. In practice, we find that each Concrete random variable focuses its probability mass on a single input, yielding a set of exactly k selected genes. The binary mask layer is similar to a layer used in the CAE approach [14], but we find that our parameterization, which uses a simple element-wise multiplication rather than a matrix multiplication, provides slightly better results (Supp. Fig. 9).

We also introduce the *binary gates layer* as a memory-efficient tool to eliminate a large number of input features from a deep learning model. Similar to previous work [26], we learn a Binary Concrete random variable for each input feature, denoted as $B_i \sim \text{BinConcrete}(\beta_i, \tau)$, and we use these "gate" variables to perform element-wise multiplication with the corresponding genes (Supp. Fig. 7B). In the Binary Concrete distribution, each parameter $\beta_i > 0$ represents an unnormalized probability, $\tau > 0$ represents a temperature value, and each random variable B_i converges to a Bernoulli random variable with probability parameter $\frac{\beta_i}{1+\beta_i}$ as $\tau \to 0$ [27].

The binary gates layer output is given by the element-wise product $\mathbf{b} \odot \mathbf{x}$, where $\mathbf{x}$ is the input and $\mathbf{b} = [b_1, \ldots, b_d] \in \mathbb{R}^d$ is a vector of samples b_i from the random variables $B_i \sim \text{BinConcrete}(\beta_i, \tau)$ for $i = 1, \ldots, d$. The layer is followed by a neural network f_θ that predicts the label $\mathbf{y}$ given $\mathbf{b} \odot \mathbf{x}$. Genes are eliminated when we learn low β_i values, and we encourage this by augmenting the loss function with a penalty on the BinConcrete samples:

$$\min_{\theta,\beta} \mathbb{E}_{\mathbf{x},\mathbf{y},\mathbf{b}}\Big[\ell_\gamma\big(f_\theta(\mathbf{b} \odot \mathbf{x}), \mathbf{y}\big) + \lambda \mathbf{1}^\top \mathbf{b}\Big]. \tag{4}$$

The regularization term penalizes the number of selected genes, and the hyperparameter $\lambda > 0$ controls the trade-off between prediction accuracy and the number of genes used. Determining a precise number of genes requires choosing the correct λ value, and our implementation finds this value automatically by iteratively adjusting λ using a modified version of the secant algorithm [51]. Briefly, given a desired number of candidate genes $d' < d$, after trying a particular penalty parameter value, our method guesses the next value λ' based on the number of genes yielded by the previous λ values. In our experiments with the SSv4 and 10X datasets, we initially narrow the set of candidates to roughly 500 genes. For the experiments involving Patch-seq and MERFISH data, which involve fewer candidate genes, we directly select gene panels using the binary mask layer.

Training

When training our deep learning models, we optimize the models using Adam with the standard learning rate (10^{-3}) [52]. Over the course of training, the temperature parameter τ is geometrically annealed from a high value to a low value to encourage discrete feature selection, similar to previous work that uses the Concrete distribution [14]. For both input layers, we use an initial temperature $\tau = 10.0$ and a final temperature of $\tau = 0.01$.

After training, the parameters for the Concrete random variables are naturally learned such that the same genes are selected at every forward pass. That is, most Binary Concrete variables are deterministically equal to 0 or 1, and the Concrete variables are one-hot at the same entry in every sample. As a selection criterion for the binary gates layer, we retain all genes i such that $\frac{\beta_i}{1+\beta_i} > \frac{1}{2}$. For the binary mask layer, we select k genes using the maximum index of each vector of unnormalized probabilities α_i for $i = 1, \ldots, k$.

The binary mask layer is necessary for selecting a specific number of genes, but the binary gates layer is preferable for eliminating a large number of genes due to memory constraints. The binary mask layer has

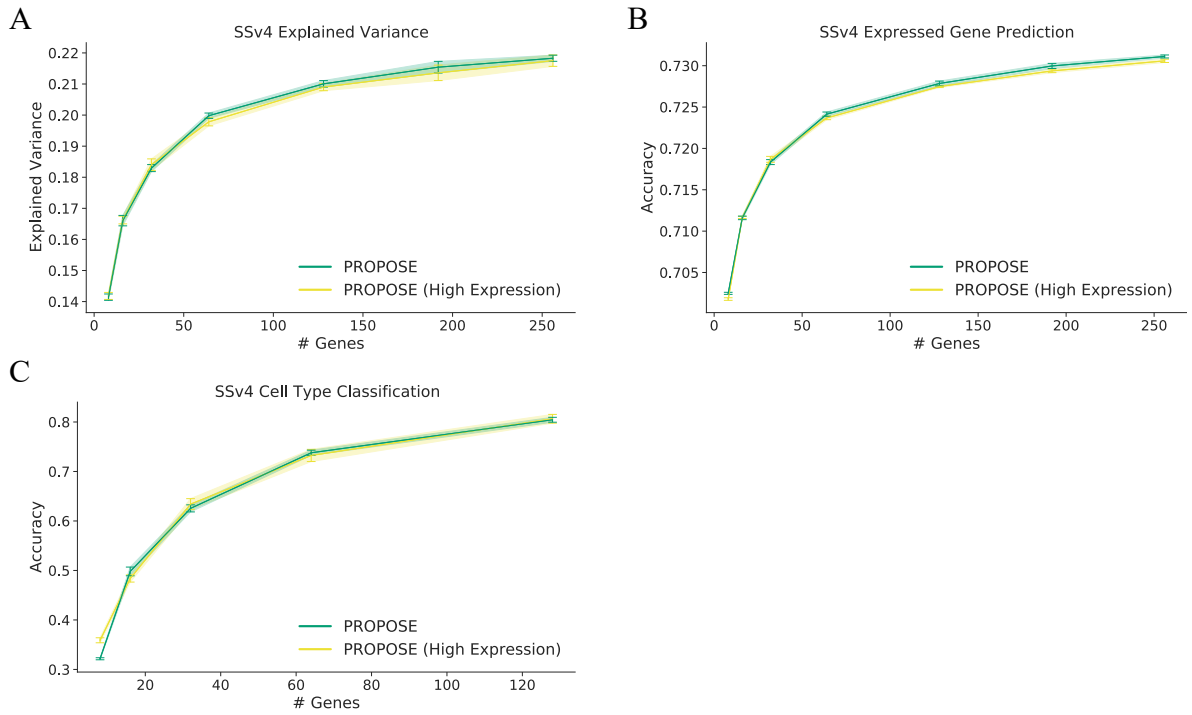

Figure 8: **Restricting PROPOSE to highly expressed genes.** PROPOSE can incorporate expert experimenter knowledge, for example by considering only genes with high expression levels (whose maximum expression level is above the median); this ensures that the transcripts are more visible during the FISH experiment. The results show that even with this restriction on the gene candidates, PROPOSE is able to identify informative gene panels and its performance is nearly identical. **A**, Explained variance for gene panels of different sizes. **B**, Expressed gene prediction accuracy. **C**, Cell type classification accuracy.

$k \times d$ learnable parameters when selecting k genes, whereas the binary gates layer has only d. The binary mask layer is therefore difficult to apply directly in scenarios where d is large and k is moderate (a large number of candidates and a relatively large gene panel), and the binary gates layer is useful for our datasets with $d = 10{,}000$ total candidate genes.

In practice, the outcome from running PROPOSE may differ across trials when the random seed is not fixed. For our experiments, we run five trials and use the gene panel that achieves the best validation loss. Our implementation is available online: `https://github.com/iancovert/propose/`.

Expert knowledge and supervision

PROPOSE can be used in a purely data-driven fashion or it can incorporate expert knowledge. For example, the candidate genes can be restricted to those with known biological function or other desirable properties, or the gene panel can be forced to include certain hand-picked genes. When such modifications are made to the gene candidates, PROPOSE finds the best available gene panel while automatically accounting for any pre-selected genes.

For example, it can be useful to consider only genes with high expression levels, because these transcripts are easier to identify in a FISH experiment. Supp. Fig. 8 shows that when PROPOSE is restricted to using only genes whose maximum expression level is higher than the median value, we are able to achieve roughly the same performance as when choosing from among all 10,000 candidate genes. In general, filtering steps

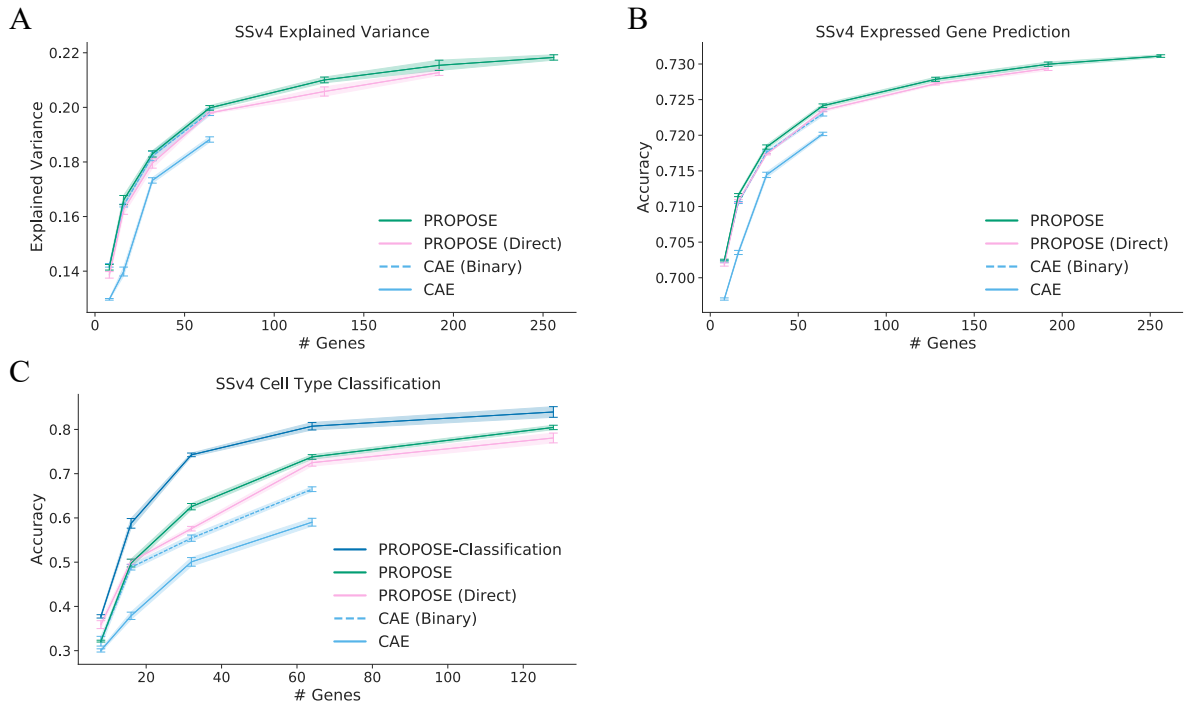

Figure 9: **PROPOSE ablations.** PROPOSE performs better than the CAE on expression profile reconstruction and cell type classification tasks, both with and without binarization. We also find that PROPOSE performs slightly better when initially narrowing the set of candidate genes, rather than proceeding directly to the selection step (PROPOSE Direct). Notably, the narrowing step enables PROPOSE to select larger panels, which is not possible with the CAE or PROPOSE (Direct) due to GPU memory constraints. **A**, Explained variance for the SSv4 dataset. **B**, Expressed gene prediction accuracy for the SSv4 dataset. **C**, Cell type classification accuracy for the SSv4 dataset.

that preserve most of the gene candidates should not prevent PROPOSE from finding highly informative panels.

Rather than using PROPOSE in an unsupervised manner, we can also incorporate supervision to align the gene panel with specific experimental aims. If the FISH study has a specific prediction objective, such as characterizing electrophysiological properties, and accompanying labels are available for the reference scRNA-seq data, PROPOSE can incorporate this information into the selection procedure. Adapting PROPOSE is straightforward, requiring only a change in the prediction target and loss function, as we demonstrate with PROPOSE-Classification (multiclass cross entropy loss, Fig. 3) and PROPOSE-Ephys (mean squared error loss, Fig. 4).

Ablation experiments

We tested several variants of PROPOSE to validate our design choices. First, we compared PROPOSE to a version that skips the initial step of narrowing the set of candidate genes. The results are slightly worse (Supp. Fig. 9), which we attribute to the beneficial effect of iteratively reducing the number of genes. Next, we compared PROPOSE to the CAE approach [14], which is closely related but differs in its choice of loss function (mean squared error rather than hurdle loss) and feature selection layer. We find that the CAE underperforms PROPOSE across our three evaluation metrics with the SSv4 data. Most significantly, the CAE underperforms at cell type classification, which is consistent with recent work that highlights the importance of properly modeling gene dropouts [23, 24]. Finally, we observe that incorporating gene

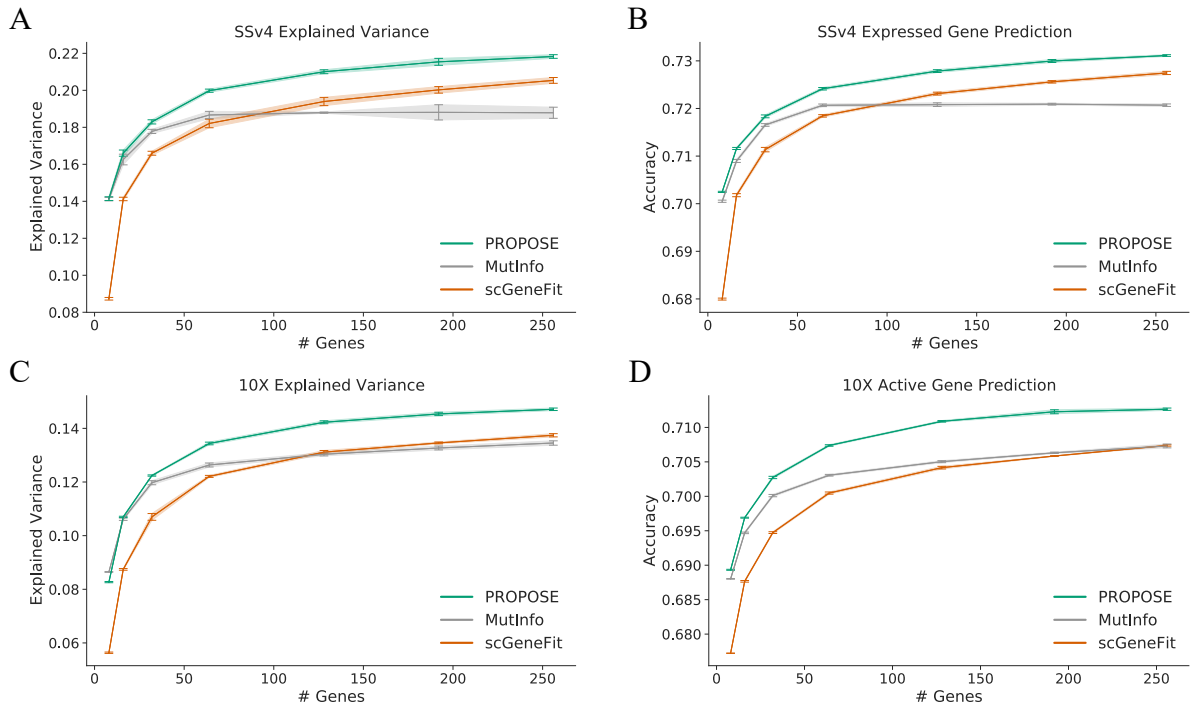

Figure 10: **Expression profile reconstruction results for supervised gene selection methods.** Here, we compare PROPOSE with the two supervised methods that leverage cell type labels, scGeneFit and MutInfo, on the tasks of predicting log-normalized expression counts (**A**, **C**) and whether each gene is expressed (**B**, **D**). **A**, Explained variance for gene panels with the SSv4 dataset. **B**, Expressed gene prediction accuracy for the SSv4 dataset. **C**, Explained variance for gene panels with the 10X dataset. **D**, Expressed gene prediction accuracy for the 10X dataset.

expression binarization improves the CAE's performance on our metrics, but that the binarized CAE still does not match PROPOSE's performance.

GPU memory usage can be a concern when selecting relatively large gene panels from a large number of candidate genes: both the binary mask layer and the selection layer used by the CAE have $k \times d$ learnable parameters when selecting k genes from among d candidates, and they produce large tensors during training. As a result, the CAE causes a memory overflow for panels containing 128 or more genes; this issue can be avoided by reducing the minibatch size during training, aggregating gradients across multiple forward passes or parallelizing training across multiple GPUs, but we reflect this obstacle in our results for illustrative purposes (Fig. 9). Similarly, PROPOSE causes a memory overflow for panels of 256 genes when we do not initially narrow the set of candidate genes. Thus, the initial narrowing step with the binary gates layer is useful not only for achieving a small performance improvement, but for helping PROPOSE scale to large panels and datasets.

Additional results

Here, we provide several additional results evaluating PROPOSE relative to the baseline selection methods. First, we present results for the expression profile reconstruction metrics using the supervised methods, scGeneFit and MutInfo, which leverage cell type labels during the selection stage. Supp. Fig. 10 shows their explained variance and expressed gene prediction accuracy for the SSv4 and 10X datasets. MutInfo is competitive for small gene panels, but neither method is able to match PROPOSE's performance.

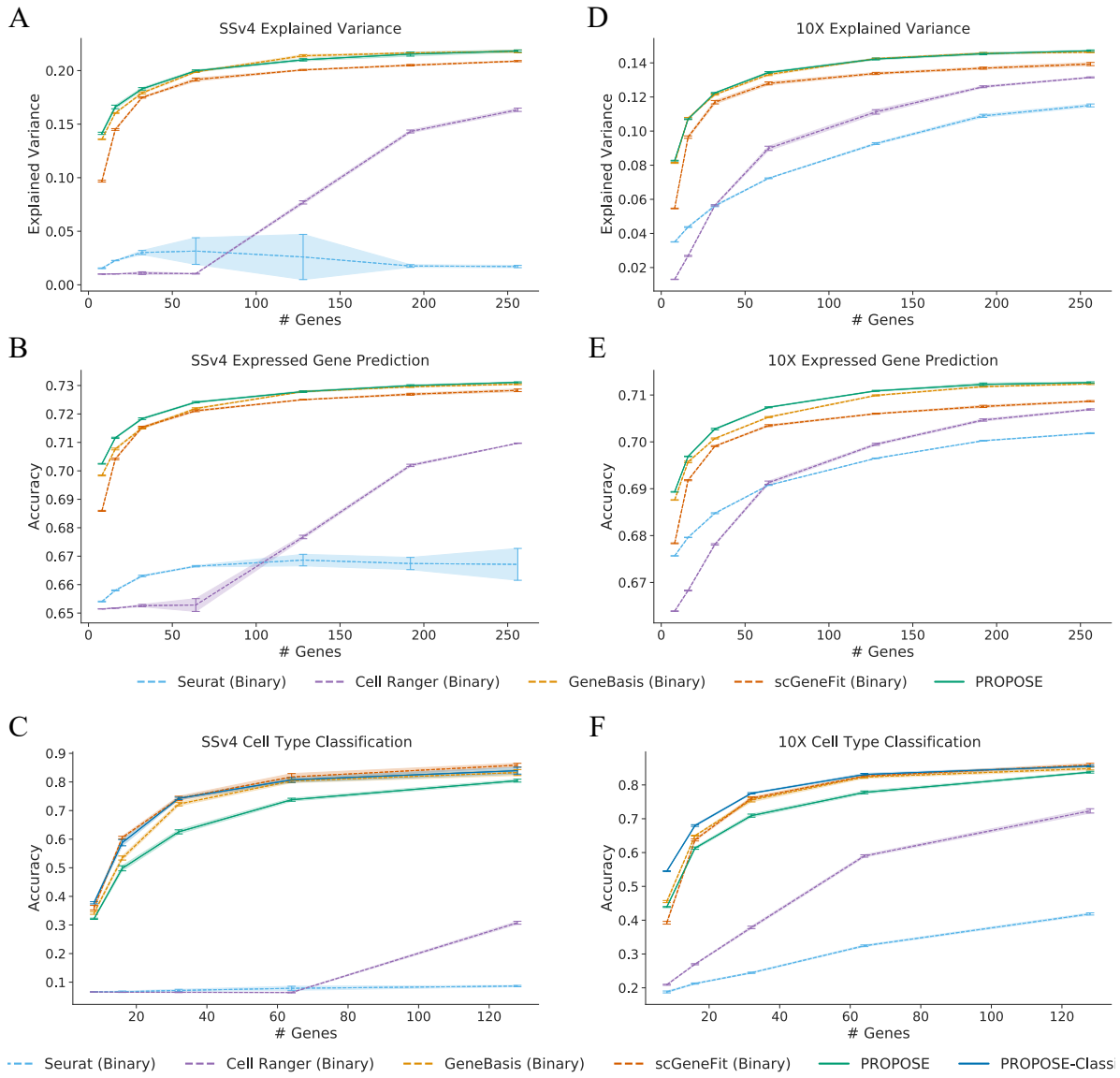

Figure 11: **Introducing binarization into existing gene selection methods.** Binarizing gene expression counts before selecting the gene panel can enable certain baseline methods to perform better in our expression profile reconstruction and cell type classification tasks. GeneBasis and scGeneFit become more competitive with PROPOSE and PROPOSE-Classification, while Seurat and Cell Ranger perform worse. **A**, Explained variance for gene panels with the SSv4 dataset. **B**, Expressed gene prediction accuracy for gene panels with the SSv4 dataset. **C**, Cell type classification accuracy for gene panels with the SSv4 dataset. **D**, Explained variance for gene panels with the 10X dataset. **E**, Expressed gene prediction accuracy for gene panels with the 10X dataset. **F**, Cell type classification accuracy for gene panels with the 10X dataset.

Next, we examine whether introducing binarization into the baseline methods enables them to achieve better performance on our evaluation metrics, which simulate the use of gene panels in a FISH study. Supp. Fig. 11 shows that binarization makes scGeneFit and geneBasis more competitive with PROPOSE, but that Seurat and Cell Ranger perform worse; the latter result is unsurprising, because variance and dispersion values are significantly less informative for binary data.

As an intuitive visualization of how PROPOSE can distinguish cell types using only binarized expression levels, we plot the frequency of each gene being expressed within each cell type. Supp. Fig. 12 displays this

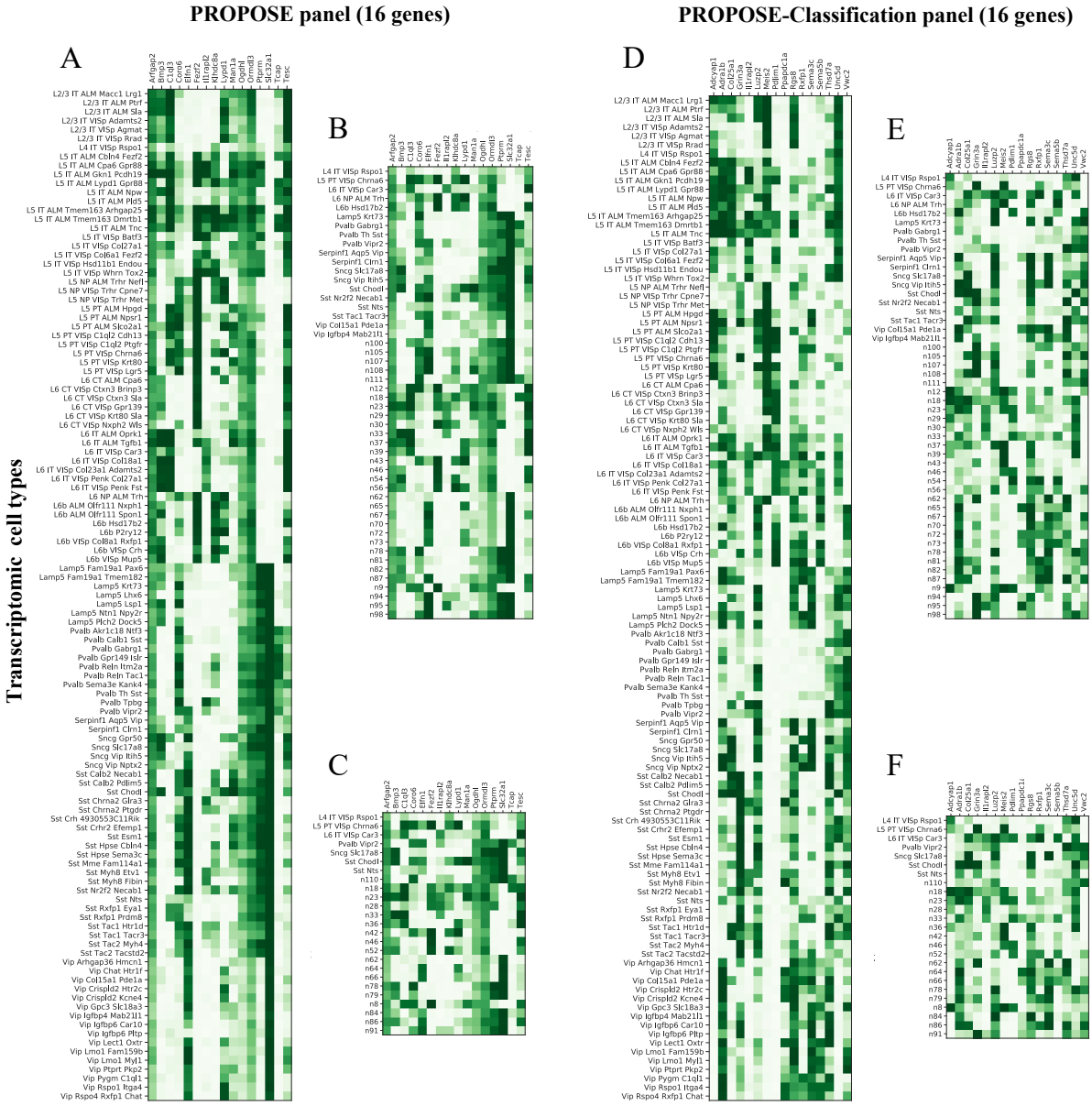

Figure 12: **PROPOSE gene panel expression frequency by cell type.** Using the SSv4 dataset, we plot the portion of the time that each gene in a panel is expressed within each cell type (white indicates never expressed, dark green indicates always expressed). Examining these matrices for the PROPOSE and PROPOSE-Classification panels containing 16 genes, using the full set of transcriptomic cell types as well as small numbers of subclasses, we can identify unique expression patterns for each cell type. **A**, Expression matrix for PROPOSE gene panel with all cell types. **B**, Expression matrix for PROPOSE gene panel with 50 subclasses. **C**, Expression matrix for PROPOSE gene panel with 25 subclasses. **D**, Expression matrix for PROPOSE-Classification gene panel with all cell types. **E**, Expression matrix for PROPOSE-Classification gene panel with 50 subclasses. **F**, Expression matrix for PROPOSE-Classification gene panel with 25 subclasses.

expression matrix for the full set of transcriptomic cell types in the SSv4 dataset, as well as for 50 and 25 subclasses, using the PROPOSE and PROPOSE-Classification panels containing 16 genes. The results show that each cell type has a unique expression pattern, enabling us to distinguish cell types using a surprisingly small number of genes. All 2^{16} expression patterns across the 16 genes are technically possible, but with

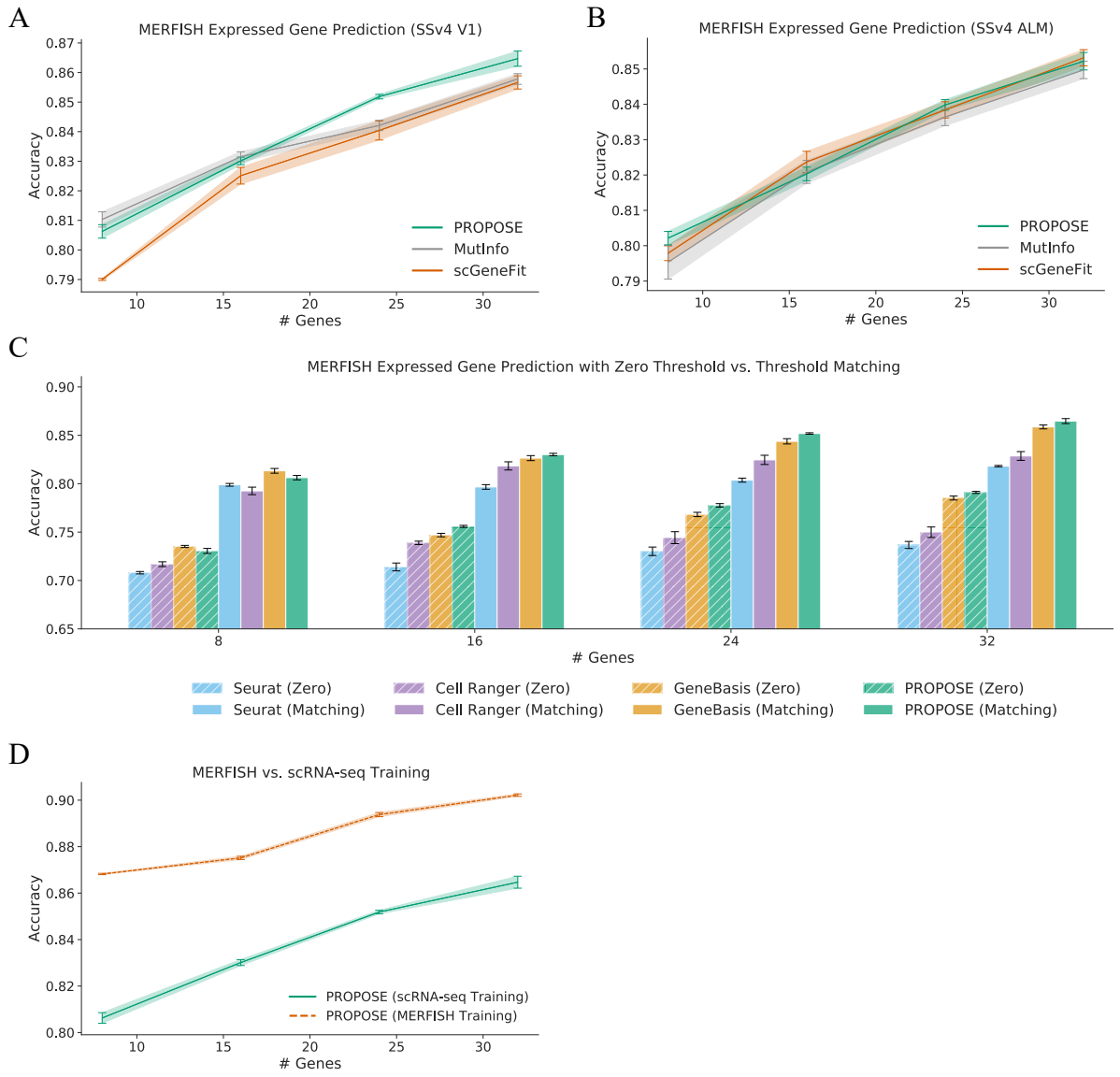

Figure 13: **Additional MERFISH expression imputation results.** Here, we report results for the supervised gene selection methods, we investigate the importance of the thresholding procedure for MERFISH expression counts, and we examine the degree of accuracy lost due to domain shift between the scRNA-seq and MERFISH datasets. **A**, Imputation accuracy for panels selected by each method when using the SSv4 V1 scRNA-seq data; PROPOSE outperforms the supervised baselines, except for MutInfo with small panel sizes. **B**, Imputation accuracy for panels selected by each method when using the SSv4 ALM scRNA-seq data; PROPOSE achieves similar performance to the supervised baselines. **C**, Imputation accuracy when using a zero threshold for MERFISH expression counts, versus when using a threshold that matches the scRNA-seq quantile; matching the threshold significantly improves the accuracy for all methods across all panel sizes. **D**, Comparing imputation accuracy when selecting genes using V1 cells but training on V1 scRNA-seq versus MOp MERFISH data; training on in-domain MERFISH data, which is free of domain shift but inaccessible in practice, reveals an accuracy loss of 4-6% due to training on scRNA-seq.

these panels, many genes are expressed nearly all of the time or none of the time within each cell type. For example, we find that with all 113 transcriptomic cell types, 67% of the entries are expressed either under 20% or over 80% of the time (Supp. Fig. 12A); this figure is even higher at 74% for PROPOSE-Classification (Supp. Fig. 12D).

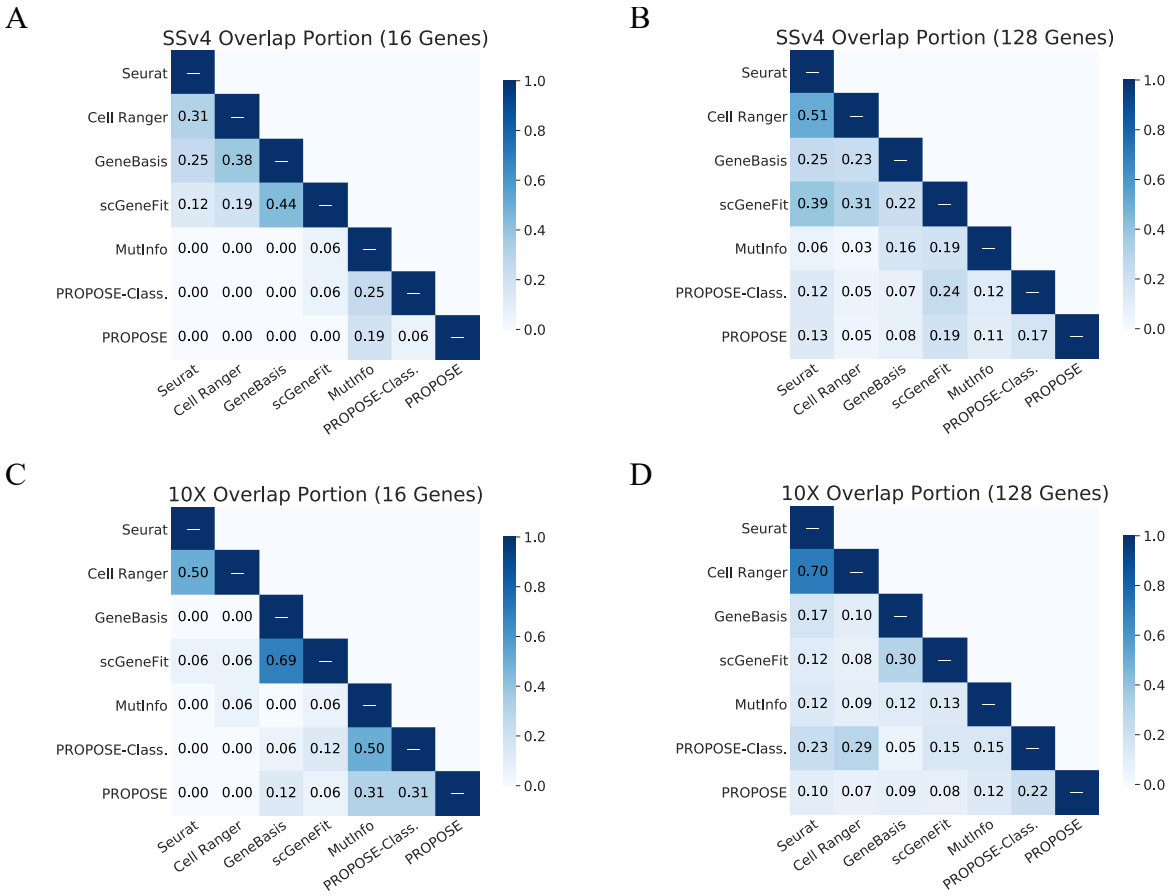

Figure 14: **Diversity in gene panels of different sizes.** Using the SSv4 and 10X dataset, we examine the overlap between panels containing 16 and 128 genes. For these panel sizes, the genes identified by PROPOSE remain distinct from those selected by existing methods. **A**, Portion of overlapping genes between panels of 16 genes for the SSv4 dataset. **B**, Portion of overlapping genes between panels of 128 genes for the SSv4 dataset. **C**, Portion of overlapping genes between panels of 16 genes for the 10X dataset. **D**, Portion of overlapping genes between panels of 128 genes for the 10X dataset.

For the MERFISH gene imputation experiments, we provide several additional results in Supp. Fig. 13. First, Supp. Figs. 13A-B show the imputation accuracy for the supervised methods, scGeneFit and MutInfo. We find that when using the SSv4 V1 scRNA-seq data, PROPOSE and MutInfo reach comparable accuracy for small gene panels, but that PROPOSE achieves higher accuracy with larger panels; when using the SSv4 ALM scRNA-seq data, the three methods offer comparable accuracy for panels of all sizes. Next, we examine the importance of using a carefully chosen threshold when binarizing the MERFISH gene expression counts. If we naively use a threshold value of zero, the imputation accuracy is significantly lower than if we choose a threshold to match the quantile that zero represents in the scRNA-seq data (Supp. Fig. 13C). Finally, we use the fact that we have a dataset of MERFISH expression levels to quantify the accuracy lost due to training on out-of-domain scRNA-seq data. We find that training with the in-domain MERFISH data yields an accuracy improvement of 4-6% (Supp. Fig. 13D), suggesting that our binarization step does not completely remove the effects of the domain shift relative to the SSv4 V1 scRNA-seq data.

Regarding the diversity in gene panels between different methods, we report the gene panel overlap for panels of size 16 and 128 (Supp. Fig. 14), in addition to the results for panels of 32 genes shown in Fig. 6. For

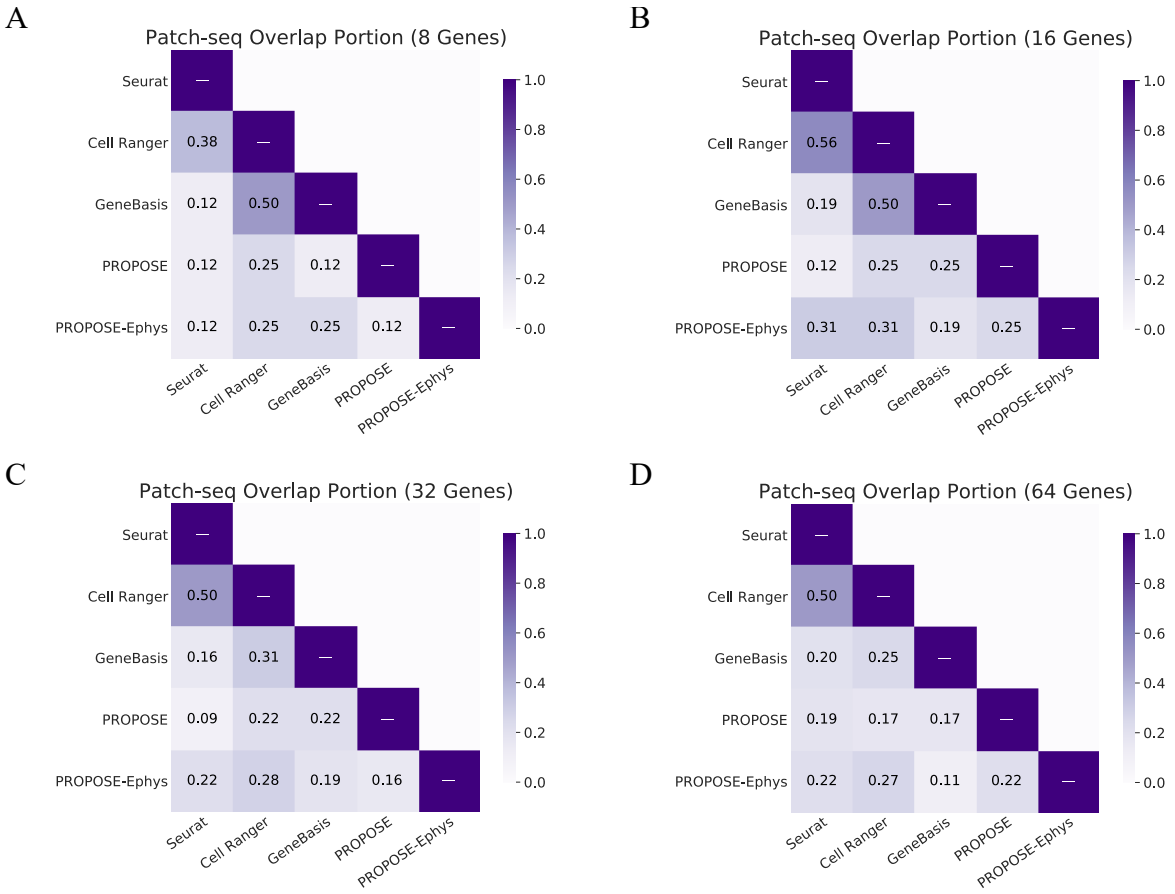

Figure 15: **Diversity in Patch-seq gene panels.** We find that for panels of all sizes, PROPOSE-Ephys selects distinct genes from the remaining methods. **A**, Portion of overlapping genes for panels of 8 genes. **B**, Portion of overlapping genes for panels of 16 genes. **C**, Portion of overlapping genes for panels of 32 genes. **D**, Portion of overlapping genes for panels of 64 genes.

panels of these sizes, we still find that PROPOSE selects genes that are largely different than those identified by other methods. We also find several pairs of similar methods that appeared in the main text results as well: scGeneFit and geneBasis, Seurat and Cell Ranger, and MutInfo and PROPOSE-Classification.

Similarly, we report gene panel overlap for the experiments with Patch-seq data (Fig. 15). The results show that PROPOSE-Ephys selects up to roughly 30% of the same genes as other methods, with Cell Ranger being the most similar for all panel sizes. PROPOSE and Cell Ranger are the next most effective methods in this experiment after PROPOSE-Ephys, but the three methods select largely distinct gene panels.

Finally, we examine to what extent the variability in PROPOSE's selections across trials impacts the performance in our evaluation metrics (Supp. Fig. 16). Using five independent trials with the SSv4 dataset, we examine the mean performance across trials, the gap between the minimum and maximum performance, and the performance from the single trial chosen according to its validation loss (calculated using the PROPOSE reconstruction model's hurdle loss). We find that the variability in performance is relatively small, particularly for the explained variance metric (Supp. Fig. 16A). Choosing a single trial based on its validation loss tends to result in performance above the mean for smaller panels ($\leq$16 genes), but with larger panels it can be slightly less effective than choosing a single trial at random. These results suggest that the significant redundancy in genome-wide expression profiles enables the selection of diverse panels with nearly

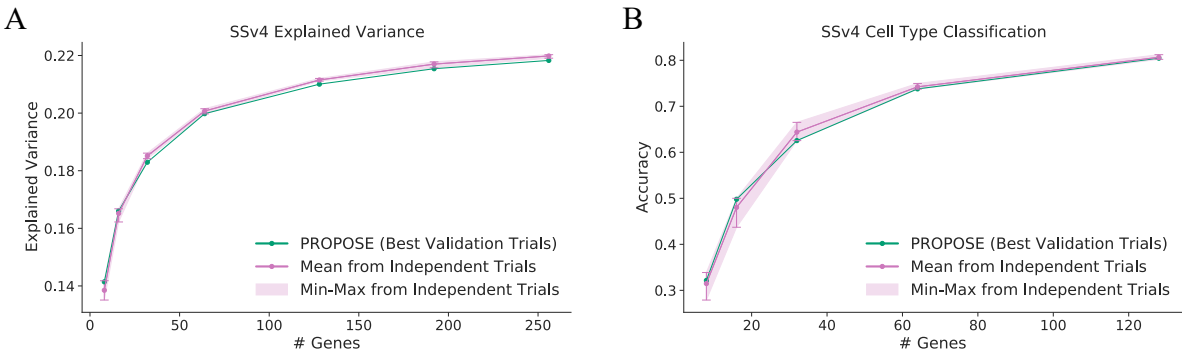

Figure 16: **PROPOSE performance variability across runs.** PROPOSE can select different panels in each trial, but the range of performance in our evaluation metrics is relatively narrow. We compare the mean performance across five trials to the minimum and maximum performance, as well as the performance from the single trial chosen according to its validation loss. **A**, Explained variance across five trials for the SSv4 dataset. **B**, Cell type classification accuracy across five trials for the SSv4 dataset.

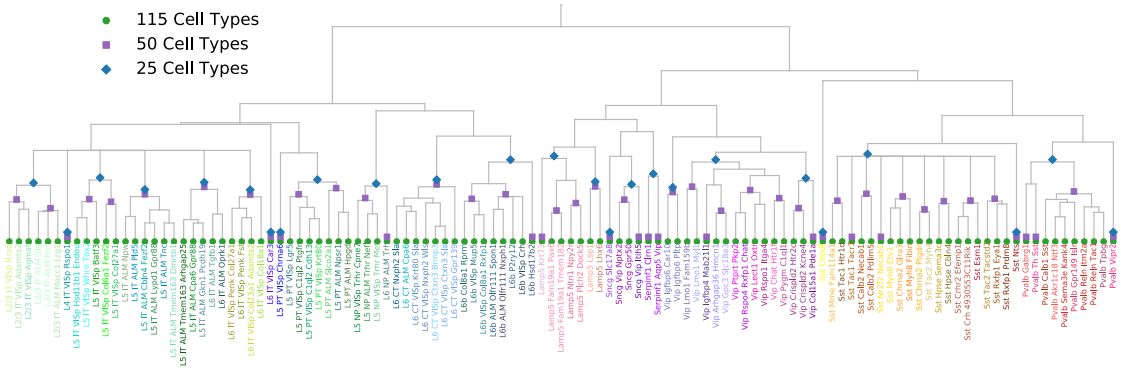

Figure 17: **Transcriptomic hierarchy for the SSv4 V1/ALM neuronal cells.** The cell types for this dataset are defined by a transcriptomic hierarchy containing 115 total cell types, represented here as leaf nodes. To simplify the cell type classification task, we consider merging cell types according to the transcriptomic hierarchy, yielding 50 or 25 more clearly defined subclasses.

equivalent information content.

Evaluation metrics

The variance explained by a gene panel is measured by training a neural network to predict the full expression profile using the selected genes, and then measuring the mean squared error on the test set. For all methods, we used CPM normalized and logarithmized expression counts as the prediction target. To give the resulting error value an intuitive scale, we calculate the explained variance as follows: we subtract the mean squared error from the total variance, and then present the difference as a portion of the total variance.

The cell type classification performance is the accuracy of a gradient-boosted decision tree (GBDT) model trained with a multi-class cross entropy loss. The cell type with the highest classification probability is taken as the predicted class. To evaluate the classification performance with cell subclasses for the SSv4 dataset, we trained models with cell types merged according to their order in the transcriptomic hierarchy (Supp. Fig. 17).

The expressed gene prediction accuracy is measured using the test set accuracy of a neural network trained with a cross entropy loss that is additive across the target genes. Our results in the main text for the

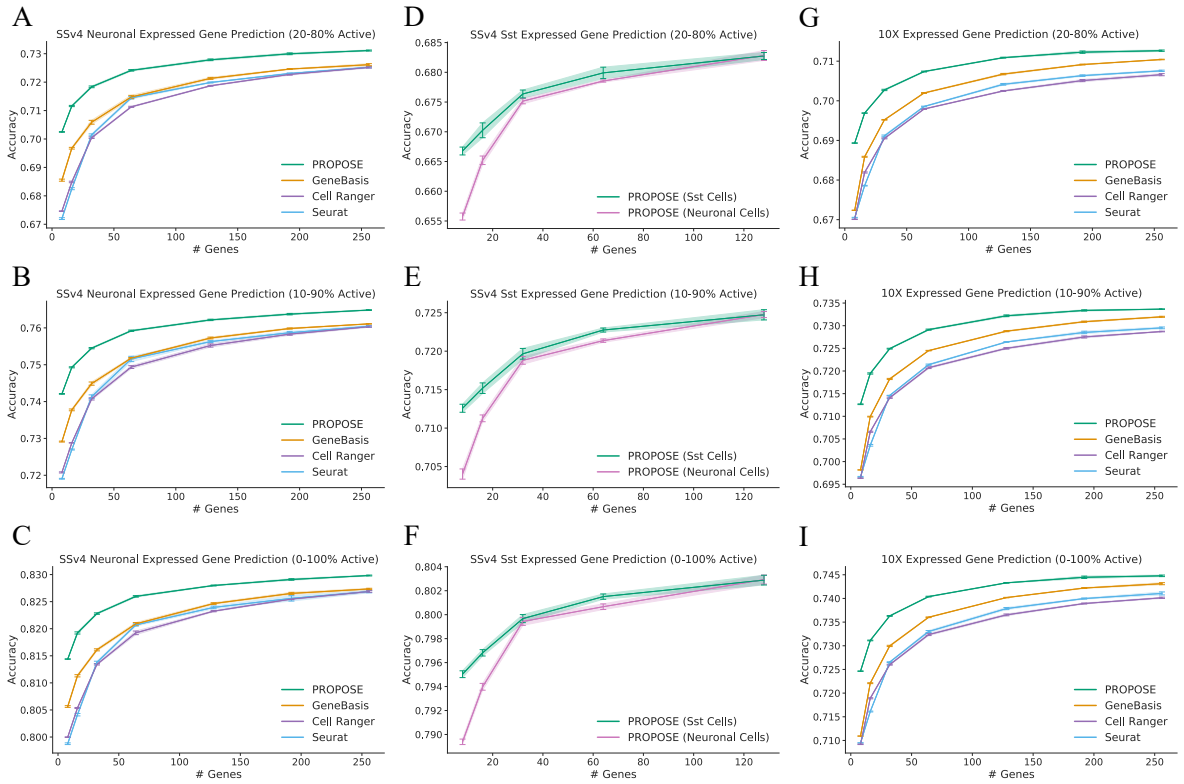

Figure 18: **Expressed gene prediction with different expression cutoffs.** Genes that are expressed in either all cells or no cells are easiest to predict, and focusing on these genes can inflate the expressed gene prediction accuracy. We examine how the accuracy changes as we apply different cutoffs to determine the more difficult-to-predict genes. **A**, SSv4 neuronal accuracy for genes expressed in 20-80% of cells. **B**, SSv4 neuronal accuracy for genes expressed in 10-90% of cells. **C**, SSv4 neuronal accuracy for genes expressed in 0-100% of cells (all genes). **D**, SSv4 Sst accuracy for genes expressed in 20-80% of cells. **E**, SSv4 Sst accuracy for genes expressed in 10-90% of cells. **F**, SSv4 Sst accuracy for genes expressed in 0-100% of cells. **G**, 10X accuracy for genes expressed in 20-80% of cells. **H**, 10X accuracy for genes expressed in 10-90% of cells. **I**, 10X accuracy for genes expressed in 0-100% of cells (all genes).

SSv4 and 10X datasets focus on genes that are expressed in 20-80% of cells, ignoring the remaining genes whose expression is easiest to predict. The results with different ranges of 10-90% and 0-100% are shown in Supp. Fig. 18. The trend between methods does not depend on the cutoff, but including the easier genes inflates the overall accuracy while shrinking the gap between methods.

Models and hyperparameters

Each dataset is split into training, validation and test sets, and the test set was not used except for when evaluating model performance. Our hyperparameter choices, including early stopping for the downstream tasks, were all made using the validation set. Our model and hyperparameter choices are summarized in Table 1: we used either multi-layer perceptron (MLP) or LightGBM models in each case [53]. Confidence intervals for the downstream tasks were determined by training models with five bootstrapped training sets and measuring the test set performance across these models.

Table 1: Model and hyperparameter choices.

Model	Description
PROPOSE (SSv4, 10X, Patch-seq)	MLP (4 × 128 units, ReLU activations), mbsize 128, 500 epochs
PROPOSE (SSv4 Sst)	MLP (2 × 128 units, ReLU activations), mbsize 1024, 4000 epochs
PROPOSE (SSv4 V1 → MERFISH, SSv4 ALM → MERFISH)	MLP (2 × 128 units, ReLU activations), mbsize 128, 250 epochs
PROPOSE-Classification (SSv4, 10X)	MLP (2 × 128 units, ReLU activations), mbsize 1024, 500 epochs
Explained variance (SSv4, SSv4 Sst, 10X)	MLP (2 × 128 units, ReLU activations), mbsize 512, 100 epochs (with early stopping)
Expressed gene prediction (SSv4, SSv4 Sst, 10X)	MLP (2 × 128 units, ReLU activations), mbsize 256, 100 epochs (with early stopping)
Expressed gene prediction (SSv4 V1 → MERFISH, SSv4 ALM → MERFISH)	MLP (2 × 128 units, ReLU activations), mbsize 256, 100 epochs (with early stopping)
Cell type classification (SSv4, 10X)	LightGBM, learning rate 0.05, max 10,000 boosters (with early stopping)
Electrophysiological properties (Patch-seq)	MLP (2 × 128 units, ReLU activations), mbsize 256, 100 epochs (with early stopping)

Data pre-processing

Due to the technical noise in scRNA-seq data, we applied CPM normalization [49] to the raw measurements and then applied the log1p operation. These values serve as prediction targets for PROPOSE, whereas the inputs to our various models are binarized expression counts. For both the V1/ALM SmartSeq v4 and M1 10X datasets, we restricted our analysis to the 10,000 transcripts with the highest variance. We used exon counts for the high-resolution SmartSeq v4 dataset and the sum of intron and exon counts for the lower resolution 10X data. For the MOp MERFISH dataset, the only pre-processing we applied was binarization to enable our models trained on scRNA-seq to generalize. When working with the MERFISH data, we analyze 253 genes that appear in both the MOp and SSv4 datasets rather than the 258 genes described in the original paper [6].

Expression quantization

We quantized gene expression levels in order to train models using scRNA-seq data that can transfer to FISH data. There is a complex domain shift between these two technologies, but the quantized gene expression values should be similar if we assume that the transformation is monotonic. When applying models trained with scRNA-seq on FISH data in practice, we recommend using the threshold matching approach, i.e., finding the quantile that the scRNA-seq threshold represents in the scRNA-seq measurements (we use a threshold value of zero) and then identifying a matching threshold in the FISH data.

Baseline methods

For the Seurat and Cell Ranger gene selection protocols, we used the implementations available in the Scanpy [31] package: `https://github.com/theislab/scanpy`. Seurat was run with raw gene expres-

Table 2: Summary of datasets.

Species	Region	Technology	Num. Genes	Num Cells.	Annotations Used
Mouse	Primary visual (V1), anterior lateral motor (ALM)	SmartSeq v4	45,768 (10,000 in experiments)	22,160 neuronal, 2,701 Sst	Transcriptomic cell types (115 total)
Human	Motor cortex (M1)	10X	50,281 (10,000 in experiments)	72,629	Transcriptomic cell types (117 total)
Mouse	Primary visual (V1)	Patch-seq	1,252	3,411	Electrophysiological profiles (68 features)
Mouse	Primary motor (MOp)	MERFISH	258 (253 in experiments)	280,327	None

sion counts and Cell Ranger with logarithmized counts. For the scGeneFit method, we used the authors' implementation with logarithmized expression counts, and we used the default hyperparameters: `https://github.com/solevillar/scGeneFit-python`. For the geneBasis method, we used the authors' R implementation: `https://github.com/MarioniLab/geneBasisR`. Finally, for the MutInfo method, we implemented the greedy forward selection algorithm described in prior work [34, 54] using the hyperparameter value $\beta = 1$ to account for gene correlations.

Datasets

Our datasets are summarized in Table 2, including the species, brain region, and relevant annotations for our experiments. The V1/ALM SmartSeq mouse neocortex data is available at `https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-v1-and-alm-smart-seq`. The M1 10X data is available at `https://portal.brain-map.org/atlases-and-data/rnaseq/human-m1-10x`. The Patch-seq data is available at `https://github.com/AllenInstitute/coupledAE-patchseq`. The MOp MERFISH data is available at `https://download.brainimagelibrary.org/02/26/02265ddb0dae51de/`.

References

[1] Andrea M Femino, Fredric S Fay, Kevin Fogarty, and Robert H Singer. Visualization of single RNA transcripts in situ. *Science*, 280(5363):585–590, 1998.

[2] Simone Codeluppi, Lars E Borm, Amit Zeisel, Gioele La Manno, Josina A van Lunteren, Camilla I Svensson, and Sten Linnarsson. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nature Methods*, 15(11):932–935, 2018.

[3] Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233), 2015.

[4] Xiao Wang, William E Allen, Matthew A Wright, Emily L Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400), 2018.

[5] Sheel Shah, Eric Lubeck, Wen Zhou, and Long Cai. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron*, 92(2):342–357, 2016.

[6] Meng Zhang, Stephen W Eichhorn, Brian Zingg, Zizhen Yao, Kaelan Cotter, Hongkui Zeng, Hongwei Dong, and Xiaowei Zhuang. Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. *Nature*, 598(7879):137–143, 2021.

[7] Jennie L Close, Brian R Long, and Hongkui Zeng. Spatially resolved transcriptomics in neuroscience. *Nature Methods*, 18(1):23–25, 2021.

[8] Hanqing Liu, Jingtian Zhou, Wei Tian, Chongyuan Luo, Anna Bartlett, Andrew Aldridge, Jacinta Lucero, Julia K Osteen, Joseph R Nery, Huaming Chen, et al. DNA methylation atlas of the mouse brain at single-cell resolution. *Nature*, 598(7879):120–128, 2021.

[9] Nathan W Gouwens, Staci A Sorensen, Fahimeh Baftizadeh, Agata Budzillo, Brian R Lee, Tim Jarsky, Lauren Alfiler, Katherine Baker, Eliza Barkan, Kyla Berry, et al. Integrated morphoelectric and transcriptomic classification of cortical GABAergic cells. *Cell*, 183(4):935–953, 2020.

[10] Federico Scala, Dmitry Kobak, Matteo Bernabucci, Yves Bernaerts, Cathryn René Cadwell, Jesus Ramon Castro, Leonard Hartmanis, Xiaolong Jiang, Sophie Laturnus, Elanine Miranda, et al. Phenotypic variation of transcriptomic cell types in mouse motor cortex. *Nature*, 598(7879):144–150, 2021.

[11] Zizhen Yao, Hanqing Liu, Fangming Xie, Stephan Fischer, Ricky S Adkins, Andrew I Aldridge, Seth A Ament, Anna Bartlett, M Margarita Behrens, Koen Van den Berge, et al. A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature*, 598(7879):103–110, 2021.

[12] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

[13] Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.

[14] Abubakar Abid, Muhammad Fatih Balin, and James Zou. Concrete autoencoders for differentiable feature selection and reconstruction. *arXiv preprint arXiv:1901.09346*, 2019.

[15] Brian Aevermann, Yun Zhang, Mark Novotny, Trygve Bakken, Jeremy Miller, Rebecca Hodge, Boudewijn Lelieveldt, Ed Lein, and Richard H Scheuermann. NS-Forest: A machine learning method for the objective identification of minimum marker gene combinations for cell type determination from single cell RNA sequencing. *bioRxiv*, 2020.

[16] Sheel Shah, Yodai Takei, Wen Zhou, Eric Lubeck, Jina Yun, Chee-Huat Linus Eng, Noushin Koulena, Christopher Cronin, Christoph Karp, Eric J Liaw, et al. Dynamics and spatial genomics of the nascent transcriptome by intron seqFISH. *Cell*, 174(2):363–376, 2018.

[17] Romain Lopez, Achille Nazaret, Maxime Langevin, Jules Samaran, Jeffrey Regier, Michael I Jordan, and Nir Yosef. A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. *arXiv preprint arXiv:1905.02269*, 2019.

[18] Jonathan Liu, Vanessa Tran, Venkata Naga Pranathi Vemuri, Ashley Byrne, Michael Borja, Snigdha Agarwal, Ruofan Wang, Kyle Awayan, Abhishek Murti, Aris Taychameekiatchai, et al. Comparative analysis of MERFISH spatial transcriptomics with bulk and single-cell RNA sequencing. *bioRxiv*, 2022.

[19] Oscar Clivio, Romain Lopez, Jeffrey Regier, Adam Gayoso, Michael I Jordan, and Nir Yosef. Detecting zero-inflated genes in single-cell transcriptomics data. *bioRxiv*, page 794875, 2019.

[20] Peng Qiu. Embracing the dropouts in single-cell RNA-seq analysis. *Nature Communications*, 11(1):1–9, 2020.

[21] Abhishek Sarkar and Matthew Stephens. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nature Genetics*, 53(6):770–777, 2021.

[22] Ian T Jolliffe. Principal components in regression analysis. In *Principal component analysis*, pages 129–155. Springer, 1986.

[23] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018.

[24] Gökcen Eraslan, Lukas M Simon, Maria Mircea, Nikola S Mueller, and Fabian J Theis. Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications*, 10(1):1–14, 2019.

[25] Andrew McDavid. *Statistical Hurdle Models for Single Cell Gene Expression: Differential Expression and Graphical Modeling*. PhD thesis, University of Washington, 2016.

[26] Chun-Hao Chang, Ladislav Rampasek, and Anna Goldenberg. Dropout feature ranking for deep learning models. *arXiv preprint arXiv:1712.08645*, 2017.

[27] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The Concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

[28] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-Softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[29] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

[30] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):1–12, 2017.

[31] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):1–5, 2018.

[32] Alsu Missarova, Jaison Jain, Andrew Butler, Shila Ghazanfar, Tim Stuart, Maigan Brusko, Clive Wasserfall, Harry Nick, Todd Brusko, Mark Atkinson, et al. GeneBasis: an iterative approach for unsupervised selection of targeted gene panels from scrna-seq. *Genome Biology*, 22(1):1–22, 2021.

[33] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[34] Xiang Li, Wei Li, Xiaodong Xu, and Wei Hu. Cell classification using convolutional neural networks in medical hyperspectral imagery. In *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, pages 501–504. IEEE, 2017.

[35] Bianca Dumitrascu, Soledad Villar, Dustin G Mixon, and Barbara E Engelhardt. Optimal marker gene selection for cell type discrimination in single cell analyses. *Nature Communications*, 12(1):1–8, 2021.

[36] Daniel Ramsköld, Shujun Luo, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid R Faridani, Gregory A Daniels, Irina Khrebtukova, Jeanne F Loring, Louise C Laurent, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology*, 30(8):777, 2012.

[37] Bosiljka Tasic, Zizhen Yao, Lucas T Graybuck, Kimberly A Smith, Thuc Nghi Nguyen, Darren Bertagnolli, Jeff Goldy, Emma Garren, Michael N Economo, Sarada Viswanathan, et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, 563(7729):72–78, 2018.

[38] Trygve E Bakken, Nikolas L Jorstad, Qiwen Hu, Blue B Lake, Wei Tian, Brian E Kalmbach, Megan Crow, Rebecca D Hodge, Fenna M Krienen, Staci A Sorensen, et al. Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature*, 598(7879):111–119, 2021.

[39] Ruoxin Li and Gerald Quon. scBFA: modeling detection patterns to mitigate technical noise in large-scale single-cell genomics data. *Genome Biology*, 20(1):1–20, 2019.

[40] Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, 568(7751):235–239, 2019.

[41] Geoffrey Stanley, Ozgun Gokce, Robert C Malenka, Thomas C Südhof, and Stephen R Quake. Continuous and discrete neuron types of the adult murine striatum. *Neuron*, 105(4):688–699, 2020.

[42] Cole Trapnell. Defining cell types and states with single-cell genomics. *Genome Research*, 25(10):1491–1498, 2015.

[43] Rohan Gala, Agata Budzillo, Fahimeh Baftizadeh, Jeremy Miller, Nathan Gouwens, Anton Arkhipov, Gabe Murphy, Bosiljka Tasic, Hongkui Zeng, Michael Hawrylycz, et al. Consistent cross-modal identification of cortical neurons with coupled autoencoders. *Nature Computational Science*, 1(2):120–127, 2021.

[44] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29), 2018.

[45] Matthew Amodio, David Van Dijk, Krishnan Srinivasan, William S Chen, Hussein Mohsen, Kevin R Moon, Allison Campbell, Yujiao Zhao, Xiaomei Wang, Manjunatha Venkataswamy, et al. Exploring single-cell data with deep multitasking neural networks. *Nature Methods*, pages 1–7, 2019.

[46] Cédric Arisdakessian, Olivier Poirion, Breck Yunits, Xun Zhu, and Lana X Garmire. DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. *Genome Biology*, 20(1):1–14, 2019.

[47] Emma Pierson and Christopher Yau. ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, 16(1):241, 2015.

[48] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications*, 9(1):1–17, 2018.

[49] Malte D Luecken and Fabian J Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, 2019.

[50] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16(1):278, 2015.

[51] James F Epperson. *An introduction to numerical methods and analysis*. John Wiley & Sons, 2021.

[52] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[53] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30:3146–3154, 2017.

[54] Roberto Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4):537–550, 1994.

Acknowledgements

We thank members of the Lee Lab and Jennie Close for helpful feedback on the manuscript. This work was funded by NSF DBI-1552309 and DBI-1759487 (I.C. and S.-I.L.), NIH R35-GM-128638 and R01-NIA-AG-061132 (I.C. and S.-I.L.), the Allen Institute (R.G., K.S. and U.S.), the Howard Hughes Medical Institute (T.W. and K.S.) and the Helen Hay Whitney Foundation (T.W.). R.G., K.S. and U.S. wish to thank the Allen Institute for Brain Science founder, Paul G. Allen, for his vision, encouragement and support.

Competing interests

The authors declare no competing interests.