Published in final edited form as:

Nat Comput Sci. 2021 February ; 1(2): 120–127. doi:10.1038/s43588-021-00030-1.

Consistent cross-modal identification of cortical neurons with coupled autoencoders

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

Allen Institute, Seattle, WA, USA.

Abstract

Consistent identification of neurons in different experimental modalities is a key problem in neuroscience. Although methods to perform multimodal measurements in the same set of single neurons have become available, parsing complex relationships across different modalities to uncover neuronal identity is a growing challenge. Here we present an optimization framework to learn coordinated representations of multimodal data and apply it to a large multimodal dataset profiling mouse cortical interneurons. Our approach reveals strong alignment between transcriptomic and electrophysiological characterizations, enables accurate cross-modal data prediction, and identifies cell types that are consistent across modalities.

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]. This has prompted major consortia such as the BRAIN Initiative Cell Census Network (BICCN) to seek a comprehensive characterization of cell types and their function[7]. However, the effectiveness of such approaches is predicated on the existence of cellular identities that manifest consistently across different observation modalities, and our ability to identify them. Although single-cell RNA sequencing (scRNA-seq) experiments have uncovered a detailed transcriptomic organization of cortical cells in the mouse brain[8,9], emerging experimental techniques now enable concurrent characterization of multiple aspects of neuronal identity and function[7]. For example, MERFISH[10] can provide paired in situ measurements of anatomy and gene expression of multiple neurons in intact tissue, and the Patch-seq protocol[11] can characterize morphology, electrophysiology, and gene expression

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to R.G. or U.S. rohang@alleninstitute.org; uygars@alleninstitute.org.

Author contributions

R.G. and U.S. designed the methodology and wrote the manuscript. R.G. wrote software and performed formal analysis. A.B., F.B., J.M., N.G., A.A., and G.M. performed data curation and pre-processing. R.G., B.T. M.H. and U.S. wrote the manuscript revisions. R.G., A.B., F.B., J.M., N.G., G.M., B.T., H.Z., M.H., and U.S. conceptualized the study.

Competing interests

The authors declare no competing interests.

Extended data is available for this paper at https://doi.org/10.1038/s43588-021-00030-1.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-021-00030-1.

of single neurons in tissue slices. Aligning modalities in such paired reference datasets offers the opportunity to move towards a unified, multimodal view of cellular diversity and potentially enable translation of individual measurements across modalities with high fidelity.

Aligning multimodal data for cell-type research is challenging due to the complexity of biological relationships between modalities, difficulties in measuring signal and quantifying noise in each modality, and the high-dimensional nature of measurements. We present a deep neural network-based methodology referred to as coupled autoencoders to perform alignment for paired datasets, demonstrate its utility for the multimodal cell-type identification problem, and provide an unsupervised, data-driven characterization of GABAergic cell type diversity, which has been a central problem in neurobiology[5,12-14]. Classical approaches to group GABAergic cells on the basis of anatomy, physiology, and so on alone typically disagree on both the number and identity of cell types[5], presumably as the relative importance of the features within an observation modality is unknown. Yet, unequivocal identification of interneurons is essential to elucidate the brain circuits that they participate in. Moreover, discordant results cast doubt on the hypothesis that neurons have unique identities, whereby different experiments reveal different facets of those identities, potentially through complicated transformations and noise processes. Here we focus on the two modalities with the largest number of paired samples in a recent Patch-seq dataset[14]. There are neither overlapping features nor known associations across the two modalities. Using the fact that the same samples are measured in each modality, our goal is to formulate cell identities that are consistent across these modalities.

Results

Coupled autoencoders consist of multiple autoencoder networks, each of which comprises 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 (Fig. 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, hyperparameter λ (Methods) controls the relative importance of achieving accurate reconstructions versus learning representations that are similar across modalities.

If a common latent representation exists across transcriptomic and electrophysiological measurements that captures salient characteristics of neurons in the individual data modalities, an important consequence will be that unimodal electrophysiological measurements of interneurons can be used to predict gene expression and vice versa. This ability to translate measurements across modalities may enable researchers to test cell-type-specific hypotheses without performing costly and potentially intractable multimodal experiments. The ability to align these modalities would strongly support 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[15,16] can align the major cell classes to some extent, a more detailed alignment of features and cell types may require nonlinear transformations.

A further consequence of such aligned representations would be the ability to identify (without supervision) cell types of key classes such as GABAergic interneurons in the mouse visual cortex that are consistent across transcriptomic and electrophysiological characterizations of this neuron population. The level of agreement between those clusters and a reference transcriptomic taxonomy of cortical cell types[8], and the degree of perturbation of cluster boundaries with respect to that reference taxonomy, can enhance both practical and conceptual aspects of our understanding of cell types.

Aligned three-dimensional representations z_t and z_e for the high-dimensional transcriptomic and electrophysiological profiles X_t and X_e obtained with coupled autoencoders are shown in Fig. 1b,c. Cells labeled according to the reference transcriptomic taxonomy (Extended Data Figs. 1 and 2) cluster together in representations of both observation modalities. Moreover, the representations largely preserve hierarchical relationships between cell types of the reference taxonomy. 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 (see Methods) of the reference taxonomy from z_t and z_e in Fig. 1d,e, and data reconstruction performance in Fig. 1f. First consider the uncoupled ($\lambda_{te} = 0$) setting in which each autoencoder performs nonlinear dimensionality reduction independently for its respective input modality. Representations based on the transcriptomic data alone (z_t, $\lambda_{te} = 0$) are best suited for supervised cell-type classification using quadratic discriminant analysis (QDA), leading to 0.74 ± 0.05 accuracy for leaf node cell-type labels (Fig. 1d). This is not surprising as the reference transcriptomic taxonomy was derived from analyses of gene expression alone. Electrophysiological profiles are expected to be noisy, and of lower resolution compared with transcriptomic profiles[17]. Nevertheless, in Fig. 1e, classifiers based on representations of electrophysiology alone (z_t, $\lambda_{te} = 0$) predict leaf node cell-type labels with an accuracy of 0.31 ± 0.04 (where the chance level is 0.03). To add context, a model trained only to classify leaf node cell types on the basis of electrophysiological profiles alone led to an accuracy of 0.52 ± 0.04 (see Methods). Finally, within-modality reconstruction accuracies of uncoupled representations provide an upper limit for both within- and cross-modal reconstructions that may be achieved with three-dimensional representations obtained with coupled autoencoders.

To evaluate whether complicated, nonlinear transformations underlie the relationship between the transcriptomic and electrophysiological features of neurons, we considered the performance of linear methods (principal components canonical correlation analysis or PC-CCA) and coupled autoencoders with $\lambda_{te} \in \{0.5, 1.0\}$ at these tasks, with the representation dimensionality set to three. The weak dependence on λ_{te} in Fig. 1 and Extended Data Fig. 3 suggests that our method is robust with respect to this hyperparameter. Data augmentation (see Methods) further stabilizes the coupling (Extended Data Fig. 4). A latent dimensionality of $3 \leq d \leq 10$ can capture the variability in the dataset (Extended Data Fig. 5). We choose $d = 3$ to minimize the number of parameters needed for downstream tasks. We note that the Patch-seq experiment provides—to the extent of experimental measurement—perfect knowledge of anchors between the modalities by virtue of paired measurements. In this setting, the popular tool Seurat[18] uses a variant of linear CCA to achieve alignment, for which the performance is expected to be comparable with baselines considered here. Results in Fig. 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 predicted more reliably compared with linear methods. Importantly, the within-modality reconstruction error is comparable to that obtained in the uncoupled setting, demonstrating that coupled representations enable alignment across modalities while faithfully compressing the individual data modalities.

Cross-modal data prediction (Extended Data Figs. 6 and 7) is a key computational tool for identifying corresponding properties of cell types and guiding the design of new experiments. We considered a subset of genes that underlie recently discovered cell type specific paracrine signaling pathways in the cortex[19]. The Patch-seq transcriptomic data shows these cell-type-specific gene expression patterns (Fig. 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. The striking similarity of these expression patterns (Pearson's $r = 0.89 \pm 0.10$, mean ± s.d. over cell types; Fig. 2b and Extended Data Fig. 6) demonstrates the effectiveness of coupled autoencoders at the cross-modal prediction task at a granular level. Similar results were obtained for GABAergic cell-type marker genes (Supplementary Fig. 1). Neuropeptide precursor genes and their cognate G-protein-coupled receptors have widespread expression in the cortex and are implicated to form cell-type-specific broadcast communication networks[19,20]. The high degree of similarity in Fig. 2a,b therefore provides indirect evidence for coordination between intrinsic cellular electrophysiology and circuit-level neurotransmitter networks in the cortex (a link between electrophysiology and cell adhesion molecules was previously studied using scRNA-seq[21]).

We considered cross-modal prediction of electrophysiological features in an analogous manner, pooling values of the features on a per-cell-type basis, and focusing on features that are captured by the compressed representation well (within-modality reconstruction $R^2 > 0.42$; Extended Data Fig. 7). Although the results of Fig. 1d,e already suggest that the electrophysiology features are not as specific to transcriptomic cell types, we can nevertheless identify cell type specific patterns (Fig. 2c). The cross-modal reconstruction of these features also matches the data (Pearson's $r = 0.98 \pm 0.02$, mean ± s.d. over cell types), reinforcing the result that gene expression can explain many intrinsic electrophysiological features accurately, and that coupled autoencoders are a powerful starting point to unravel such nonlinear relationships. Moreover, the per-feature prediction accuracy can help uncover the features that are important for neuronal identity (for example, *Vip*, *Sst, Npy1r,* and *Oprd1,* the up–down ratio of the action potential, and the rheobase current; Extended Data Figs. 6 and 7)

We directly tested the idea that pre-trained coupled autoencoders can be used to predict unobserved cross-modal features in independent experiments by using two recent Patch-seq datasets[22,23], which include 107 and 524 inhibitory neurons from the mouse motor cortex, respectively. We applied a coupled autoencoder trained on the dataset in this work without extra training to predict the transcriptomic labels and electrophysiological properties of these neurons from their transcriptomic profiles. The results in Supplementary Figs. 2 and 3, and Extended Data Figs. 8 and 9 show that this approach yields accurate prediction of cell-type labels and certain electrophysiological properties, despite a 5% mismatch between the gene lists, differences in electrophysiology protocols and brain regions.

Although clustering of individual modalities into cell-type taxonomies shows general correspondence, a strategy for consensus clustering is less clear. The notion of a finite set of cell types can be formalized as a statistical mixture model, whereby 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. We explored the extent to which such a model can explain the data consistently across modalities. We performed unsupervised clustering on coordinated representations obtained with the coupled autoencoder to explain both modalities. Figure 3a shows the distribution (32.19 ± 3.16, mean ± s.d.) of optimal number of Gaussian mixture components over representations obtained with different network initializations. We take the ceiling of the mean (33) as the number of clusters that can be consistently defined with coordinated representations, and refer to this de novo clustering as consensus clusters. Figure 3b and Supplementary Fig. 4 demonstrate that the same consensus cluster can be assigned to neurons with high frequency, based on observing either the transcriptomic or electrophysiological (but not both) modality. Although 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. As a metric of the consensus between assignments across modalities, we calculated the ratio of clusters for which the diagonal entry of the contingency matrix is at least as large as the off-diagonals of the corresponding row and column. We obtained $c_{\mathrm{ref}} = 0.26 \pm 0.01$ for the reference labels, whereas we obtained $c_{\mathrm{con}} = 0.87 \pm 0.04$ for the consensus clusters on all cells. We obtained $c_{\mathrm{ref}} = 0.26 \pm 0.01$ and $c_{\mathrm{con}} = 0.58 \pm 0.03$ for the test cells (see Methods and Supplementary Fig. 4). These results suggest that consensus clusters can be used to produce cell-type assignments for which there is a better agreement across experimental modalities, compared with the reference taxonomy.

The consensus clusters are also consistent with the reference taxonomy, although not to the degree of all leaf node labels (Fig. 3c). This can indicate over-split (for example, Lamp5 Plch2 Dock5, Lamp5 Lsp1), under-split (for example, Sst Calb2 Pdlim5) and not-tight (for example, Vip Lmo1 Myl1, Sst Mme Fam114a1) cell types in the reference taxonomy. To quantify the degree to which different label sets represent the underlying data, we report the average silhouette score for test samples (not used to train the coupled autoencoder or the mixture model) for each label, using z_{t} and z_{e}, and compare consensus clusters against those of the reference transcriptomic taxonomy (Fig. 3d,e). A negative silhouette score suggests an unreliable cluster (grayed out labels in Fig. 3b,c). Not only do consensus clusters capture the structure of the data better than the reference labels on z_{t} (the average silhouette score is $S_{\mathrm{con}} = 0.24 \pm 0.01$ for consensus clusters and $S_{\mathrm{ref}} = 0.14 \pm 0.01$ for reference labels; mean ± s.d. over the five best initializations), but also on z_{e} ($S_{\mathrm{con}} = 0.04 \pm 0.02$, $S_{\mathrm{ref}} = -0.13 \pm 0.01$). For the reference taxonomy, we repeated this analysis using standalone (rather than aligned) representations and for different hierarchical mergings of the taxonomy with at least 33 labels, and obtained similar results (see Methods, Extended Data Fig. 10). These results suggest that consensus clusters are a more identifiable characterization of the joint transcriptomic and electrophysiological diversity of interneurons than one based on transcriptomics alone.

Discussion

Our analysis of what is so far the largest multimodal Patch-seq dataset of cortical GABAergic neurons with unsupervised clustering on coordinated representations reveals approximately 33 clusters that can be defined consistently with transcriptomic and electrophysiological measurements, providing a deeper association of these modalities than previously explained. Beyond inferring cell types, coupled autoencoders trained on reference datasets can serve as efficient translators for experiments using a single observation modality to infer neuronal properties in other modalities. This capability can provide indirect evidence for/against hypotheses that are hard to test, such as predicting the expression levels of genes regulating ion channels of interest, purely from observations of intrinsic electrophysiology.

An intriguing and essential issue regarding cell types is to what extent they are inherently discrete entities or representatives of a continuum[24]. A mixture model allows types to overlap each other in the representation space so long as the cluster centres 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 approximately five well-known subclasses but less than the >50 partitions suggested by scRNA-seq data alone. A potential caveat is that scRNA-seq in the Patch-seq experiment is noisier than standalone transcriptomic measurements[14]. Coupled autoencoders can jointly analyze multiple modalities. Future work can incorporate additional observation modalities (for example, morphology, connectivity) to improve our understanding of neuronal identity.

Finally, dataset size plays an important role in all of 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; further analysis of consensus or transcriptomic clusters (Fig. 3c) should take sample size into account. The number of cortical GABAergic interneuron types is therefore likely to grow and the number of consensus clusters in Fig. 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 the literature as cross-modal retrieval, multimodal alignment, multiview representation learning[25-27]. Deep learning methods such as DeepCCA[28,29] and correspondence autoencoders[30] are promising approaches to achieve multimodal data alignment, but have had limited success in associating complex neural datasets (see the Supplementary Information for an overview of modern data alignment approaches). Our coupled autoencoder networks are related architectures with key improvements to scaling of representations, that are critical for the overall quality of learned representations[31].

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

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

where $\mathscr{E}_i$ and $\mathscr{D}_i$ denote the encoding and decoding networks, respectively, for the ith 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 $X = \{(x_{s1}, x_{s2}, \ldots, x_{sK}), s \in \mathscr{S}\}$, we define the loss due to $\boldsymbol{\Phi}$ as

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

That is, each autoencoding agent (Fig. 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 $\widetilde{X}$ and (2) mismatches between the representations learned by the different agents. A slightly more general treatment can be found in ref.[31].

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

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

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

Representations collapse for common scaling function choices.

Proofs for the following propositions can be found in ref.[31].

Proposition 1. Let $f_{ij} = 1$. Representations of the coupled autoencoder that minimize the loss in equation (2) satisfy $\|z_{si}\| < \epsilon$ for any norm $\| \cdot \|$, input set X, $\epsilon > 0$ and all s,i.

Proposition 2. Let f_{ij} implement batch normalization[32]. Representations of the coupled autoencoder that minimize the loss in equation (2) satisfy $| \, z_{si}(m) - z_{si}(\bar{m}) \, | \; < \epsilon$ for all $s, i, m, \bar{m}$, and $\epsilon > 0$, with probability 1.

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 (t and e), the loss function according to equation (2) consists of two reconstruction error terms and a single coupling error term. For a single sample s,

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

where $z_{st} = \mathscr{E}_{\mathrm{t}}(x_{st})$ and $z_{se} = \mathscr{E}_{\mathrm{e}}(x_{se})$. Here x_{st} denotes gene expression vector for sample s and x_{se} denotes the concatenated sparse principal component (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. The hyperparameters α_{t}, α_{e} and λ_{te} explicitly control this trade-off in our formulation (Extended Data Fig. 3 shows behavior over a range of these values with the Patch-seq dataset). We set all three parameters to 1.0 for all central analyses in this manuscript.

Although not used explicitly in this study, we would like to point out that the deterministic view in equation (4) is equivalent to maximizing log-likelihood of a discriminative probabilistic model for independent and identically distributed observations[31]:

$$\begin{aligned} \sum_{s \in S} \log \mathrm{p}(\mathrm{x}_{\mathrm{st}}, \mathrm{x}_{\mathrm{se}}, \hat{\mathrm{z}}_{\mathrm{st}} \mid \hat{\mathrm{z}}_{\mathrm{se}}) &= \sum_{s \in S} \log \mathrm{p}(\mathrm{x}_{\mathrm{st}} \mid \hat{\mathrm{z}}_{\mathrm{st}}, \hat{\mathrm{z}}_{\mathrm{se}}) + \log \mathrm{p}(\mathrm{x}_{\mathrm{se}} \mid \hat{\mathrm{z}}_{\mathrm{se}}) + \log \mathrm{p}(\hat{\mathrm{z}}_{\mathrm{st}} \mid \hat{\mathrm{z}}_{\mathrm{se}}) \\ &= \sum_{s \in S} \log \mathrm{p}(\mathrm{x}_{\mathrm{st}} \mid \hat{\mathrm{z}}_{\mathrm{st}}) + \log \mathrm{p}(\mathrm{x}_{\mathrm{se}} \mid \hat{\mathrm{z}}_{\mathrm{se}}) + \log \mathrm{p}(\hat{\mathrm{z}}_{\mathrm{st}} \mid \hat{\mathrm{z}}_{\mathrm{se}}), \end{aligned} \tag{5}$$

where we assume that x_{se} is independent of x_{st} and $\hat{z}_{st}$ given $\hat{z}_{se}$, and that x_{st} is independent of $\hat{z}_{se}$ given $\hat{z}_{st}$. By modeling the conditional probabilities $p(x_{st} \mid \hat{z}_{st})$, $p(x_{se} \mid \hat{z}_{se})$ and $p(\hat{z}_{st} \mid \hat{z}_{se})$ as multivariate normal distributions with diagonal covariances $x_{st} \mid \hat{z}_{st} \sim \mathscr{N}(\tilde{x}_{st}, \sigma_{\mathrm{t}}^2 I)$, $x_{se} \mid \hat{z}_{se} \sim \mathscr{N}(\tilde{x}_{se}, \sigma_{\mathrm{e}}^2 I)$ and $\hat{z}_{st} \mid \hat{z}_{se} \sim \mathscr{N}(\hat{z}_{se}, \lambda^{-1} I)$, we can write equation (5) as

$$\sum_{s \in S} \log \mathrm{p}(\mathrm{x}_{\mathrm{st}}, \mathrm{x}_{\mathrm{se}}, \hat{\mathrm{z}}_{\mathrm{st}} \mid \hat{\mathrm{z}}_{\mathrm{se}}) = \frac{-1}{2} \sum_{s \in S} \sigma_{\mathrm{t}}^{-2} \| x_{st} - \tilde{x}_{st} \|_2^2 + \sigma_{\mathrm{e}}^{-2} \| x_{se} - \tilde{x}_{se} \|_2^2 + \lambda \, \| \hat{z}_{st} - \hat{z}_{se} \|_2^2 + \mathrm{const}\,. \tag{6}$$

Comparing the loss function in equation (6) with equation (4), we see that $\alpha_{\mathrm{t}} \approx \sigma_{\mathrm{t}}^{-2}$ and $\alpha_{\mathrm{e}} \approx \sigma_{\mathrm{e}}^{-2}$ are related to the noise in the measurements, and $\lambda_{\mathrm{te}} \approx \lambda$ denotes the precision in cross-modal latent variable estimation.

The two modalities t and e are interchangeable in equation (5), and Fig. 1b suggests that the individual cell types may be well-approximated by hyperellipsoids; fitting a Gaussian mixture model to the encodings therefore provides an efficient prior distribution for $p(\hat{z}_{st})$ (or $p(\hat{z}_{se})$) and produces a generative model for multimodal datasets.

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[33] and used Bernoulli noise (that is, Dropout[34]) to augment repeated presentations of the transcriptomic 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 as 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 using both the representation obtained by the encoder network of the same modality, as well as that obtained by the encoder for the other modality. This was done to improve performance of cross-modal prediction. Such a way of calculating the reconstruction loss can be viewed as an augmentation strategy for the decoder networks that considerably improves the accuracy of cross-modal prediction (Extended Data Fig. 4).

Linear baselines.

Canonical correlation analysis is a standard linear method to align low-dimensional representations[15]. To optimize the performance with linear methods, we first used principal 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 Fig. 1. The dimensionality of CCA representations was chosen to match the dimensionality obtained with coupled autoencoders (dimensionality = 3). The inverse CCA and PCA transformations were used to reconstruct data from the representations for the within- and across-modality cases in Fig. 1f. Reconstruction performance for different representation dimensionality is compared in Extended Data Fig. 5.

Supervised cell-type classification.

Label sets obtained at different resolutions of the reference transcriptomic taxonomy were used as ground truth to evaluate representations. The different resolutions correspond to different horizontal levels of the reference taxonomy hierarchy in Extended Data Fig. 1. 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; thus, at the lowest resolution, there is just a single class (n59) that encompasses all neurons, and at the highest resolution there are 53 classes that are cell-type labels (Extended Data Fig. 1). We used QDA[15] to perform classification with 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 the accuracy values shown in Fig. 1d,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, as QDA classifier parameters for those poorly represented classes would be unreliable. The results were pooled across the folds for the remaining number of classes (that is, QDA components) in Fig. 1d,e. The architecture of the neural network only trained to classify cell types at the highest resolution of the taxonomy using only electrophysiological profiles used the same encoder network as for the X_e autoencoder, except that the output was a 53 way classification. The network was trained with the standard cross-entropy loss for classification.

Unsupervised clustering and consensus clusters.

For this analysis, 80% of the cells were used for training and the remaining 20% served as the test set. The training and test sets had similar distributions of the cell-type labels based on the reference taxonomy. The coupled autoencoder ($\lambda_{te} = 1.0$) was initialized 21 times to obtain as many different representations. For each representation, Gaussian mixture models with different numbers of components (15 to 45 in steps of 1) were fit on z_t utilizing only the training set. The model with the lowest value of the Bayesian information criterion[15.] on the training set was used to determine the optimal number of components. The distribution for optimal number of mixture components across the 21 different representations was binned using the Freedman–Diaconis rule[35] (Fig. 3a). Based on this distribution (32.19 ± 3.16, mean ± s.d.), we use the ceiling of the mean as the number of clusters that can be consistently defined with coordinated representations. These 33 mixture components are referred to as consensus clusters. We picked the representation from coupled autoencoder model with the best total reconstruction error to show results on the test cells. The 33 component Gaussian 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 Fig. 3b (Supplementary Fig. 5 shows the individual BIC plots for the different representations, and offers a comparison with a similar analysis using PC-CCA representations). 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 Fig. 3b,c reflects this optimal match.

Evaluating the agreement and consistency of cluster assignments.

We calculated the following fraction to evaluate the agreement between assignments across the experimental modalities on a per-cluster basis:

$$c = \frac{\sum_{m=1}^{M} \mathbf{1}(C(m,m)) \geq \phi \max\{\max_{i} C(m,i), \max_{i} C(i,m)\}}{M},$$

where M denotes the number of clusters, C denotes the contingency matrix for which the fraction is calculated and **1** denotes the indicator function; $\mathbf{1}(a) = 1$ if a holds, otherwise $\mathbf{1}(a) = 0$; c_{con} (c_{ref}) refers to this quantity when calculated with the contingency matrix calculated over the consensus labels (reference taxonomy merged to 33 labels). We set the scalar factor $\phi \geq 1$ to $\phi = 1$ in the reported experiments. We also found that the conclusion that $c_{\mathrm{con}} > c_{\mathrm{ref}}$ is not sensitive to this factor over a broad range of values (1 to 5). To obtain uncertainty estimates, we selected the best 5 of the 21 networks used for training on the same 80% of the dataset with different initializations, on the basis of the lowest total reconstruction error on the test set. We performed clustering with 33 mixture components to obtain the labels and calculate the corresponding contingency matrices. For any other label set, we use the representation to train QDA classifiers and assign labels to the test set to obtain the contingency matrix.

We used silhouette analysis to evaluate the ability of a set of labels in representing the underlying data. Although the silhouette values reported in the main text, S_{con} and S_{ref} are averages over all involved samples, the values reported in the corresponding figures represent averages on a per-cluster basis. To obtain uncertainty estimates, we selected the best 5 of the 21 networks used for training on the same 80% of the dataset with different initializations, on the basis of the lowest total reconstruction error on the test set. Silhouette scores were obtained for the different label sets using the test set representations obtained from these networks.

Patch-seq dataset.

We used the transcriptomic and electrophysiological profiles of GABAergic interneurons from mouse visual cortex of a recent Patch-seq dataset[14]. Briefly, neurons were patched with biocytin-filled electrodes with which the electrophysiological responses to a series of hyperpolarizing and depolarizing current injections were recorded, nuclear and cytosolic mRNA was extracted, reverse transcribed and the resulting cDNA was sequenced using SMART-Seq v4 (ref. [14]). Among the 3,708 cortical GABAergic interneurons reported in that work, axonal and dedritic morphology of only 350 cells was reconstructed. We restricted our analysis to the two modalities with largest number of samples, and dropped the morphology modality altogether. The 3,708 cells were also mapped to the reference transcriptomic hierarchical taxonomy[8] with different levels of confidence. We discarded cells which were annotated as inconsistent[14] based on this confidence level, leaving 3,411 cells for which both transcriptomic and electrophysiological profiles were available. The relevant taxonomy and abundances of cells per type for well-representated types (at least ten samples per type) are shown in Extended Data Figs. 1 and 2.

A set of 1,252 genes used as input for the analyses in this study. The gene selection procedure included two filtering steps. The first step excluded genes where the primary source of variation in gene expression is unlikely to be related to its cell-type identity. Specifically, we removed (1) genes that are highly expressed in non-neuronal cells, (2) genes with reported sex or mitochondrial associations, and (3) genes that are much more highly expressed in Patch-seq data versus fluorescence-activated cell-sorting data (or vice versa) and therefore may be associated with the experimental platform[14]. We also removed gene models and some other families of unannotated genes that may be difficult to interpret. The second filtering step used the β score, which is a published measure of how binary a gene is with respect to cell types[36]. Higher β indicates that for each cell type a gene is either expressed (with CPM > 1) or unexpressed in most cells. We removed all genes with $\beta <$ 0.4, leaving a total of 1,252 genes in the analysis after both filtering steps. Gene expression values were CPM normalized and then $\log_e(\bullet + 1)$ transformed.

Fourty-four sPCs were extracted to summarize the time series data from different portions of the electrophysiology measurement protocol[14]. An additional 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, whereas the remaining features were individually normalized to have zero mean and unit norm. We have used the following abbreviations to name electrophysiological features: action potential (AP), first action potential (AP1), inter-spike interval (ISI), threshold (thr.), sub-threshold (subthr.), instantaneous (inst.), frequency (freq.), spike (spk.), square (sq.), stimulus (stim.), amplitude (amp.), membrane potential (v), membrane potential time derivative (dv), current (i). Note that the coupled autoencoder trains to minimize the mean squared error for both modalities. Normalization of individual features on the electrophysiology side can therefore affect the R^2 calculations (the autoencoder may also act as a denoiser and impute genes that suffer from experimental drop-out which can affect the observed R^2 values as well.) Data were 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.

Cross-modal reconstruction of inhibitory cell-type marker genes.

We compiled a set of marker genes for inhibitory cell types from (according to Figs. 5e,f in ref.[8]) to showcase the cross-modal reconstruction ability of our model. Out of these, 38 genes were available in the set of 1,252 genes used as input and cross-modal reconstructions for these are shown in Supplementary Fig. 1. Note that six of these marker genes are also present in the list of neuropeptide precursor gene list shown in Fig. 2.

Supervised approach to identify consistent cell types based on reference taxonomy.

Our experiments suggest that one can identify approximately 33 clusters with in this dataset using a completely unsupervised approach. To argue that clusters obtained in this manner are more consistent across modalities, we compare the contingency matrix of Fig. 3b with one obtained with a supervised approach, utilizing the labels of the reference taxonomy (Supplementary Fig. 6). Accordingly, we first merged the reference hierarchy of Extended Data Fig. 1 to obtain 33 class labels. With these labels, we trained separate

QDA classifiers on three-dimensional representations z_t and z_e from the uncoupled (λ_{te} = 0) autoencoders. Note that this is equivalent to performing dimensionality reduction with separate autoencoders trained on the individual modalities. First, we show a representative contingency matrix for the ground truth transcriptomic labels, and the labels predicted by the classifier for test cells. The dominant diagonal in Supplementary Fig. 6a shows that the uncoupled three-dimensional representations capture the transcriptomic classification very well, whereas Supplementary Fig. 6b shows the contingency matrix for label predictions with z_t and z_e. We observe that for certain classes, classifiers trained on different modalities never lead to identical labels (zeros along the diagonal). We quantify this by defining a fraction of co-occupied labels, f_{co}; f_{co} = 0.88 for Fig. 3, whereas f_{co} = 0.42 ± 0.02 for the supervised approach devised here (mean ± s.d, sixfold cross-validation).

Application as a cross-modality translator for unimodal data.

We used two different published Patch-seq datasets with transcriptomic and electrophysiological profiles of GABAergic cells (Scala et al. 2019[22] and Scala et al. 2020[23]) to demonstrate the utility of coupled autoencoders to serve as translators for unimodal data. The Patch-seq data used in the main text is used as the reference dataset, and is referred to as the Gouwens et al. 2020 dataset. The Scala et al. 2019 dataset consists of 107 neurons, whereas Scala et al. 2020 dataset consists of 524 neurons sampled in mouse motor cortex.

Previous studies have shown that the cell type diversity of inhibitory neurons is essentially conserved across brain areas[8]; thus, although both the Scala et al. 2019 and the Scala et al. 2020 datasets profile neurons from mouse motor cortex, we can hope to use the mouse visual cortex Patch-seq dataset used in this study to serve as a meaningful reference. Nearly 5% of the genes that were used as input for the coupled autoencoders were either not measured, or were missing from the transcriptomic profiles of neurons in both these datasets. At the time of network training, we zeroed out the genes expression values at random both to mimic gene dropout[33,34] and to increase the robustness against non-identical input gene lists. We were therefore able to use the pre-trained network without additional training to make predictions with these datasets.

The Scala et al. 2020 dataset was obtained from the public repository related to this work at https://github.com/berenslab/mini-atlas. The dataset consists of gene expression profiles that were mapped to the reference taxonomy considered here. We relied on this mapping to select 524 cells that were confidently mapped to the inhibitory types that were well sampled in the Gouwens et al. dataset. In particular, we filtered out cells that were mapped to a single leaf node of the reference taxonomy with less than 80% confidence.

Results for inferring transcriptomic cell types and predicting electrophysiological features from gene expression with pre-trained coupled autoencoders are shown in Extended Data Figs. 8 and 9, and Supplementary Figs. 2 and 3 The electrophysiology measurement protocols in both these datasets differ from the one in the Allen Institute dataset. In particular, differences in the temperature and internal/external solutions with which experiments were conducted are expected to contribute to differences in estimated parameters across the datasets.

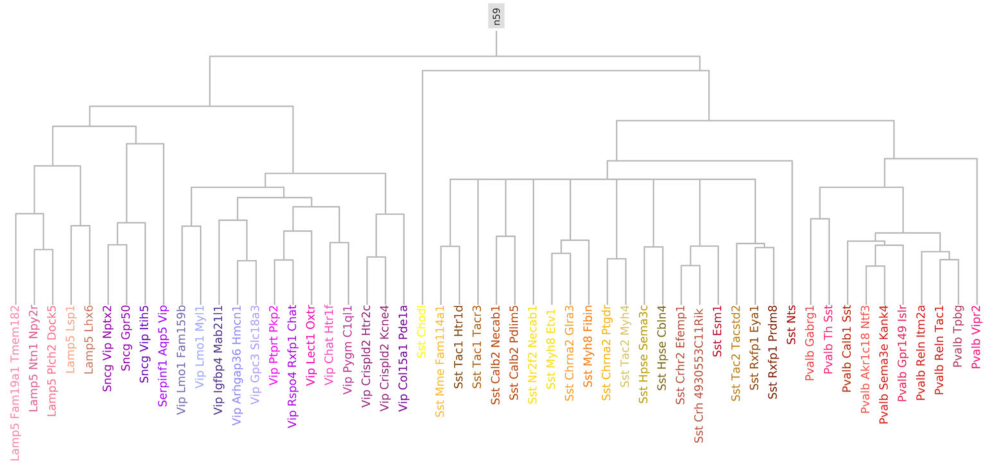
Data availability

The Patch-seq transcriptomic data are available at http://data.nemoarchive.org/other/grant/AIBS_patchseq/transcriptome/scell/SMARTseq/processed/analysis/20200611/, whereas the electrophysiological data are available at https://dandiarchive.org/dandiset/000020. For the Scala et al. 2019 dataset, the sequencing data are available under accession no. GSE134378, whereas the electrophysiological data are available at https://doi.org/10.5281/zenodo.3336165. The Scala et al. 2020 dataset was obtained from the public repository related to this work at https://github.com/berenslab/mini-atlas. Source Data are available with this paper.

Code availability

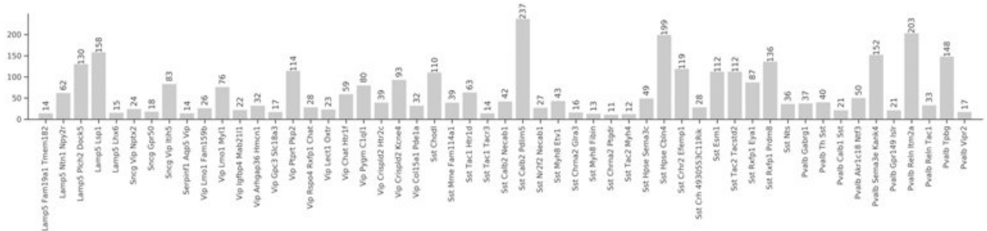
Code for the coupled autoencoder implementation and analysis are available at https://github.com/AllenInstitute/coupledAE-patchseq. An interactive version of the code base is provided in ref. [37].

Extended Data



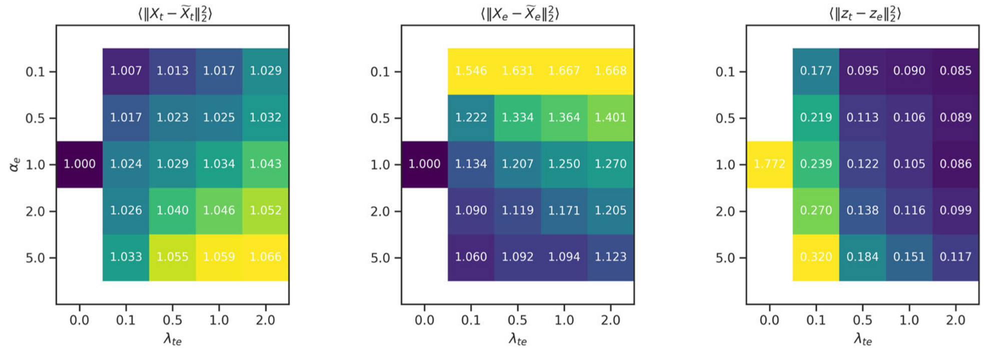
Extended Data Fig. 1 |. Reference taxonomy for well-represented GABAergic neurons.

Cells were mapped to the complete hierarchical classification tree for cortical cells with a marker gene based procedure. Here we show a subset of the full hierarchical tree, which consists of only those leaf nodes that are well-represented (n≥10) in the Patch-seq dataset. At the highest resolution, this tree consists of 53 cell type labels. The lowest resolution view consists of a single label (n59) which comprises of all GABAergic cortical neurons.



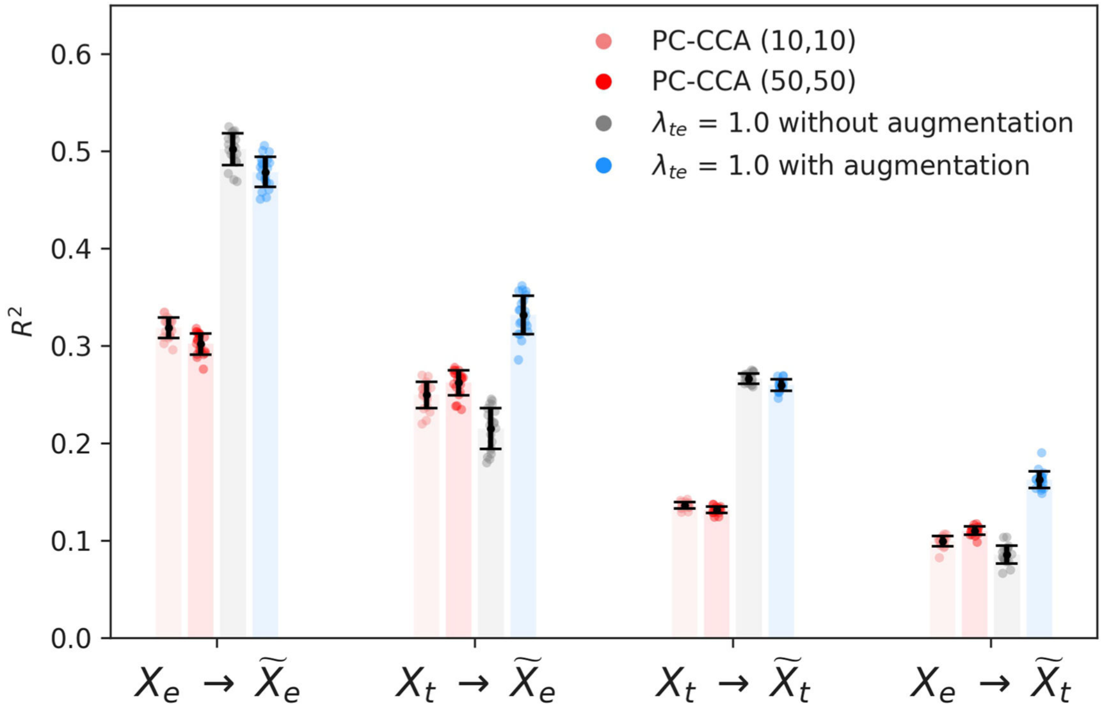
Extended Data Fig. 2 |. Cell type distribution.

The distribution of samples according to the reference hierarchy cell type label assignment. Types with less than 10 samples are not shown.



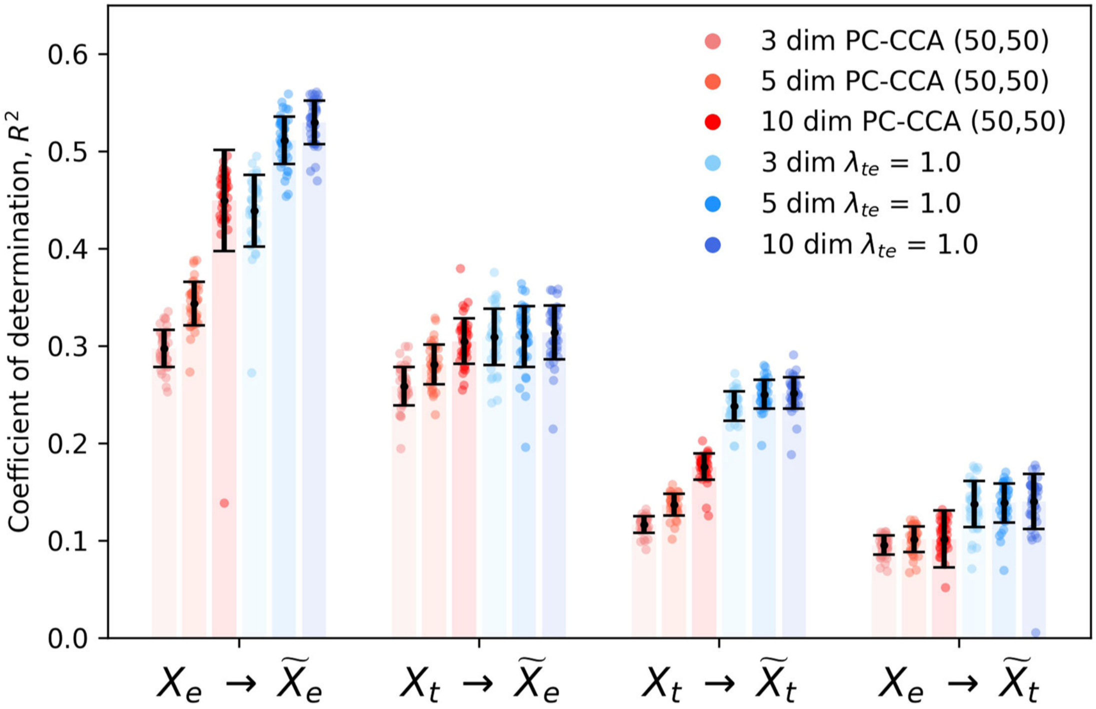
Extended Data Fig. 3 |. Hyper-parameter search.

(Left and center) Reconstruction errors relative to the value over uncoupled networks, and (Right) coupling error over different values for α_e and λ_{te} averages over validation sets. The value for α_t was set to 1.0 and representation dimensionality was set to 3 for these experiments. As coupling is increased, the reconstruction error increases illustrating the trade-off between coupling and reconstruction accuracy.



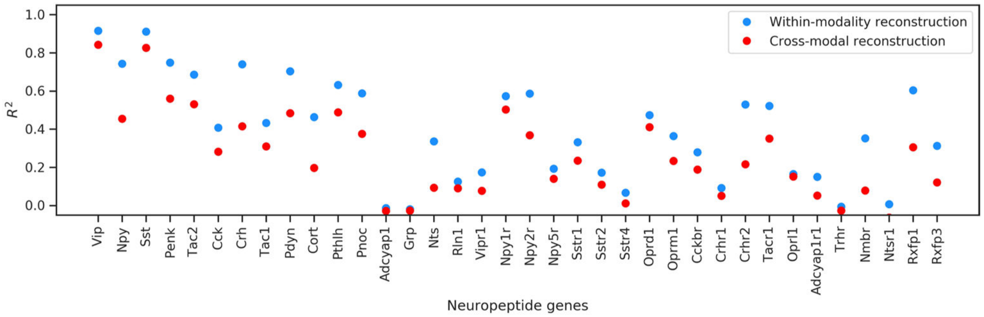
Extended Data Fig. 4 |. Decoder augmentation improves cross-modal prediction accuracy.

We use cross modal representations to augment the input for decoder subnetworks while training. Reconstruction performance as measured by the coefficient of determination (R^2) for linear baselines (PC-CCA), and coupled autoencoders with- and without- augmentation. Error bars show standard deviation over 20 cross validation folds.



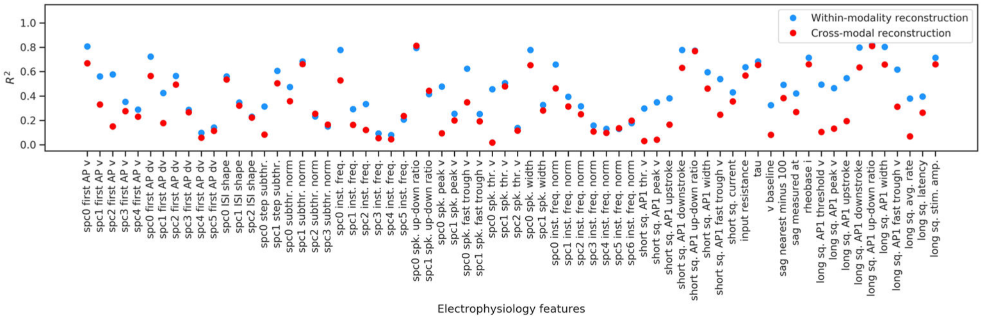
Extended Data Fig. 5 |. Effect of latent space dimensionality on reconstruction performance.

errors for coupled autoencoder and linear baseline for different latent space dimensionality dim ∈ {3, 5, 10}. Coupled autoencoders reconstruct the data more accurately than linear baselines ($p < 10^{-4}$, two-sided Wilcoxon signed-rank test). The only exception is for $X_t \rightarrow \widetilde{X}_e$ with dimensionality set to 10, where the null hypothesis cannot be rejected. We would like the dimensionality to be as low as possible for downstream tasks such as clustering and classification with limited data, and as high enough for good performance at tasks such as data imputation or cross-modal data prediction.



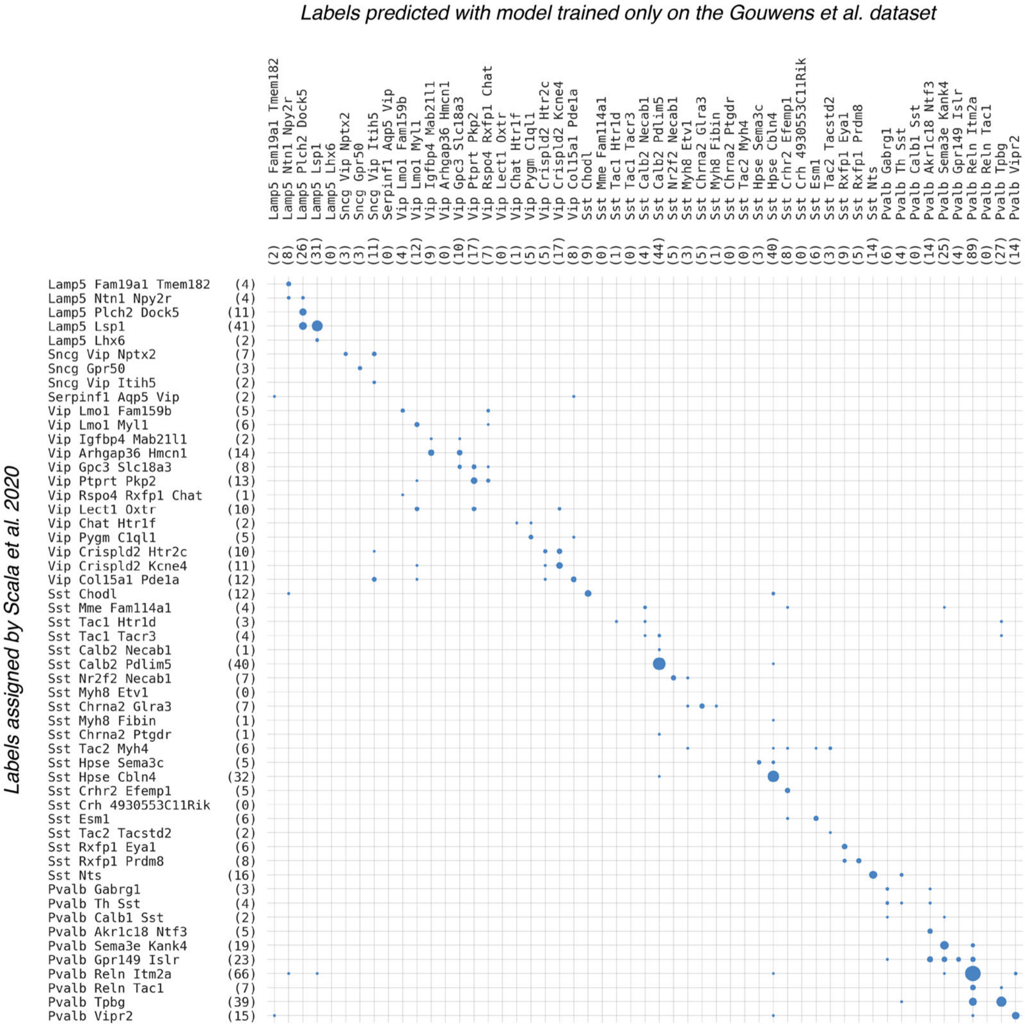
Extended Data Fig. 6 |. Reconstruction of gene expression using coordinated representations.

Within-modality reconstructions for individual genes are decoded from the coordinated λ_{te} = 1.0 representation z_t obtained for the transcriptomic data. Cross-modal reconstructions are obtained from the corresponding z_e, which is the representation for the electrophysiological data. The cross-modal reconstructions are comparable to within-modality reconstructions, and a majority of the neuropeptide precursor genes are reconstructed well, as suggested by the high coefficient of determination (R^2) values.reconstructed well, as suggested by the high coefficient of determination (R^2) values.



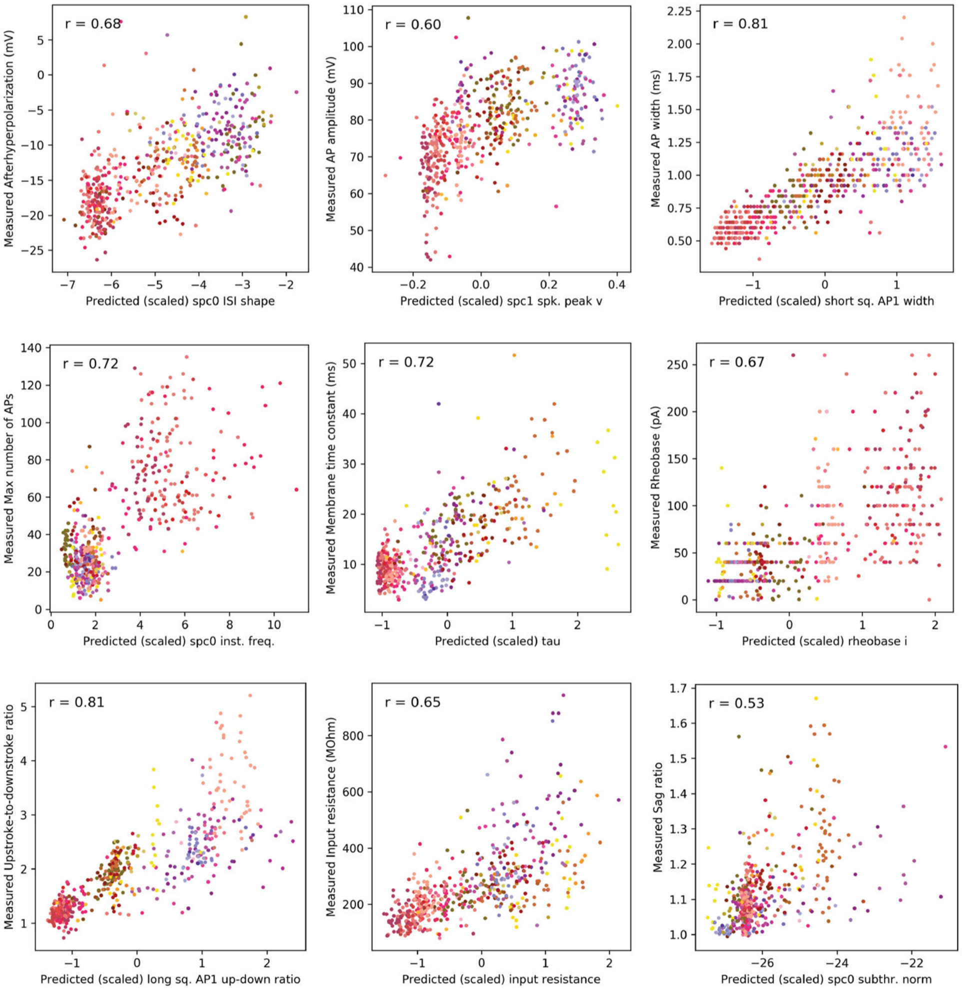
Extended Data Fig. 7 |. Reconstruction of electrophysiological features using coordinated representations.

The within-modality reconstructions for electrophysiological features are decoded from the coordinated $\lambda_{te} = 1.0$ representation z_e obtained for the electrophysiological data. Cross-modal reconstructions are obtained from the corresponding z_t, which is the representation for the transcriptomic data. Features that are reconstructed well in the within-modality case are analyzed in the context of transcriptomic cell types in the main text.



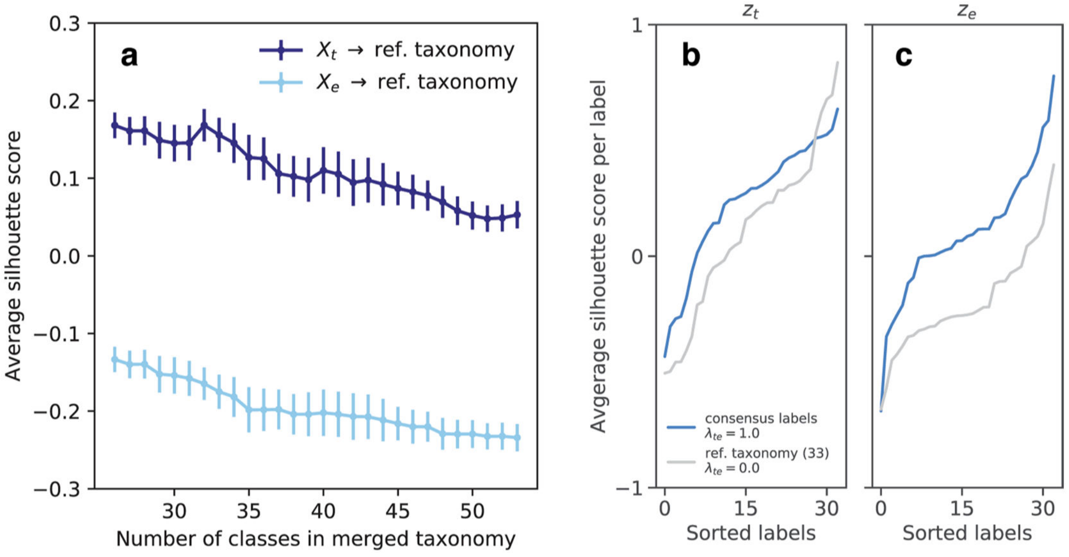
Extended Data Fig. 8 |. Predicting cell types based on gene expression.

Gene expression profiles of the 524 inhibitory neurons Scala et al. 2020 dataset were used to obtain 3-d representations without additional training of the coupled autoencoder trained on the Gouwens et al. dataset. QDA classifiers trained to predict cell types for the Gouwens et al. dataset were thereafter used to predict labels for the cells in the Scala et al. 2020 dataset. The contingency matrix comparing the predicted cell types and the cell types assigned by Scala et al. is shown. Overall accuracy of label prediction is 66%, with many inaccuracies being accounted for by closely related types.



Extended Data Fig. 9 |. Predicting electrophysiological properties from gene expression.

Gene expression profiles for 524 inhibitory neurons in the Scala et al. 2020 dataset were used as input for the coupled autoencoder that was trained only with the Gouwens et al. dataset. The electrophysiological measurements were not measured the same way in the two datasets; cross-modal setting only allows predictions for electrophysiological features of the Gouwens et al. dataset for cells in the Scala et al. dataset. There is a strong correlation (Pearson's *r* is shown on each plot) for many related measurements across the datasets. Cells are colored according to the cell type assignments of Scala et al. 2020, who mapped them to the same reference taxonomy that is used throughout this study.



Extended Data Fig. 10 |. Reference taxonomy labels do not partition the data well.

Average silhouette scores for test samples, for successive mergings of the reference taxonomy with uncoupled representations do not indicate any particularly favorable number of clusters. Error bars show mean ± SD over 5 best initializations (based on reconstruction accuracy) of single modality (uncoupled) autoencoders operating on X_t and X_e. Here the uncoupled representations z_t and z_e serve as low dimensional representations of the standalone data. The per-label silhouette score for the 33-merged reference taxonomy labels with uncoupled representations performs worse than consensus cluster labels on both, z_t (**b**) and z_e (**c**).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We wish to thank the Allen Institute for Brain Science founder, P. G. Allen, for his vision, encouragement and support. This work was supported by the NIH grant 1RF1MH123220-01.

References

1. Tremblay R, Lee S & Rudy B Gabaergic interneurons in the neocortex: from cellular properties to circuits. Neuron 91, 260–292 (2016). [PubMed: 27477017]
2. Zeng H & Sanes JR Neuronal cell-type classification: challenges, opportunities and the path forward. Nature Rev. Neurosci 18, 530 (2017). [PubMed: 28775344]
3. Paul A et al. Transcriptional architecture of synaptic communication delineates gabaergic neuron identity. Cell 171, 522–539 (2017). [PubMed: 28942923]
4. Huang ZJ & Paul A The diversity of gabaergic neurons and neural communication elements. Nat. Rev. Neurosci 20, 563–572 (2019). [PubMed: 31222186]
5. Ascoli GA et al. Petilla terminology: nomenclature of features of gabaergic interneurons of the cerebral cortex. Nat. Rev. Neurosci 9, 557 (2008). [PubMed: 18568015]
6. Berens P & Euler T Neuronal diversity in the retina. e-Neuroforum 23, 93–101 (2017).

7. Adkins RS et al. A multimodal cell census and atlas of the mammalian primary motor cortex. Preprint at https://www.biorxiv.org/content 10.1101/2020.10.19.343129v1 (2020).

8. Tasic B et al. Shared and distinct transcriptomic cell types across neocortical areas. Nature 563, 72–78 (2018). [PubMed: 30382198]

9. Zeisel A et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 347, 1138–1142 (2015). [PubMed: 25700174]

10. Chen KH et al. Spatially resolved, highly multiplexed RNA profiling in single cells. Science 348, aaa6090 (2015). [PubMed: 25858977]

11. Cadwell CR et al. Multimodal profiling of single-cell morphology, electrophysiology, and gene expression using Patch-seq. Nat Protoc. 12, 2531–2553 (2017). [PubMed: 29189773]

12. Somogyi P, Tamas G, Lujan R & Buhl EH Salient features of synaptic organisation in the cerebral cortex. Brain Res. Rev 26, 113–135 (1998). [PubMed: 9651498]

13. DeFelipe J et al. New insights into the classification and nomenclature of cortical gabaergic interneurons. Nat. Rev. Neurosci 14, 202–216 (2013). [PubMed: 23385869]

14. Gowens NW et al. Integrated morphoelectric and transcriptomic classification of cortical GABAergic cells. Cell 183, 935–953 (2020). [PubMed: 33186530]

15. Hastie T, Tibshirani R & Friedman J The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Springer, 2009).

16. Kobak D et al. Sparse reduced-rank regression for exploratory visualization of multimodal data sets. Preprint at https://www.biorxiv.org/content 10.1101/302208v2 (2019).

17. Gouwens NW et al. Classification of electrophysiological and morphological neuron types in the mouse visual cortex. Nat. Neurosci 22, 1182–1195 (2019). [PubMed: 31209381]

18. Stuart T et al. Comprehensive integration of single-cell data. Cell 177, 1888–1902 (2019). [PubMed: 31178118]

19. Smith SJ et al. Single-cell transcriptomic evidence for dense intracortical neuropeptide networks. eLife 8, e47889 (2019). [PubMed: 31710287]

20. Smith SJ, Hawrylycz M, Rossier J & Sümbül U New light on cortical neuropeptides and synaptic network plasticity. Curr. Opin. Neurobiol 63, 176–188 (2020). [PubMed: 32679509]

21. Földy C et al. Single-cell rnaseq reveals cell adhesion molecule profiles in electrophysiologically defined neurons. Proc. Natl Acad. Sci. USA 113, E5222–E5231 (2016). [PubMed: 27531958]

22. Scala F et al. Layer 4 of mouse neocortex differs in cell types and circuit organization between sensory areas. Nat. Commun 10, 4174 (2019). [PubMed: 31519874]

23. Scala F et al. Phenotypic variation of transcriptomic cell types in mouse motor cortex. Nature 10.1038/s41586-020-2907-3 (2020).

24. Harris KD et al. Classes and continua of hippocampal ca1 inhibitory neurons revealed by single-cell transcriptomics. PLoS Biol. 16, e2006387 (2018). [PubMed: 29912866]

25. Baltrušaitis T, Ahuja C & Morency L-P Multimodal machine learning: a survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell 41, 423–443 (2018). [PubMed: 29994351]

26. Li Y, Yang M & Zhang Z A survey of multi-view representation learning. IEEE Trans. Knowl. Data Eng 31, 1863–1883 (2018).

27. Wang K, Yin Q, Wang W, Wu S & Wang L A comprehensive survey on cross-modal retrieval. Preprint at https://arxiv.org/abs/1607.06215 (2016).

28. Andrew G, Arora R, Bilmes J & Livescu K Deep canonical correlation analysis. In International Conference on Machine Learning 1247–1255 (JMLR, 2013).

29. Wang W, Arora R, Livescu K & Bilmes J On deep multi-view representation learning. In International Conference on Machine Learning 1083–1092 (JMLR, 2015).

30. Feng F, Wang X & Li R Cross-modal retrieval with correspondence autoencoder. In Proc. 22nd ACM International Conference on Multimedia 7–16 (ACM, 2014).

31. Gala R et al. A coupled autoencoder approach for multi-modal analysis of cell types. In Advances in Neural Information Processing Systems 9263–9272 (Curran Associates, 2019).

32. Ioffe S & Szegedy C Batch normalization: accelerating deep network training by reducing internal covariate shift. In Proc. 32nd International Conference on International Conference on Machine Learning Vol. 37 448–456 (JMLR, 2015).

33. Kharchenko PV, Silberstein L & Scadden DT Bayesian approach to single-cell differential expression analysis. Nat. Methods 11, 740 (2014). [PubMed: 24836921]

34. Srivastava N, Hinton G, Krizhevsky A, Sutskever I & Salakhutdinov R Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res 15, 1929–1958 (2014).

35. Freedman D & Diaconis P On the histogram as a density estimator: L2 theory. Zeitschrift Wahrsch. Verwandte Gebiete 57, 453–476 (1981).

36. Bakken TE et al. Single- nucleus and single-cell transcriptomes compared in matched cortical cell types. PLoS One 13, e0209648 (2018). [PubMed: 30586455]

37. Gala R et al. Consistent Cross-modal Identification of Cortical Neurons with Coupled Autoencoders (CodeOcean, 2020); 10.24433/CO.4098627.v1

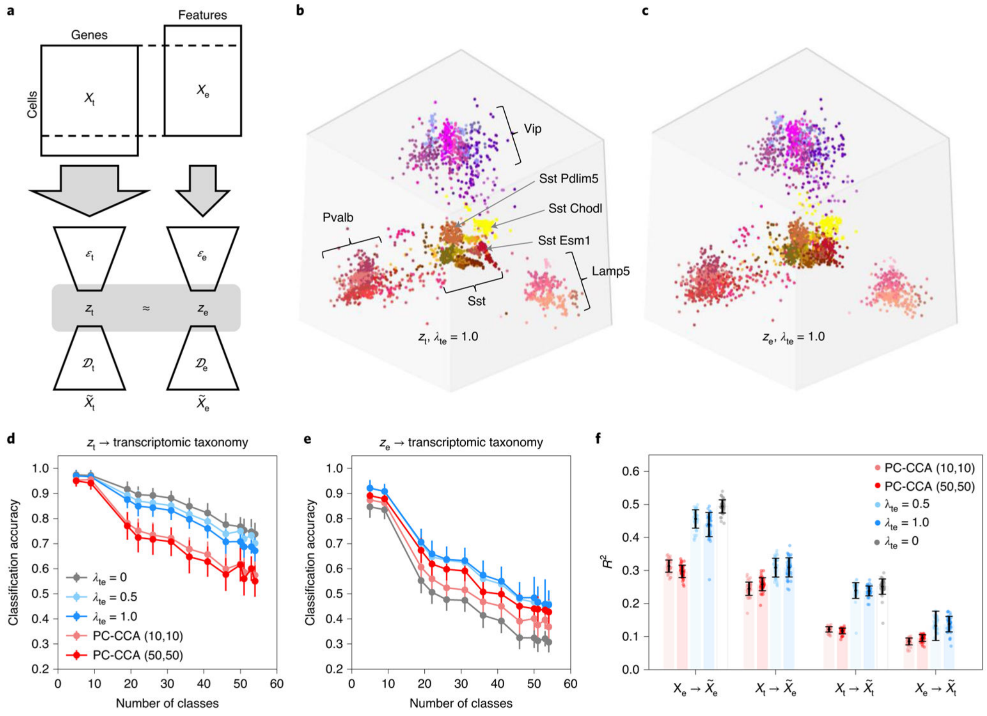


Fig. 1 |. Coordinated representations of transcriptomic and electrophysiological profiles with coupled autoencoders.

a, A schematic showing the coupled autoencoder architecture for Patch-seq data. Encoders ($\mathscr{E}$) compress input data (X) into low-dimensional representations (z), whereas decoders ($\mathscr{D}$) reconstruct data ($\tilde{X}$) from representations. The coupling penalty in the loss function encourages representations to be similar across the transcriptomic (t) and electrophysiology (e) modalities. **b**,**c**, Three-dimensional coordinated representations of the transcriptomic (**b**) and electrophysiological (**c**) datasets. Each point represents a single cell, which is colored by its cell-type membership according to the reference transcriptomic taxonomy. **d**,**e**, Performance on supervised cell-type classification tasks at different resolutions of the reference taxonomy. Classification with QDA is performed using three-dimensional representations of the transcriptomic (**d**) and electrophysiological (**e**) datasets obtained with coupled autoencoders and with linear methods. **f**, Performance on within-modality ($X_e \to \tilde{X}_e$ and $X_t \to \tilde{X}_t$) and cross-modality ($X_t \to \tilde{X}_e$ and $X_e \to \tilde{X}_t$) reconstruction tasks. Uncoupled representations are not suitable for cross-modal tasks. Error bars show mean ± s.d. over 43-fold cross-validation for panels **d**–**f**. Note that there are 1,252 genes versus 68 electrophysiology features in the dataset over which **f** is calculated.

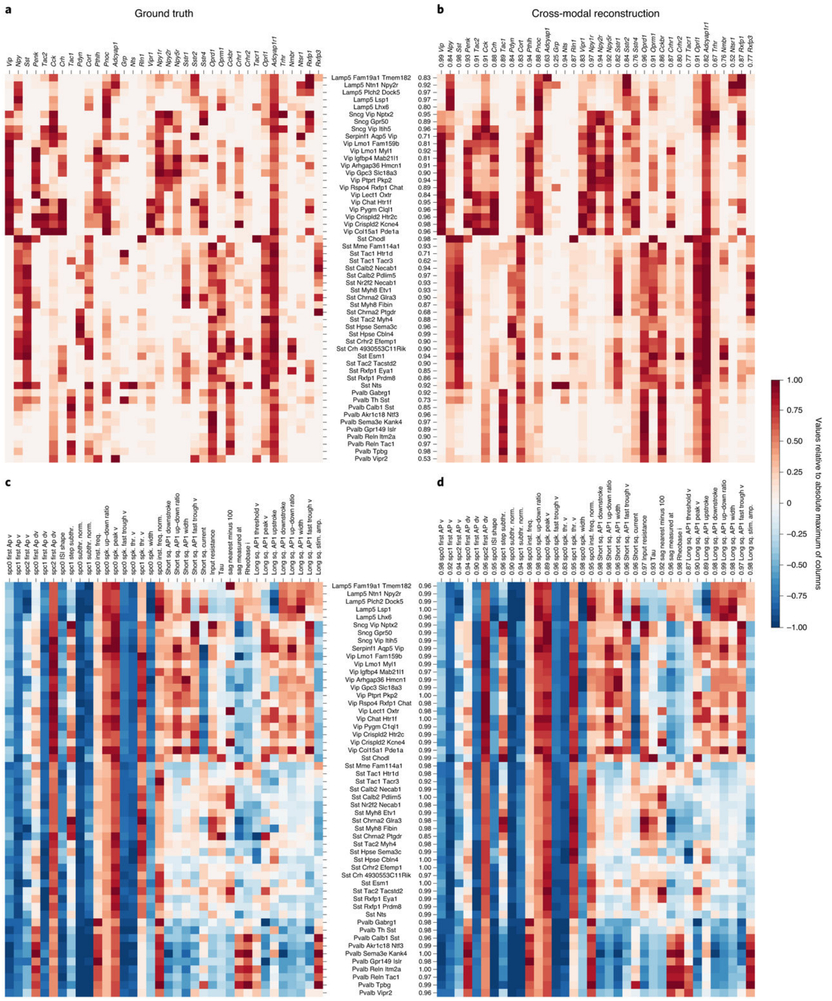


Fig. 2 |. Cross-modal reconstructions capture cell-type-specific gene expression patterns and electrophysiological features.

a, Gene expression levels averaged over samples reference taxonomy cell types, normalized per gene by the maximum value of each column. **b**, The cell-type specificity of different genes is captured well by cross-modal prediction of gene expression profiles from electrophysiological features (Pearson's $r = 0.89 \pm 0.10$, mean ± s.d. over cell types). **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 (Pearson's $r = 0.98 \pm 0.02$, mean ± s.d. over cell types). Cell-type and feature-wise fidelity of reconstructions are quantified with Pearson's r for each row and column in **b** and **d** as compared to ground truth in **a** and **c**.

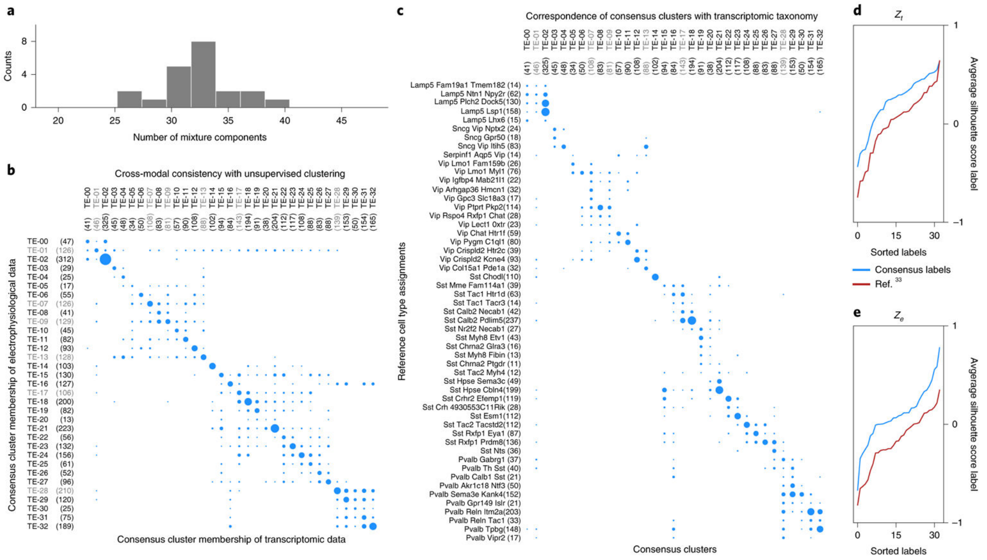


Fig. 3 |. Deriving a consensus cell-type clustering.

a, Unsupervised clustering using Gaussian Mixtures on the coordinated representation z_t and BIC-based model selection suggests 33.0 consensus clusters (32.19 ± 3.16, mean ± s.d.). **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 with unsupervised cluster assignments show that these unsupervised clusters have substantial overlap with known transcriptomic cell types. The number of cells for each label are indicated within parentheses next to the label, and area of the dots is proportional to number cells in the scatter plots of **b** and **c**. **d**, Consensus clusters are compared with an equal number of cell classes obtained by merging the reference hierarchical taxonomy, using silhouette analysis based on coupled representations. Average per-cluster silhouette values for test cells are larger for consensus cluster labels. Clusters for which the silhouette score is less than zero in **d** and **e** are grayed out in panels **b** and **c**.