Consistent cross-modal identification of cortical neurons with coupled autoencoders

Rohan Gala[1], Agata Budzillo[1], Fahimeh Baftizadeh[1], Jeremy Miller[1], Nathan Gouwens[1], Anton Arkhipov[1], Bosiljka Tasic[1], Gabe Murphy[1], Hongkui Zeng[1], Michael Hawrylycz[1], and Uygar Sümbül[1]

[1]Allen Institute, 615 Westlake Ave N, Seattle WA, USA

Abstract

Consistent identification of neurons and neuronal cell types across different observation modalities is an important problem in neuroscience. Here, we present an optimization framework to learn coordinated representations of multimodal data, and apply it to a large Patch-seq dataset of mouse cortical interneurons. Our approach reveals strong alignment between transcriptomic and electrophysiological profiles of neurons, enables accurate cross-modal data prediction, and identifies cell types that are consistent across modalities.

Keywords

neuronal cell type, Patch-seq, multimodal, cross-modal, coupled autoencoder

Highlights

Coupled autoencoders for multimodal assignment, Analysis of Patch-seq data consisting of more than 3000 cells

The characterization of cell types in the brain is an ongoing challenge in contemporary neuroscience. Describing and analyzing neuronal circuits using cell types can help simplify their complexity and unravel their role in healthy and pathological brain function.[1–6] However, the effectiveness of such approaches rests on the existence of cellular identities that manifest consistently across different observation modalities, and our ability to identify them. Recent single cell RNA sequencing (scRNA-seq) experiments have provided a detailed window into the transcriptomic organization of cortical cells in the mouse brain.[7,8] Technological developments have enabled collection of large Patch-seq datasets that include electrophysiological and transcriptomic properties for the same set of neurons.[9,10] The problem of aligning multimodal data for cell type research is challenging due to complexity of biological relationships between modalities, difficulties in measuring signal and quantifying noise in each modality, and the high dimensional nature of these datasets. Recent works to align single cell -omic measurements have largely focused on removing experimental batch effects, or on estimating correspondences between individual samples across unpaired modalities.[11,12] For Patchseq-data, there are neither overlapping features nor known associations across the modalities. However the same samples are measured in each modality, and our goal is to formulate consistent cell identities. We present a new deep neural network based methodology referred to as *coupled autoencoders* that addresses the issue of data alignment, and demonstrate its utility for the multimodal cell type identification problem using a Patch-seq dataset with transcriptomic and electrophysiological profiles of 3,411 mouse cortical interneurons.[9]

Coupled autoencoders consist of multiple autoencoder networks, each of which consists of encoder and decoder subnetworks. These subnetworks are nonlinear transformations that project input data into a low dimensional representation, and back to the input data space respectively, Figure 1a. In learning these transformations, the goal is to simultaneously maximize reconstruction accuracy for each data modality as well as similarity across representations for the different modalities. In particular, hyper-parameter λ controls the relative importance of achieving accurate reconstructions versus learning representations that are similar across modalities.

We find that low-dimensional representations of transcriptomics and electrophysiological measurements can be aligned to a high degree, while capturing salient characteristics of neurons in the individual data modalities. This strongly supports the hypothesis that molecular and electrophysiological properties of individual neurons are closely related, reflecting attributes of a common cell type, albeit through a complicated mapping. Importantly, although linear transformations[13,14] can align the major cell classes, a more detailed alignment of features and cell types is revealed only through non-linear transformations that avoid pathological

representations.

Using the aligned representations, we show that unsupervised clustering can identify ∼33 classes of GABAergic interneurons in the mouse visual cortex that are consistent across transcriptomic and electrophysiological characterizations of this neuron population. Additionally, these classes are in agreement with a reference transcriptomic taxonomy of cortical cell types.[7] Our method is general and can be extended to accommodate additional modalities of interest such as morphology and connectivity, as the datasets mature. We further demonstrate how coupled autoencoders trained on a reference dataset such as the one in this study can serve as a dictionary for smaller, single modality datasets to accurately identify cell types as well as predict data for unobserved modalities.

Aligned 3-d representations z_t and z_e for the transcriptomic and electrophysiological profiles for the high-dimensional observation vectors X_t and X_e obtained with coupled autoencoders are shown in Figure 1b-c. Cells labeled according to the reference taxonomy (see Figure S1) cluster together in representations of both observation modalities. Moreover, the representations largely preserve hierarchical relationships between cell types of the reference taxonomy. For example, in Figure 1b-c various cell types of the Sst class appear close together, while remaining well-separated from cell types of other classes such as Pvalb, Vip, and Lamp5.

Representations obtained with coupled autoencoders may be used to perform a variety of downstream analyses on complex datasets. We considered supervised classification accuracy in predicting cell type labels at different resolutions (Methods) of the reference taxonomy from z_t and z_e in Figures 1d-e, and data reconstruction performance in Figure 1f. First, we orient the reader with results for the uncoupled setting ($\lambda_\mathrm{te} = 0.0$) at each of these tasks. In Figure 1d, we note that the representations based on the transcriptomic data alone are best suited for supervised cell type classification using QDA, leading to >70% accuracy for leaf node cell type labels. This is not surprising, since the reference taxonomy was derived from analyses of gene expression alone. Electrophysiological profiles are expected to be noisy, and of lower resolution compared to transcriptomic profiles.[15] Nevertheless in Figure 1e, classifiers based on representations of electrophysiology alone predict leaf node cell type labels with ∼30% accuracy (chance level is ∼3%). Lastly, the within-modality reconstruction accuracy of uncoupled representations in Figure 1f provides an upper limit for both, within- and cross-modal reconstructions that may be achieved with 3-d representations obtained with coupled autoencoders.

To evaluate whether complicated, non-linear transformations underlie the relationship between the transcriptomic and electrophysiological features of neurons, we considered the performance

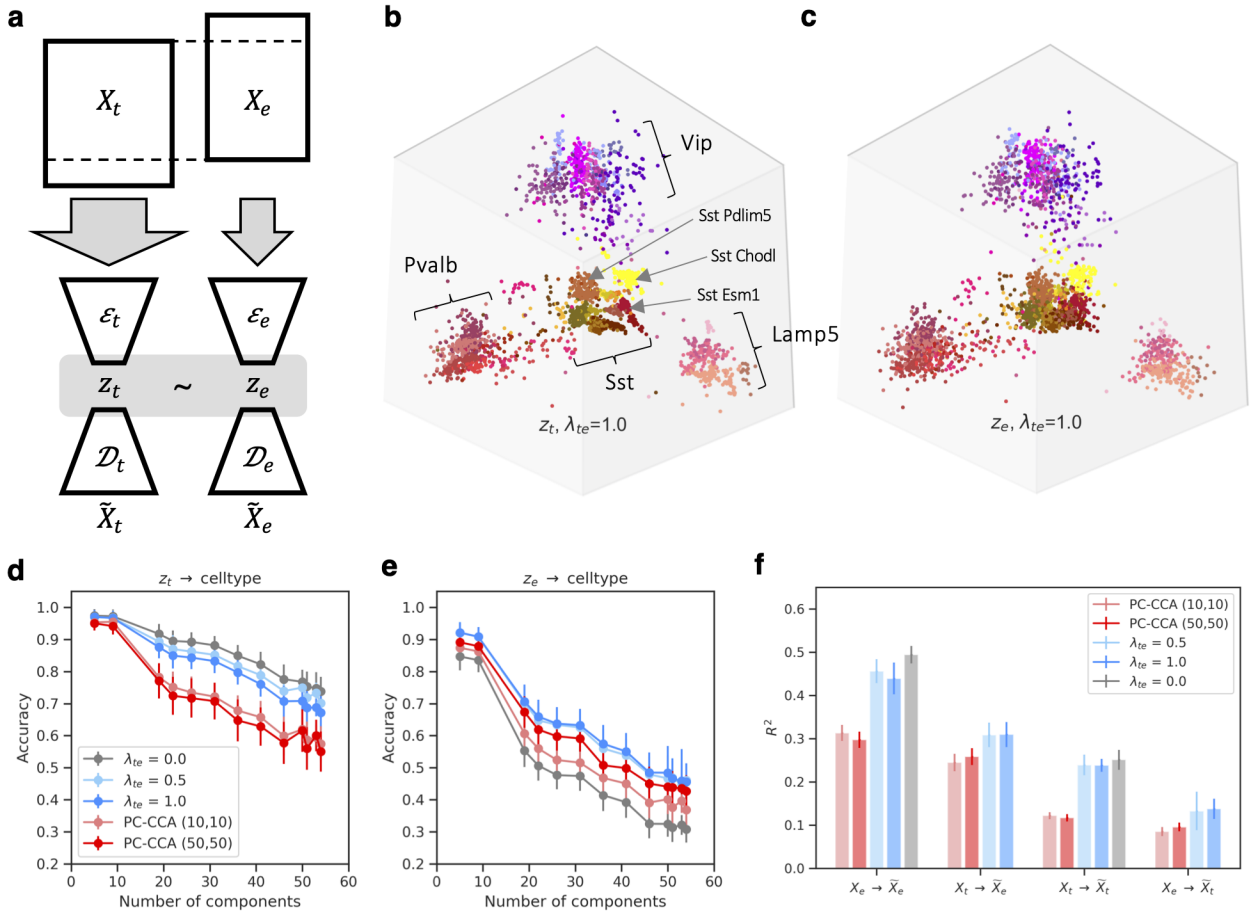

Figure 1: **Coordinated representations of transcriptomic and electrophysiological profiles with coupled autoencoders** (a) Schematic showing the coupled autoencoder architecture for Patch-seq data. Encoders ($\mathcal{E}$) compress input data (X) into low dimensional representations (z). Decoders ($\mathcal{D}$) reconstruct data ($\widetilde{X}$) from representations. The coupling penalty in the loss function encourages representations to be similar across the transcriptomic (t) and electrophysiology (e) modalities. (b-c) 3-d coordinated representations of the transcriptomic and electrophysiological datasets. Each point represents a single cell, colored by the reference hierarchy leaf node to which the cell was mapped to. (d-e) Supervised cell type classification with QDA at different resolutions of the reference hierarchy that were based on 3-d representations obtained with coupled autoencoders and with linear methods. (f) Reconstruction performance as measured with coefficient of determination in the within-modality ($X_\mathrm{e} \to \widetilde{X}_\mathrm{e}$ and $X_\mathrm{t} \to \widetilde{X}_\mathrm{t}$), and in the cross-modality ($X_\mathrm{t} \to \widetilde{X}_\mathrm{e}$ and $X_\mathrm{e} \to \widetilde{X}_\mathrm{t}$) cases. Error bars show (mean $\pm$ SD, 43-fold cross validation) for panels (d-f)

of linear methods (PC-CCA), and coupled autoencoders with $\lambda_{\text{te}} \in \{0.5, 1.0\}$ at these tasks, with the representation dimensionality set to 3. We note that the Patch-seq experiment provides perfect knowledge of *anchors* between the modalities by virtue of paired recordings. In this setting, the popular tool Seurat[16] uses a variant of linear CCA to achieve alignment, for which the performance is expected to be comparable to baselines considered here. Results in Figure 1d-f show that coupled autoencoders learn well-aligned representations of transcriptomic and electrophysiology data, such that cell type labels can be predicted with better accuracy, and the cross-modal data can be inferred more reliably compared to linear methods. Importantly, the within-modality reconstruction error is comparable to that obtained in the uncoupled setting, suggesting that the representations compress the individual data modalities with high fidelity.

Cross-modal data prediction is a key computational tool for identifying corresponding properties of cell types, and in the design of new experiments. Non-linear transformations to align single cell modalities directly in the data domain have been explored before,[17] but crucially did not provide low dimensional co-ordinated representations. We considered a subset of genes that underlie recently discovered cell type specific paracrine signaling pathways in the cortex.[18] The Patch-seq transcriptomic data shows these cell type specific gene expression patterns, Figure 2a. We used only electrophysiology features to infer the expression patterns for all genes in the cross-modal setting, and show results for the same subset of genes as before in Figure 2b. The striking similarity of these expression patterns (Pearson's r=0.89±0.10, mean±SD over cell types) not only demonstrates the effectiveness of coupled autoencoders at the cross-modal prediction task at a granular level, but also suggests that intrinsic electrophysiology contains information regarding neuropeptide communication networks.

We considered cross-modal prediction of electrophysiological features in an analogous manner, pooling values of the features on a per cell type basis. We considered electrophysiological features that are captured by the compressed representation well (within-modality reconstruction $R^2 > 0.25$, Figure S6). While results of Figure 1d-e already suggest that the electrophysiology features are not as specific to transcriptomic cell types, we can nevertheless identify cell type specific patterns, Figure 2c. The cross-modal reconstruction of these features also matches the data (Pearson's r=0.99±0.01, mean±SD over cell types), reinforcing the idea that gene expression can explain many intrinsic electrophysiological features accurately, and that coupled autoencoders are a powerful starting point to unravel such non-linear relationships.

We directly tested the idea that pre-trained coupled autoencoders can be used to predict unobserved cross-modal features in smaller independent experiments by using the Patch-seq

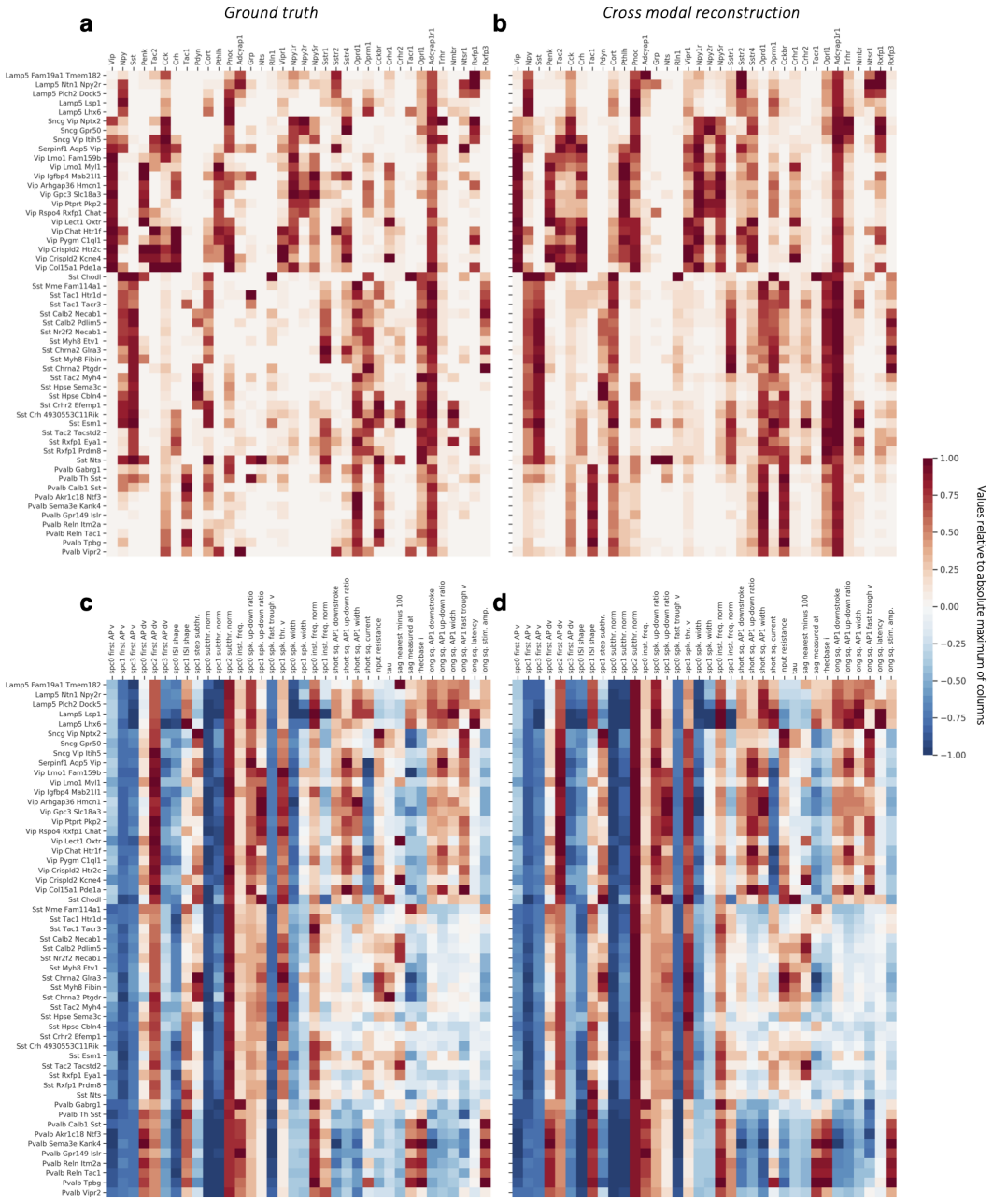

Figure 2: **Cross-modal reconstructions capture cell type specific gene expression patterns and electrophysiological features.**(a) Gene expression levels averaged over samples of individual cell types of the reference taxonomy, normalized per gene by the maximum value of each column. (b) Cell type specificity of different genes is captured well by cross-modal prediction of gene expression profiles from electrophysiological features. (c) A subset of electrophysiological features pooled by cell types shows analogous cell type specificity. (d) Cross-modal reconstructions of the electrophysiology features from gene expression profiles match the measured electrophysiology features.

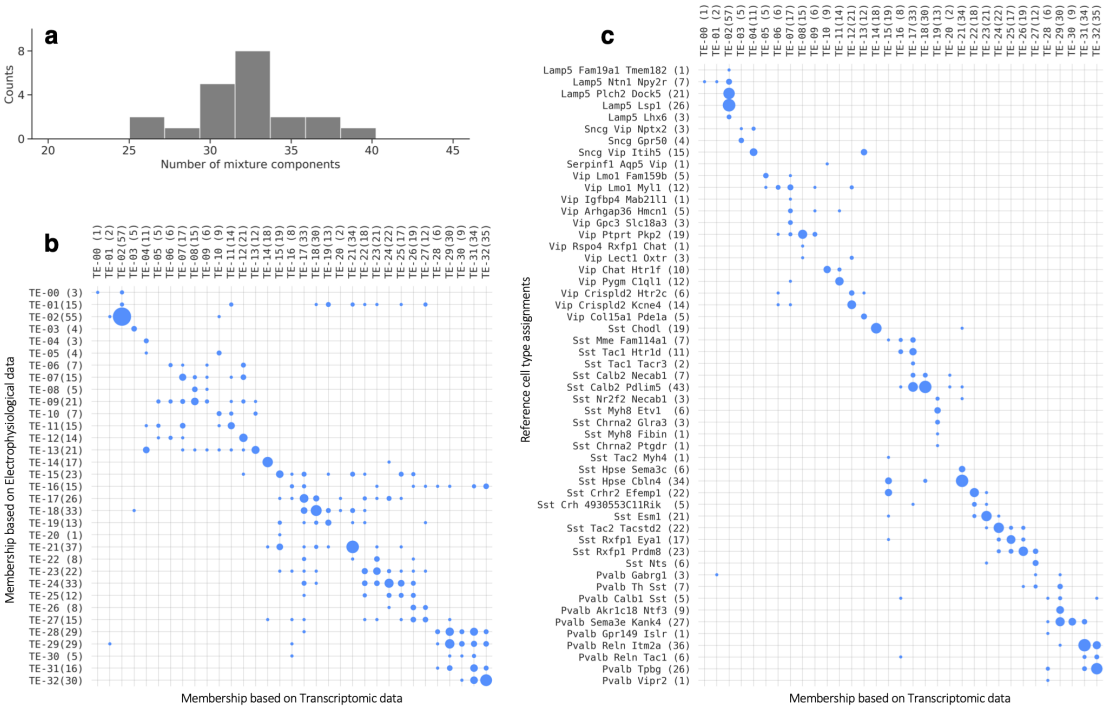

Figure 3: A proposal for consensus clusters (a) Unsupervised clustering using Gaussian Mixtures on the coordinated representation z_t and BIC based model selection suggests ∼33 consensus clusters. (b) Contingency matrix for cluster assignments based on independent, unsupervised clustering of the transcriptomic and electrophysiology representations shows that the clusters are highly consistent. (c) Contingency matrix for the leaf node cell type labels of the reference hierarchy compared to unsupervised cluster assignments show that these unsupervised clusters have substantial overlap with known transcriptomic cell types. Number of test cells for each label are indicated within parentheses next to the label, and area of the dots is proportional to the number cells in panels (b) and (c).

dataset of Scala *et al.*,[19] which includes 107 inhibitory neurons from mouse motor cortex. We applied a coupled autoencoder without additional training to predict the transcriptomic labels and electrophysiological properties of the 107 neurons from their transcriptomic profiles. Results in Figures S8 and S7 show that this approach yields accurate prediction of cell type labels and certain electrophysiological properties, despite ∼5% mismatch between the gene lists and significant differences in electrophysiology protocols.

While clustering of individual modalities into cell type taxonomies shows general correspondence, a strategy for consensus clustering is less clear. The notion of a consensus set of cell types can be formalized as a statistical mixture model. Accordingly, the observation for each cell is explained by a combination of its membership to one of a discrete number of types, and continuous variability around the type representative. Encouraged by the clustering of cells belonging

to similar transcriptomic types in Figure 1b-c, we explored the extent to which such a model can explain the data consistently across modalities. Specifically, we performed unsupervised clustering by fitting a Gaussian mixture model on coordinated representations obtained with the coupled autoencoder to explain both modalities. Figure 3a shows the distribution of optimal number of mixture components over representations obtained with different coupled autoencoder initializations. This plot suggests that the number of clusters that can be consistently defined with coordinated representations has a tight distribution around ∼33. We refer to this *de novo* clustering of the data as consensus clusters. Figure 3b demonstrates that the same consensus cluster can be assigned to neurons not used during training with high frequency, based on observing either the transcriptomic or electrophysiological (but not both) modality. While the dominant diagonal of this contingency matrix indicates the success of this notion of consistent, multimodal cortical cell types, the off-diagonal entries point to imperfections of this view, either due to experimental noise and limitations of experimental characterization, or due to imperfection of the model itself.

Lastly, the consensus clusters are also consistent with the reference transcriptomic taxonomy, Figure 3c. This might suggest over-splitting in the transcriptomic taxonomy and help identify transcriptomic "super-clusters" of GABAergic neurons, as well as point towards the limitations of the dataset, such as having too few samples for certain transcriptomic labels (see Figure S2) to support a mixture component.

In this study, we have presented a principled way to align multimodal observations of neuronal data and define clusters that are consistent across data modalities. Our analysis of the largest multimodal Patch-seq dataset to date with an unsupervised clustering on coordinated representations reveals ∼33 clusters that can be defined consistently with transcriptomic and electrophysiological measurements of cortical GABAergic neurons. We demonstrated that coupled autoencoders trained on reference datasets can serve as efficient look up tables for smaller, single modality neuron characterization to not only infer cell types, but also their properties in other modalities. Refining this ability will enable the design of new kinds of experiments.

An intriguing and essential issue regarding cell types is whether they should be considered as discrete entities or as a continuum.[20] Here, we tested a mixture model view on multimodal data, which allows for types to overlap each other in the representation space so long as the cluster centers are more dominant than the peripheries. With this model, mouse visual cortex interneuron Patch-seq data suggests the existence of ∼33 clusters, more than the ∼5 well-known subclasses but less than the > 50 partitions suggested by scRNA-seq data alone.

Finally, dataset size plays an important role in all our results. More samples can allow the use of larger representation space dimensionality and improve cross-modal data prediction. Similarly, clustering is ill-defined for cell types with too few samples. Therefore, the number of cortical GABAergic interneuron types is likely to grow, and the number of consensus clusters in Figure 3 more likely represents an under-count of the diversity when the notion of cell types is considered as a mixture model.

Methods

Coupled autoencoders

Approaches to discover and extract relationships in multimodal datasets are discussed in literature as cross-modal retrieval, multimodal alignment, multi-view representation learning.[21–23] Deep learning methods such as DeepCCA[24,25] and correspondence autoencoders[26] are promising approaches to achieve multimodal data alignment, but have had limited success in associating complex neural datasets. Our coupled autoencoder networks are related architectures with key improvements to scaling of representations that are critical for the overall quality of learned representations.[27]

We first describe the general coupled autoencoder framework. Then, we show its application to the Patch-seq dataset. For K observation modalities, we represent the coupled autoencoder by

$$\Phi = (\{(\mathcal{E}_i, \mathcal{D}_i, \alpha_i)\}_{1\leq i\leq K}, \lambda), \tag{1}$$

where $\mathcal{E}_i$ and $\mathcal{D}_i$ denote the encoding and decoding networks for the i-th observation modality, α_i sets the relative importance of the different modalities, and $\lambda \geq 0$ sets the relative importance of representation fidelity within observation modalities versus the alignment of different representations.

For a set of paired observations $X = \{(x_{s1}, x_{s2}, \ldots, x_{sK}),\ s \in S\}$, we define the loss due to Φ as

$$L(X;\Phi) = \sum_{s\in S}\left[\sum_{i=1}^{K}\alpha_i\|x_{si} - \mathcal{D}_i(\mathcal{E}_i(x_{si}))\|_2^2 + \lambda \sum_{\substack{i,j\in K,\\ i<j}} \frac{\|\mathcal{E}_i(x_{si}) - \mathcal{E}_j(x_{sj})\|_2^2}{f_{ij}(X)}\right]. \tag{2}$$

That is, each autoencoding agent (Figure 1a) within the coupled architecture processes a separate data modality and optimizes a loss function that consists of penalties for (1) the

discrepancies between the actual input X and reconstructed input $\tilde{X}$ (2) mismatches between the representations learned by the different agents. (A slightly more general treatment can be found in Ref.[27])

In Eq. 2, the functional form of the denominator f_{ij} that scales the mean squared difference between representations of the same sample based on the different data modalities, is crucial to learn good quality representations. Common choices for f_{ij} lead to pathological solutions, i.e. the latent representations collapses into a zero- or one-dimensional space (see Propositions in Supplementary Methods). To avoid such pathological solutions, we propose using:

$$f_{ij}(X) = \min(\sigma^2_{\min}(Z_i), \sigma^2_{\min}(Z_j)) \tag{3}$$

where $\sigma_{\min}(Z_i)$ denotes the minimum singular value of the matrix Z_i, which consists of rows $Z_i(s,:) = z_{si}$ where $z_{si} = \mathcal{E}_i(x_{si})$. In practice, we perform stochastic gradient descent and calculate f_{ij} by its mini-batch approximation. Scaling the coupling loss term in this manner approximates whitening by the full covariance matrix well, and also is practically important when the batch size is small or representation dimensionality is large, regimes where calculating the full covariance matrix would be unreliable and computationally expensive.

Application to the Patch-seq dataset

We use the fact that the same neurons were profiled with both modalities to obtain aligned, low-dimensional representations of the gene expression profiles and electrophysiological features. In the case of just these two data modalities, transcriptomics (t) and electrophysiology (e), the loss function according to Eq. 2 consists of two reconstruction error terms, and a single coupling error term. For a single sample s,

$$L((x_{st}, x_{se})) = \alpha_{\mathrm{t}}\|x_{st} - \mathcal{D}_{\mathrm{t}}(\mathcal{E}_{\mathrm{t}}(x_{st}))\|_2^2 + \alpha_{\mathrm{e}}\|x_{se} - \mathcal{D}_{\mathrm{e}}(\mathcal{E}_{\mathrm{e}}(x_{se}))\|_2^2 + \lambda_{\mathrm{te}}\frac{\|z_{st} - z_{se}\|_2^2}{f_{\mathrm{te}}(X)}, \tag{4}$$

where $z_{st} = \mathcal{E}_{\mathrm{t}}(x_{st})$ and $z_{se} = \mathcal{E}_{\mathrm{e}}(x_{se})$. Here x_{st} denotes gene expression vector for sample s and x_{se} denotes the concatenated sPC and physiological feature measurement vectors for the same sample. The interplay between the accuracy with which the representations capture the individual data modality, versus how well the representations are aligned is a fundamental trade-off that any attempt to define consistent multimodal cell types must resolve (see Supplementary Material for an equivalent formulation in the probabilistic setting). The hyper-parameters α_{t}, α_{e} and λ_{te} explicitly control this trade-off in our formulation (Figure S3).

Data augmentation

Data augmentation is important to regularize the networks and alleviate overfitting, particularly when the dataset size is small. We mimicked the biological dropout phenomenon[28] and used Bernoulli noise (i.e., Dropout[29]) to augment repeated presentations of the transcriptiomic vectors while training. This strategy also renders the network robust to partial mismatches in gene lists, and reduces dependence of the representations and reconstructions on specific marker genes. The individual electrophysiological features have unequal variances, since the total variance in the sPC is normalized on a per-experiment basis. We therefore used additive Gaussian noise with variance proportional to that of the individual features to augment the electrophysiological vectors while training the network. The reconstruction loss for the decoders was calculated with both, the representation obtained by the encoder network of the same modality, and that obtained by the encoder for the other modality. This was done to improve performance of cross-modal prediction. We view this way of calculating the reconstruction loss function as an augmentation strategy for the decoder networks.

Linear baselines

Canonical correlation analysis (CCA) is a standard linear method to align low dimensional representations.[13] To optimize the performance with linear methods, we first used principle component analysis (PCA) to reduce the dimensionality of individual data modalities, followed by CCA to achieve aligned representations across the modalities. The number of dimensions to which the transcriptomic and electrophysiology data were reduced to with PCA is indicated as a tuple in the legends of Figure 1. The dimensionality of CCA representations was chosen to match the dimensionality obtained with coupled autoencoders (dim=3). The inverse CCA and PCA transformations were used to reconstruct data from the representations both, for the the within- and across- modality cases in Figure 1f.

Supervised cell type classification

Label sets obtained at different resolutions of the reference taxonomy were used as ground truth to evaluate representations. The different resolutions correspond to different horizontal levels of the reference taxonomy hierarchy in Figure S1. Starting from the leaf node cell type labels, each cell is assigned the parent node label based on the set of labels that remains at a given level of the hierarchy. Quadratic Discriminant Analysis (QDA)[13] was used to train classifiers on

the representations obtained with coupled autoencoders or CCA, and used to predict the cell type labels for all such label sets. Cells that were not used to train the coupled autoencoder were used to obtain accuracy values shown in Figure 1(d-e) using a k=43 fold cross validation approach. Validation folds were obtained such the class distribution in each fold was similar to that for the overall dataset. Classes with $n \leq 10$ samples in the dataset were discarded from the analysis. Similarly, classes for which there were less than n=6 samples in the training set of any fold were discarded from evaluation for only that fold, since QDA classifier parameters for those poorly represented classes would be unreliable. The results were pooled across the folds for the remaining number of classes (i.e. QDA components) in Figure 1(d-e).

Unsupervised clustering and consensus clusters

Gaussian mixture models with a different number of components (15 to 45 in steps of 1) were fit on z_t obtained with coupled autoencoders ($\lambda_{te} = 1.0$) for 21 different network initializations trained on the same 80% of the dataset. The remaining 20% of cells serve as the test set for this analysis. The training and test sets had similar distributions of the cell type labels based on the reference taxonomy. Each mixture model fit was initialized 50 times and fit until convergence. For the representation from each network initialization, we used Bayesian Information Criterion[13] (BIC) to perform model selection. The distribution for optimal number of mixture components across the 21 different representations was binned using the Freedman-Diaconis rule,[30] Figure 3a. Based on this distribution we estimated the number of clusters that can be consistently defined with coordinated representations to be 33. We picked the model with the lowest reconstruction error, and refer to the mixture model with 33 components fitted on z_t as consensus clusters. The fitted mixture model was then used to assign consensus cluster labels to test cells based on z_t, as well as based on z_e. The consensus cluster assignments obtained in this manner are compared in Figure 3b. We used the Hungarian algorithm to match the consensus clusters with leaf node cell types of the reference taxonomy, using the negative of the contingency matrix based on training cells as the cost function. The order of the consensus clusters in Figure 3b-c reflects this optimal match.

Patch-seq dataset

We used the transcriptomic and electrophysiological profiles of 3,411 GABAergic interneurons from mouse visual cortex of a recent Patch-seq dataset.[9] The dataset includes cell type labels that were obtained by mapping the gene expression profiles to a reference taxonomy.[7] The

relevant taxonomy, and abundances of cells per type are shown in Figure S1 and Figure S2. There are 59 cell types at the highest resolution (i.e. leaf nodes) of this reference taxonomy. A set of 1,252 genes after removing genes related to mitochondria and sex were used as input for the analyses in this study. Gene expression values were CPM normalized, and then $\log_e(\bullet + 1)$ transformed. 44 sparse principle components (sPC) were extracted to summarize the time series data from different portions of the electrophysiology measurement protocol.[9] Additionally 24 measurements of intrinsic physiology features were obtained using the IPFX library `https://ipfx.readthedocs.io/`. The sPC values were scaled to have unit variance per experiment. The remaining features were individually normalized to have zero mean and unit norm. Data was divided into k=43 folds for cross validation experiments. For the consensus cluster experiments, 20% of the cells were set aside as the test set. Different random seeds were used to train networks 21 times on the remaining 80% of the cells.

Code availability

Code for the coupled autoencoder implementation and analysis are available at `https://github.com/AllenInstitute/coupledAE-patchseq`. The coupled autoencoder was implemented using Tensorflow 2.1. Scikit-learn[31] version 0.22.2 implementations of PCA, CCA, QDA and Gaussian Mixture Models, and Scipy version 1.4.1 implementation of the Hungarian algorithm (linear sum assignment) were used to perform the analyses.

Acknowledgements

We wish to thank the Allen Institute for Brain Science founder, Paul G Allen, for his vision, encouragement and support.

References

[1] Robin Tremblay, Soohyun Lee, and Bernardo Rudy. Gabaergic interneurons in the neocortex: from cellular properties to circuits. *Neuron*, 91(2):260–292, 2016.

[2] Hongkui Zeng and Joshua R Sanes. Neuronal cell-type classification: challenges, opportunities and the path forward. *Nature Reviews Neuroscience*, 18(9):530, 2017.

[3] Anirban Paul, Megan Crow, Ricardo Raudales, Miao He, Jesse Gillis, and Z Josh Huang. Transcriptional architecture of synaptic communication delineates gabaergic neuron identity. *Cell*, 171(3):522–539, 2017.

[4] Z Josh Huang and Anirban Paul. The diversity of gabaergic neurons and neural communication elements. *Nature Reviews Neuroscience*, 20(9):563–572, 2019.

[5] Giorgio A Ascoli, Lidia Alonso-Nanclares, Stewart A Anderson, German Barrionuevo, Ruth Benavides-Piccione, Andreas Burkhalter, György Buzsáki, Bruno Cauli, Javier DeFelipe, Alfonso Fairén, et al. Petilla terminology: nomenclature of features of gabaergic interneurons of the cerebral cortex. *Nature Reviews Neuroscience*, 9(7):557, 2008.

[6] Philipp Berens and Thomas Euler. Neuronal diversity in the retina. *e-Neuroforum*, 23(2):93–101, 2017.

[7] 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.

[8] A Zeisel, AB Muñoz-Manchado, S Codeluppi, P Lönnerberg, Manno G La, A Juréus, S Marques, H Munguba, L He, C Betsholtz, C Rolny, G Castelo-Branco, J Hjerling-Leffler, and S Linnarsson. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347:1138–42, Mar 2015.

[9] Nathan W Gouwens, Staci A Sorensen, Fahimeh Baftizadeh, Agata Budzillo, Brian R Lee, Tim Jarsky, Lauren Alfiler, Anton Arkhipov, Katherine Baker, Eliza Barkan, Kyla Berry, Darren Bertagnolli, Kris Bickley, Jasmine Bomben, Thomas Braun, Krissy Brouner, Tamara Casper, Kirsten Crichton, Tanya L Daigle, Rachel Dalley, Rebecca de Frates, Nick Dee, Tsega Desta, Samuel D Lee, Nadezhda Dotson, Tom Egdorf, Lauren Ellingwood, Rachel Enstrom, Luke Esposito, Colin Farrell, David Feng, Olivia Fong, Rohan Gala, Clare Gamlin, Amanda Gary, Alexandra Glandon, Jeff Goldy, Melissa Gorham, Lucas Graybuck, Hong Gu, Kristen Hadley, Michael J Hawrylycz, Alex M Henry, DiJon Hill, Madie Hupp, Sara Kebede, Tae Kyung Kim, Lisa Kim, Matthew Kroll, Changkyu Lee, Katherine E Link, Matthew Mallory, Rusty Mann, Michelle Maxwell, Medea McGraw, Delissa McMillen, Alice Mukora, Lindsay Ng, Lydia Ng, Kiet Ngo, Philip R Nicovich, Aaron Oldre, Daniel Park, Hanchuan Peng, Osnat Penn, Thanh Pham, Alice Pom, Lydia Potekhina, Ramkumar Rajanbabu, Shea Ransford, David Reid, Christine Rimorin, Miranda Robertson, Kara Ronellenfitch, Augustin Ruiz, David Sandman, Kimberly Smith, Josef Sulc, Susan M Sunkin, Aaron Szafer, Michael Tieu, Amy Torkelson, Jessica Trinh, Herman Tung, Wayne Wakeman, Katelyn Ward,

Grace Williams, Zhi Zhou, Jonathan Ting, Uygar Sumbul, Ed Lein, Christof Koch, Zizhen Yao, Bosiljka Tasic, Jim Berg, Gabe J Murphy, and Hongkui Zeng. Toward an integrated classification of neuronal cell types: morphoelectric and transcriptomic characterization of individual GABAergic cortical neurons. *bioRxiv*, 2020.

[10] Federico Scala, Dmitry Kobak, Matteo Bernabucci, Yves Bernaerts, Cathryn R Cadwell, Jesus R Castro, Leonard Hartmanis, Xiaolong Jiang, Sophie R Laturnus, Elanine Miranda, et al. Phenotypic variation within and across transcriptomic cell types in mouse motor cortex. *bioRxiv*, 2020.

[11] Tim Stuart and Rahul Satija. Integrative single-cell analysis. *Nature Reviews Genetics*, 20(5):257–272, 2019.

[12] Maria Colomé-Tatché and Fabian J Theis. Statistical single cell multi-omics integration. *Current Opinion in Systems Biology*, 7:54–59, 2018.

[13] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[14] Dmitry Kobak, Yves Bernaerts, Marissa A Weis, Federico Scala, Andreas Tolias, and Philipp Berens. Sparse reduced-rank regression for exploratory visualization of multimodal data sets. *bioRxiv*, page 302208, 2019.

[15] NW Gouwens, SA Sorensen, J Berg, C Lee, T Jarsky, J Ting, SM Sunkin, D Feng, CA Anastassiou, E Barkan, K Bickley, N Blesie, T Braun, K Brouner, A Budzillo, S Caldejon, T Casper, D Castelli, P Chong, K Crichton, C Cuhaciyan, TL Daigle, R Dalley, N Dee, T Desta, SL Ding, S Dingman, A Doperalski, N Dotson, T Egdorf, M Fisher, Frates RA de, E Garren, M Garwood, A Gary, N Gaudreault, K Godfrey, M Gorham, H Gu, C Habel, K Hadley, J Harrington, JA Harris, A Henry, D Hill, S Josephsen, S Kebede, L Kim, M Kroll, B Lee, T Lemon, KE Link, X Liu, B Long, R Mann, M McGraw, S Mihalas, A Mukora, GJ Murphy, L Ng, K Ngo, TN Nguyen, PR Nicovich, A Oldre, D Park, S Parry, J Perkins, L Potekhina, D Reid, M Robertson, D Sandman, M Schroedter, C Slaughterbeck, G Soler-Llavina, J Sulc, A Szafer, B Tasic, N Taskin, C Teeter, N Thatra, H Tung, W Wakeman, G Williams, R Young, Z Zhou, C Farrell, H Peng, MJ Hawrylycz, E Lein, L Ng, A Arkhipov, A Bernard, JW Phillips, H Zeng, and C Koch. Classification of electrophysiological and morphological neuron types in the mouse visual cortex. *Nat Neurosci*, 22:1182–1195, Jul 2019.

[16] 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.

[17] Matthew Amodio and Smita Krishnaswamy. Magan: Aligning biological manifolds. *arXiv preprint arXiv:1803.00385*, 2018.

[18] Stephen J Smith, Uygar Sümbül, Lucas T Graybuck, Forrest Collman, Sharmishtaa Seshamani, Rohan Gala, Olga Gliko, Leila Elabbady, Jeremy A Miller, Trygve E Bakken, et al. Single-cell transcriptomic evidence for dense intracortical neuropeptide networks. *Elife*, 8:e47889, 2019.

[19] F Scala, D Kobak, S Shan, Y Bernaerts, S Laturnus, CR Cadwell, L Hartmanis, E Froudarakis, JR Castro, ZH Tan, S Papadopoulos, SS Patel, R Sandberg, P Berens, X Jiang, and AS Tolias. Layer 4 of mouse neocortex differs in cell types and circuit organization between sensory areas. *Nat Commun*, 10:4174, Sep 2019.

[20] Kenneth D Harris, Hannah Hochgerner, Nathan G Skene, Lorenza Magno, Linda Katona, Carolina Bengtsson Gonzales, Peter Somogyi, Nicoletta Kessaris, Sten Linnarsson, and Jens Hjerling-Leffler. Classes and continua of hippocampal ca1 inhibitory neurons revealed by single-cell transcriptomics. *PLoS biology*, 16(6):e2006387, 2018.

[21] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[22] Yingming Li, Ming Yang, and Zhongfei Zhang. A survey of multi-view representation learning. *IEEE transactions on knowledge and data engineering*, 31(10):1863–1883, 2018.

[23] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A Comprehensive Survey on Cross-modal Retrieval, 2016.

[24] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255, 2013.

[25] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International Conference on Machine Learning*, pages 1083–1092, 2015.

[26] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 7–16, 2014.

[27] Rohan Gala, Nathan Gouwens, Zizhen Yao, Agata Budzillo, Osnat Penn, Bosiljka Tasic, Gabe Murphy, Hongkui Zeng, and Uygar Sümbül. A coupled autoencoder approach for multi-modal analysis of cell types. In *Advances in Neural Information Processing Systems*, pages 9263–9272, 2019.

[28] Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740, 2014.

[29] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[30] David Freedman and Persi Diaconis. On the histogram as a density estimator:L 2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57(4):453–476, dec 1981.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[32] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.