Spatial Cell Type Enrichment Predicts Mouse Brain Connectivity

Shenghuan Sun[1], Justin Torok[1], Christopher Mezias[2], Daren Ma[1], and Ashish Raj[1,*]

[1]Department of Radiology, University of California, San Francisco – San Francisco, CA, United States
[2]Cold Spring Harbor Laboratory – Cold Spring Harbor, NY, United States
[*]Correspondence should be directed to Ashish Raj (ashish.raj@ucsf.edu)

ABSTRACT

A fundamental question in neuroscience is whether and how the brain's molecular, cellular and cytoarchitectonic properties govern its inter-regional structural connectivity. Several recent studies have leveraged spatial transcriptomic data to infer the mouse mesoscale connectome with high accuracy, indicating that there is a strong biological link between whole-brain gene expression and inter-regional connectivity. However, despite the fact that cells are the fundamental biological units forming and maintaining white-matter tracts, the question of whether and which regional cell-type distributions determine patterns of structural connectivity, has received scant attention. Such a link has been previously postulated, but the lack of a comprehensive spatial atlas of cell types in the mouse brain has prevented its direct investigation. Here, we utilize recent advances in whole-brain mapping of a variety of neuronal and non-neuronal subtypes using the Matrix Inversion and Subset Selection (MISS) algorithm to model inter-regional connectivity as a function of regional cell-type composition with the random forest algorithm. The machine learning model was surprisingly accurate in predicting brain-wide connectivity from cell type densities - quantitatively demonstrating, for the first time, the role that cell types play in governing connectivity. Prediction accuracy was similar across two independently sampled sets of cell types, indicating that cell-type distributions are a robust predictor of inter-regional connectivity. To uncover the individual actors in this process we undertook a thorough feature importance analysis, with surprising outcomes. We found that the single most salient cell type for reconstructing the whole-brain connectome are oligodendrocytes, the brain's myelin and fiber maintenance cells. Indeed, non-neuronal cells dominate neuronal cells as predictors of connectivity and of canonical connectome graph theoretic metrics of centrality and network "hubness", somewhat unexpected findings that merit mechanistic explorations. We found evidence of a strong distance-dependency in the cell-connectivity relationship, with layer-specific excitatory neurons contributing most for predicting long-range connectivity, while vascular and astroglia are the most salient for short-range connections. Taken together, our results provide the first demonstration of a link between whole-brain cell-type distributions and the structural connections between regions, and provide a road map for examining this relationship in other species, including humans.

Introduction

The structural connectome, which represents the density of physical projections between each pair of brain regions and is measured by such techniques as viral tracing and diffusion tensor imaging, is a coarse wiring diagram of the central nervous system[1–4]. Complex molecular processes during embryonic development encourage the formation of connections between brain regions, and later postnatal pruning results in structural connectomes with a remarkable degree of conservation between healthy individuals. A rigorous exploration of the relationships between gene expression, cell type distributions, and inter-regional connectivity can deepen our understanding of how brain circuits mature during the development of the central nervous system and how they are disrupted in neurodegenerative diseases, among other areas of inquiry. Accordingly, there is strong interest in gaining an understanding of how the gene expression and the cell type composition of each region relate to connectivity[5,6].

In particular, the correlation between regional gene expression and connectivity is well established in mice[5,7–9] and humans[10–12], but the methods used to determine this association are mainly correlative or analytic. Correlation or regression with high-dimensional input feature spaces carries a risk of overfitting, and, as a result, often fails to generalize to novel data[13]. As an alternative approach, Ji, *et al.*[14] applied random forest methods to predict the brain connectivity using gene expression with high accuracy. However, Ji, *et al.*[14] only focused on predicting the binary existence or absence of a connection and did not attempt to predict the amount of connectivity density. Additionally, several groups[5,14] report that connected regions tend to have higher correlated gene expression patterns than regions that are not, which naturally raises the question of whether the connected brain regions share common cell types. A step in this direction was taken by Huang *et al.*, who demonstrated BRICseq, a powerful technique capable of mapping individual axonal projections along with the neuronal subtypes to which they belong[15]. However, their methodology has not yet been scaled up to produce a dataset of comparable spatial coverage to

the Allen Mouse Brain Connectivity Atlas (AMBCA)[2], which is perhaps the most thorough mesoscale connectome currently available. Therefore, it is not yet clear how distributions of different types of cells - the fundamental units of connectivity - relate to the whole-brain connectome, nor have any unbiased, data-driven methods been applied to attempt to reconstruct the mouse connectome from regional cell type densities. Although the success of prior studies in using gene expression-based markers to predict connectivity suggests that cell type distributions will also be predictive of the connectome, the paucity of available whole-brain cell type distributions has made it difficult to test the hypothesis. Indeed, before the advent of spatial transcriptomics and single cell gene profiling the question would have been impossible to answer quantitatively on the whole brain level.

Here, we take advantage of these emerging technologies to develop a comprehensive data-driven computational machinery needed to address this question. We first implemented an algorithm to produce regional cell type enrichment from spatially resolved gene expression data, following a novel method we have recently developed called Matrix Inversion and Subset Selection (MISS)[16]. This method is essentially a cell type deconvolution algorithm that was shown to faithfully reproduce cell type distributions in the mouse brain using Allen Gene Expression Atlas (AGEA)[17] and publicly available single-cell RNA sequencing data[18,19]. Then, using inferred cell type enrichment distributions as input features, we applied a number of machine learning methods to reconstruct the structural connectome as given by the AMBCA[2]. Among all the models tested, the random forest (RF) algorithm outperformed other approaches at predicting both the presence or absence of a connection between any given region pair as well as the actual connectivity density values.

We were able to predict the structural connectome with a surprisingly high level of accuracy, despite that the fact that the construction of fiber connectivity is a highly complex and iterative biological process with many determinants not strictly captured by constituent cell types. We replicated our findings with a second, different set of cell type distributions inferred by MISS. Despite the two datasets having a widely different number of individual cell types, both achieved almost identical performance at the connectivity prediction task, indicating that our approach is not an artifact of a particular input feature set. Our results demonstrate quantitatively, for the first time, that regional cell type distributions underpin most of the variance in inter-regional connectivity

To uncover the individual actors in this process we undertook a thorough feature importance analysis, with both confirmatory and surprising outcomes. Strikingly, feature importance analysis for both sets implicated oligodendrocytes as the most important cell type feature for recreating whole brain inter-regional connectivity. Oligodendrocytes are the brain's myelin and fiber maintenance cells; their role in predicting connectivity is not unexpected, but their prominence in this role has not received adequate attention. A deeper dive also uncovered that non-neuronal cells generally dominate neuronal cells as predictors of connectivity, indicating a strong underlying relationship between these non-neuronal cell types and the connectome. The non-neuronal cells' densities are significantly associated with connectome graph metrics of centrality, including degree centrality and eigenvector centrality, revealing the cellular correlates of canonical hubs well known in brain graph theory. Additionally, we identified the distance between two regions as a important predictor besides the cell type enrichment information. Related to this, we found evidence of a strong distance-dependency in the cell-connectivity relationship, with layer-specific excitatory neurons contributing most for predicting long-range connectivity, while vascular and astroglia are the most salient for short-range connections. Indeed, the cell types necessary for reconstructing long-range connections are generally different from those most useful for predicting local connectivity, suggesting that these may be maintained by distinct biological pathways. Together, our findings point to hitherto under-explored role of specific cell types that play outsize roles in forming and/or maintaining connections.

Results

Overview of the study pipeline

A schematic of the analytic pipeline is displayed in **Figure 1**. We computed the regional densities for 25 neuronal and non-neuronal cell types from publicly available single-cell RNA-seq[18,20] and *in situ* hybridization[17] data using the Matrix Inversion and Subset Selection (MISS) algorithm[16] (**Figure 1-i**). We normalized these raw MISS-inferred densities to create enrichment scores to prevent the scale of these features from artificially influencing the machine learning algorithms' outputs (see **Methods**). The connectivity data we attempt to reconstruct was derived from the Allen Mouse Brain Connectivity Atlas (http://connectivity.brain-map.org)[2], which we also normalize according to standard procedures (**Figure 1-ii**; see also **Methods**). Finally, several machine learning methods were implemented to infer the whole-brain connectome from the regional cell type enrichment scores, which we evaluated quantitatively (**Figure 1-iii**).

Predicting the existence or absence of connectivity

The first question we were interested in was whether regional cell type enrichment features can be used to predict the existence or absence of connectivity between any given pair of regions. We considered the binary classification task of determining which pairs of regions connect and the prediction of connectivity density as separate because the underlying biological difference

between zero connectivity and non-zero connectivity is qualitatively different from any differences in degree of connectivity between region-pairs (see **Methods**). **Figure 2A** shows the proportions of zero and non-zero values within the ABMCA, indicating that the mouse brain connectome is approximately 64% sparse. Before using the machine learning methods, we asked whether common unsupervised clustering methods such as Principal Component Analysis (PCA) or t-Distributed Stochastic Neighbor Embedding (t-SNE) could accurately predict the presence or absence of a connection between any given region pair. However, neither approach could distinguish region-pairs that form connections from those that do not (**Figure 2B**), demonstrating that these common clustering methods might not be suitable for this task. However, we found that the random forest (RF) algorithm produced excellent classification results (**Figure 2C,D**; **Table 1**; **S. Data Table 2**)[21]. The confusion matrix in **Figure 2C** shows that the RF model predicts the existence of connectivity between pairs of regions with a accuracy of 0.79. We also found that the AUROC (Area Under the Receiving Operator Characteristic) and the AUPR (Area Under Precision-Recall curve) values for the RF algorithm were 0.85 and 0.77, respectively (**Figure 2D**). Our results using cell types parallel those using gene expression[14] and indicate that regional cell type enrichment profiles are predictive of the existence or absence of connectivity between region-pairs in the mouse brain.

Predicting connectivity density

We next turned to the task of predicting the connectivity density[2,15,22,23], which is a measure proportional to the number of axonal tracts per unit volume between any region pair. We first examined whether region-pairs with similar cell type compositions were likely to be more densely connected. **Figures 3A** and **B** (left panel) depict heat maps of the regional cross-correlation matrix with respect to cell type enrichment scores and the mouse connectome, respectively. While there is a degree of visual similarity, the two measures are only weakly correlated (Pearson's R = 0.14, Spearman's ρ = 0.08; **Figure 3C**). This agrees with previous work suggesting that coupled regions tended to have higher levels of gene expression similarity[5,14]. We conclude that inter-regional similarity in cell type enrichment profile might relate to connectivity density, but it is not sufficient to faithfully reconstruct the whole brain connectome.

Given that the connectivity density distribution is mostly comprised of very small values with a number of prominent outliers (**S. Figure 1**), we hypothesized that a nonlinear, data-driven approach could be more feasible for solving this problem. Similar to the prior binary classification task, we found that the RF model using regional cell type enrichment scores as features recreated connectivity with a high degree of accuracy (Adjusted R^2 = 0.59, Root-mean-square deviation = 0.60, 10-fold cross-validation; **Table 1**; **S. Data Table 3**). Heatmap renderings emphasize that the RF model was able to qualitatively reconstruct the brain connectome pattern as well, given the high degree of visual similarity between the cell-type-predicted connectivity and the ground truth (**Figure 3B**). We also visualize the strong agreement between the RF-predicted connectivity densities and the ground truth across the entire brain in **Figure 3D**, which exhibits a Pearson's correlation of 0.77.

In the above analyses, we separated the tasks of predicting the presence or absence of connectivity (binary classification) and predicting the density of connections among connected region-pairs (regression). From both biological and machine learning perspectives, these are distinct questions and therefore we chose to address them individually. Nevertheless, we also implemented a RF algorithm to predict connectivity density in the AMBCA without first removing unconnected region-pairs (**S. Figure 3**). As expected, we found that agreement was not as strong between ground truth and predicted connectivity when the zeroes were not first filtered out; however, the adjusted R^2 of 0.394 reflects good model performance.

Several other common machine learning algorithms were also implemented to reconstruct both the binary connectome and predict connectivity density, which, however, fail to achieve the superior performance over random forest(**S. Data Table 2, 3**).

Confirmation with an independent cell type dataset

We tested whether the random forest algorithm could also recreate whole-brain connectivity using an independently curated collection of cell types to form the input feature space. For this purpose we used MISS-inferred distributions of the scRNAseq dataset from Zeisel, *et al.*, which sampled a more diverse set of 200 cell types throughout the entire mouse brain[16,19]. We again created the heat maps of the regional cross-correlation matrix with respect to cell type enrichment. The two measures, again, are weakly correlated (Pearson's R = 0.24, Spearman's ρ = 0.13; **Figure 4A**, **S. Figure 4**). Notably, the two cell type similarity matrices created with 25 and 200 features, respectively, are strongly correlated with each other (Pearson's R = 0.79, p = 0.0; **S. Figure 4**), which we expected given the reliability of the MISS algorithm[16].

When we used this more expansive set of cell types, we were also able to produce an accurate recreation of the binarized connectome with this dataset (**Figure 4C**, **Table1**), which produced AUROC and AUPR scores of 0.87 and 0.80, respectively, following 10-fold cross-validation. This represents a modest improvement over the results with the 25-feature dataset (AUROC = 0.85, AUPR = 0.77; **Figure 3D**, **Table 1**). Similarly, the machine learning models were also successful in predicting the connectivity density between any given pair of regions with the Zeisel, *et al.* dataset, although they again only modestly outperformed the 25-type Tasic, *et al.* dataset (**Figure 4D**, **Table1**, **S. Data Tables 2** and **3**). Additionally, the high-level nonlinear models: RF, MLP and ExtraTrees all accurately recreated the connectome and outperformed models with low level non-linearity, such as Ridge and LASSO, reinforcing that nonlinear methods are better suited for reconstructing connectivity.

Feature importance analysis to identify key cellular mediators of connectivity

We next asked which cell types contribute the most to predictions of inter-regional connectivity. Unlike other machine learning models, which can give outputs whose dependencies are difficult to discern, RF models are amenable to *feature importance analysis*[21,24]. In **Figure 5A** and **B**, we show the distributions of feature importance scores across the 10 model iterations and cell-type features comprising cell "supertypes" in the Tasic, *et al.* and Zeisel, *et al.* taxonomies, respectively (refer to **S. Data Tables 4-7** for the list of cell-type names and the supertypes to which they belong). Feature importance analysis of all cell types in the random forest model indicates that oligodendrocytes (Oligo) are the most important contributors to connectivity at the whole-connectome level for both the Tasic, *et al.* and Zeisel, *et al.* datasets (**Figure 5A,B**). Oligodendrocytes are also the most highly predictive feature when analyzing the feature importance for the binary classification task of sorting region-pairs into connected and unconnected sets (**S. Figure 2**). More generally, the non-neuronal supertypes are more salient in the RF models than neuronal supertypes, with the exception of the "Chol Mono" supertype in Zeisel, *et al.*, which contains several cholinergic and monoaminergic cell types. Biologically, oligodendrocytes produce the myelin sheath insulating axons[25,26], and vascular cells (Vasc) construct both the blood-brain barrier, which protects the vulnerable central nervous system (CNS), and help provide neuronal cells with nutrients, energy, and oxygen[27–31]. Astrocytes (Astro)[32] and immune cells (Immune)[32,33] are also known to be critical for regulating neuron growth and pruning. The apparent consistency of these feature importance results between the two independently curated scRNAseq datasets suggests a true biological connection between these non-neuronal support cells and connectivity at a whole-brain level.

Selected cell types are associated with graph centrality metrics

To explore the underpinnings of the feature importance results for the non-neuronal cell types, we looked at the association between their distributions from Tasic, *et al.* and node-level graph metrics of the mouse connectome computed using `NetworkX` python package[34]. (**Figure 5C,D**; **S. Figure 5**). We highlight two metrics that convey information about the centrality and hub-tendency of each node in the connectome graph: connectivity strength/degree (sum of in- and out-degree) and eigenvector centrality, since we expect that the most central nodes are enriched in the cell types highlighted by above feature importance analysis. Interestingly, three out of the four non-neuronal cell types have significant correlations with degree, and all four have significant correlations with eigenvector centrality. The strongest correlations were observed using endothelial cells for both centrality measures, both of which were negative (**Figure 5C,D**). Voxelwise visualizations of the distributions of the Tasic, *et al.* oligodendrocytes, endothelial cells, astrocytes, and immune cells are shown in **Figure 5E**. Taken together, these results suggest that no single cell type predicts connectivity; instead all cell types must act in concert to deliver the predictive power shown in previous figures. However, non-neuronal cell types exhibit outsize importance in determining inter-regional connectivity density.

The effect of inter-regional distance on predicting connectivity density

Although adult cell-type distributions are highly informative for reconstructing the mouse connectome, the unexplained variance in the data likely comes from other biological factors. For instance, we found that there is a strong inverse relationship between inter-regional center-to-center distance and connectivity density (Spearman's ρ = -0.46; **Figure 6A**), indicating that there is a bias towards short-range connections in the mouse brain. Further, including inter-regional distance as a predictor along with the cell-type distributions produces RF models with higher R^2 values ($\Delta R^2 = 0.10$, p-value = 1.6×10^{-4}; **Figure 6B**). These results indicate that inter-regional distance contributes information that is at least partly orthogonal to that contributed by regional cell-type composition.

Consequently, we were interested in whether there was a distance dependence to cell-type feature importance, as has been suggested previously[35,36]. We therefore trained the RF algorithm on the upper and lower quartiles of connections by distance separately and determined the feature importance scores per cell supertype as above **Figure 6C-F**. The RF models achieved similar fits regardless of distance bin (R^2 = 0.62 and 0.57 for short-range and long-range connectivity, respectively) and performed comparably well to the model of whole-brain connectivity (**Table 1**). However, clear differences emerge at the level of feature importance between short-range and long-range connectivity. Although oligodendrocytes distributions from the Tasic, *et al.* and Zeisel, *et al.* datasets are not the strongest contributors to the RF model of short-range connectivity as they were for whole-brain connectivity, they remain among the top features, and generally non-neuronal cells have stronger feature importance scores than neurons, as above (**Figure 6C,E**). In particular, vascular cells (Vasc) and immune cells (Immune) exhibit the strongest contributions to short-range connectivity for the Tasic, *et al.* and Zeisel, *et al.* datasets, respectively. Of the neuronal supertypes, forebrain glutamatergic neurons (Neo Glu, Thal Glu, Hip Neo Glu) have particularly weak feature importance scores. Interestingly, this trend is reversed for reconstructing long-range connectivity: for both datasets, we found that these three neuronal cell-type distributions were consistently among the most salient features (**Figure 6D,F**). Additionally, the single-strongest feature among the Zeisel, *et al.* supertypes was striatal medium spiny neurons (MSN), which are unique to that dataset (**Figure 6F**). We summarize these results in **Figure 6G**, which shows that, for both the Tasic, *et al.* and Zeisel, *et al.* datasets: 1) non-neuronal cell types, and in particular vascular and immune cells, contribute predominantly to predicting

short-range connectivity as opposed to long-range connectivity; and 2) telencephalic glutamatergic neurons contribute little to models of short-range connectivity, but they are overrepresented among types that predict long-range connectivity. In short, while cell-type-based RF models can reconstruct short-range and long-range connectivity with a similar degree of accuracy as the whole-brain connectome, the saliency of the cell-type features markedly differs between these models.

Neuronal contributions to long-range connectivity

To explore some of the relationships between cell-type distributions and connectivity qualitatively, we show the distributions of Hip Neo Glu and MSN, the two supertypes from the Zeisel, *et al.* dataset with the highest average feature importance for predicting long-range connectivity (**Figure 7A,B**). The Hip Neo Glu class comprises twenty-four individual cell types, all of which are excitatory and located within neocortical and hippocampal regions, and the MSN class comprises six types of striatal medium spiny neuron. As expected based on their taxonomy, Hip Neo Glu cells are confined to the neocortex and hippocampus, while MSN cells are entirely within the striatum. Given the high degree of regional specificity of these cell-type classes, we also show the strongest long-range connections to and from the neocortex (**Figure 7C**) and the striatum (**Figure 7D**). More specifically, for the neocortex, these include projections to hindbrain nuclei and contralateral neocortical-neocortical connections (**Figure 7C**). The main long-range projections from the striatum originate in the olfactory tubercle and terminate in the periacqueductal gray of the midbrain, while it receives its strongest long-range inputs primarily from contralateral neocortical regions (**Figure 7D**). In this way, we can link the anatomical distributions of cell types to specific subsets of inter-regional connections.

Discussion

Summary of key results

Our results constitute one of the first applications of data-driven machine learning models for reconstructing whole brain inter-regional connectivity using spatial cell type enrichment distributions. With only 25 cell-type features[16,18], we were not only able to predict the existence or absence of connectivity (**Figure 2**) but also the density of connectivity between any given pair of regions (**Figure 3**) to a surprising level of accuracy. These results were also replicated using a different, independent set of cell type enrichment features[16,19], indicating that our performance is not an artifact of the particular input feature space chosen (**Figure 4**). *Post hoc* feature importance analyses implicated oligodendrocytes as especially critical in correctly recreating the whole brain connectome (**Figure 5**). These and other non-neuronal cell types showed a strong spatial association with topological measures of the connectome, such as node degree and eigenvector centrality. We found inter-regional distance as an important predictor of the density of brain connectivity (**Figure 6**). When feature importance was evaluated separately for short-range versus long-range connections, we found that telencephalic glutamatergic neurons appear to be far more important for recreating long-range connectivity than for short-range connectivity, while non-neuronal cell types were more important for recreating short-range connectivity. We discuss below the implications of our findings, some confirmatory and some unexpected, in the context of current literature.

Predicting binary as well as weighted connectomes

We divided our ML prediction tasks by separately predicting the absence or presence of a connection and the connectivity density between any given region pair for several reasons. First, the connectivity data are quite sparse (36% nonzero region-pairs), which can significantly impact the the ability of the model to generalize. Second, a zero connectivity density value might not necessarily mean there is no connectivity between two regions at all; rather, it might only mean the intensity was not able to pass the threshold of observability imposed by the mesoscale connectome methodology[2]. Third, from a biological perspective, connectivity density values are determined by the the amount of axonal projections from one region to another per unit volume, while the existence or absence of connectivity is partially regulated by synaptic pruning[37,38], a process that is largely supported by non-neuronal cell types during brain development[38,39]. Nevertheless, when we attempted to predict connectivity density for the whole connectome (including region-pairs with zero connectivity density), the RF model exhibited strong agreement with the ground-truth connectome, although not nearly as high as that with the zero values removed (**S. Figure 3**).

Model performance is replicated across two different scRNAseq datasets

We were able to replicate the results of our primary dataset - the 25-type Tasic, *et al.* dataset[18] - using a separate, 200-type dataset from Zeisel, *et al.*[16,19] (**Figures 3** and **4**; see also **Methods**). Interestingly, we found that the Zeisel, *et al.* dataset performed only modestly better despite containing a far more diverse array of cell types sampled from a more comprehensive set of brain regions. One possibility is that, because training accuracy is close to 1 even for the Tasic, *et al.* dataset, there is a limit to how well cell type features in the test set can reconstruct connectivity using machine learning. Fortunately, this limit occurs at a high level with excellent predictive power for both datasets. It is possible that a subset of the Zeisel, *et al.* cell types might outperform the 25 cell types from Tasic, *et al.*, but the current study design is not well-suited for exploring all

combinatorial possibilities. Alternatively, it may be that the 25 cell types inferred from the Tasic, *et al.* dataset, despite the fact that they only represent a subset of mouse neuronal diversity, provide close to maximal information content for reconstructing the brain connectivity. Nevertheless, that we were able to create models with high predictive accuracy with two sets of cell type enrichment scores coming from independently sampled scRNAseq datasets reinforces the central claim that adult cell type distributions strongly reflect the brain connectome.

Comparison with previous work

Our work is preceded by several previous attempts to model the wiring diagram of the brain. For instance, Henriksen, *et al.* modeled the mouse mesoscale connectome with graph-theory-based approaches[7], and Reimann, *et al.* built a null model for the micro-connectome integrating the macro- and mesoscale connectomics[9]. Although these studies are not directly related to our current effort, they highlight the importance of graph-theoretic features and generative models in studying the mesoscale mouse connectome. In the present study, we have focused almost exclusively on molecular or cellular signatures of connectivity, but these studies indicate that future work incorporating additional graph theoretic contributors for predicting brain wiring diagram could be fruitful.

An approach much closer to ours was taken by French, *et al.*, who built statistical models correlating the gene expression signatures of 17,530 genes in 142 anatomical regions from the Allen Brain Atlas, and identified a subset of genes that are statistically correlated with the brain's wiring diagram[5]. They found a strong association between transcriptomic data and the connectome, which motivated us to create a predictive model of whole-brain connectivity from spatially distributed biological features. Ji, *et al.* went a step further by performing machine learning to predict the existence or absence of brain connectivity from gene expression, using a previous version of the AMBCA as a target[14]. Their approach yielded a very similar predictive accuracy and AUC as the classification results we present here, and their results underscore that random forest appears to be an excellent approach whether the features are based on regional gene expression or cell type distributions. However, in addition to predicting the existence of connectivity, here we also demonstrate that cell-type densities can be used to recreate the actual connectivity density values with high accuracy. An alternative, experimental approach linking cell types to connectivity is BRICseq, which allows for the high-throughput mapping of axonal tracts alongside the transcriptomic profiling of the projecting neurons[15]. However, BRICseq has not yet been scaled up to produce a connectivity map of comparable spatial resolution and coverage as the AMBCA[2,15]. Therefore, to our knowledge, no prior approach has been able to computationally link regional cell-type composition and whole-brain connectivity.

Cell type density versus gene expression as predictors of connectivity

An important aspect common to almost all prior models that predict connectivity from gene expression features is that the selected genes are not unique between datasets, and in fact are not unique predictors of connectivity even within the same dataset[5,14]. By contrast, the identified cell types are unique by definition and parsimonious enough to be considered stable markers between and within datasets[16]. We therefore believe that, compared with using gene expression, cell type features would be more appropriate and predictive for recreating the brain connectome for the following reasons: 1) Cells are the most fundamental unit responsible for inter-regional connectivity. 2) Most neural cell types have roughly fixed functions and spatial locations in the adult brain, whereas expression for many genes is highly temporally variable. 3) Using gene expression requires informed feature selection given the sheer number of mammalian genes and gene variants. While previous authors have reported such feature selection procedures, they necessarily rely on prior assumptions or knowledge. 4) The larger the feature set (e.g., using the entire mouse transcriptome), the higher the risk of overfitting and non-generalizability. Throughout our study, we have taken care to address these challenges, and the use of a small number of cell type features was considered a means of avoiding these pitfalls. Compared with the thousands of gene features used in prior studies[5,14], the set of 25 cell types should form a more parsimonious input feature space.

Oligodendrocytes are disproportionately associated with whole-brain connectivity patterns

Our analysis demonstrated the importance of oligodendricyte cell types in recreating the whole-brain connectome. Oligodendrocytes are the most predictive feature in the random forest model for both datasets (**Figure 5**), and are also among the highly predictive features when analyzing the feature importance for the classification task (**S. Figure 2**). Biologically, oligodendrocytes produce the myelin sheath insulating neuronal axons[25,26]. They help protect the vulnerable axons from parenchymal chemokines and cytokines, and ensure the fast and efficient movement of action potentials[25,26,40]. Dysfunction of oligodendrocytes can interfere with normal micro-structure and functional connectivity in the mouse brain[41]. Oligodendrocyte myelination was also shown in previous work to be able to regulate the loss of synapses[42]. Moreover, recent work from Buchanan, *et al.* showed that oligodendrocyte precursor cells can prune axons in the mouse neocortex[43]. When we modeled short-range and long-range connectivity separately, we found that while oligodendrocytes contributed strongly to short-range connectivity, they were somewhat less informative for reconstructing long-range connectivity (**Figure 6C-G**). Overall, our results underscore the critical role this cell type plays in maintaining white matter integrity.

Non-neuronal cells contribute to whole-brain and short-range connectivity

Our analysis also demonstrated the importance of non-neuronal cell types in recreating the whole-brain connectome. Brain vascular cells compose the blood-brain barrier, which protects the vulnerable central nervous system (CNS), and they interact with the CNS for supporting neuronal cells with nutrients, energy, and oxygen[27–31]. Their breakdown was found to be strongly correlated with brain connectivity disruption and cognitive defects[27,31]. Brain endothelial cells are also involved in the process of neurovascular coupling[44,45], whereby local neural activity stimulates subsequent blood flow changes in the corresponding downstream locations[45,46]. The fact that endothelial cells are more important for short- and medium-range connections but not for long-range ones supports a role in local circuit maintenance rather than long projections. We also found that immune cell and astrocytes play an outsize role in predicting connectivity compared to neuronal cell types. Previous studies have indicated that there is an association between inflammation and functional brain connectivity[33,47]. Similarly, astrocytes, the most abundant glial cells in the CNS, have critical impact in maintaining many physiological functions of neurons. Germane to this investigation, previous experimental work has shown the existence of bidirectional interactions between astrocytes and synapses[48].

Further, we found that non-neuronal cell types contribute disproportionately to predicting short-range connectivity. Of these, vascular cells were the most important supertype for the Tasic, *et al.* dataset and immune cells were the most important supertype for the Zeisel, *et al.* dataset, although all non-neuronal supertypes tended to have higher feature importance scores than most of the neuronal supertypes (**Figure 6C,E**). There are multiple reasons why these non-neuronal cells have higher feature importance scores for predicting short-range as opposed to long-range connectivity. Generally, many non-neuronal cell types are thought to impact and interact with neighboring neuronal cell bodies in the gray matter, which may result in the mediation of more local, short-range connectivity. Alternatively, it is possible that non-neuronal cells, in their various roles supporting neuronal function, are important in the formation and maintenance of all connections in the CNS (**Figure 5A-B**). However, given that feature importance is a relative measure of the model information provided by a given feature, non-neuronal cell types contribute at most moderately to the long-range models of connectivity because certain neuronal cell types have an outsize distance-dependent effect (see below; **Figure 6C-G**). The distance dependence of cell-type contributions to connectivity is an important line of inquiry for future studies.

Non-neuronal cell density strongly corresponds to node-level graph metrics

The biological functions of oligodendrocytes and other non-neuronal cell types in maintaining network architecture are further supported by our finding that these cell types have a strong spatial correspondence to node-level measures of graph centrality (**Figure 5C** and **D**, **S. Figure 5**). Our chosen graph centrality measures were degree and eigenvector centrality; both convey information about the centrality of each node in the connectome. More specifically, a node that has high values for these centrality measures is expected to be a prominent hub of the network. We therefore expected that the most central nodes should also be enriched in the cell types highlighted by our feature importance analysis. The fact that our selected cell types are strongly associated with common graph theoretic measures of node centrality mirrors the similar relationship between graph topology and gene expression in mouse[8,14] and human[10,11,49] brains.

Our findings show that non-neuronal cells' densities are significantly associated with connectome graph metrics of centrality, including degree centrality and eigenvector centrality, revealing the cellular correlates of canonical hubs well known in brain graph theory. While no prior study has interrogated cell types directly, indirect evidence from gene expression analysis variously implicates non-neuronal (metabolic and oxidative) processes in governing network properties[8,49]. Specifically, Fulcher, *et al.* note: "The high transcriptional coupling associated with hub connectivity is driven by genes regulating the oxidative synthesis and metabolism of ATP - the primary energetic currency of neuronal communication. This genetic signature contrasts that identified for neuronal connectivity in general, which is driven by genes regulating neuronal, synaptic, and axonal structure and function."[8] Our results highlighting oligodendrocyte cells, which is involved in neural energetics and maintenance, support this line of reasoning.

Neuronal subtypes differentially mediate long-range connectivity

In addition to oligodendrocytes, we found that telencephalic glutamatergic neurons and striatal medium spiny neurons were among the most salient classes of cell types, but only for predicting long-range connections (**Figure 6C-F**). The former are well known to project to remote locations within and outside of the neocortex (**Figure 7A,C**), and therefore their prominence in long-range but not shorter connections is consistent with their neurobiology. It is particularly striking that the telencephalic glutamatergic cell supertypes in both the Tasic, *et al.* and Zeisel, *et al.* datasets (Neo Glu and Hip Neo Glu, respectively) are also among the least important features for predicting short-range connectivity (**Figure 6C,E**), suggesting that these neurons predominantly engage in long-range connections. Similarly, the high feature importance of medium spiny neurons is concordant with their function, as these are long-range-projecting, inhibitory neurons. Medium spiny neurons comprise a significant fraction of neurons in the striatum and are involved in dopamine signaling; notably, these neurons selectively exhibit altered behavior in several psychiatric disorders[50]. When we look at the feature importance of individual cell types between the two

datasets, we see a similar pattern as we do at the supertype level (**Figure 6G**). In particular, telencephalic glutamatergic neurons contribute weakly to predicting short-range connectivity and are overrepresented among types with high feature importance scores for predicting long-range connectivity. Since telencephalic glutamatergic neurons comprise many of the long-range, inter-regional connections of the brain, the distance dependence we observed is biologically plausible. Taken together, these results suggest that the formation and maintenance of brain connections requires a wide array of cell types. However, we caution that this kind of feature importance analysis will benefit from further experimental work to elucidate in more detail the biological roles of the identified cell types with respect to connectivity.

Limitations

A primary limitation of the current work is that regional cell type enrichment scores in the adult mouse brain can only explain neural inter-regional connectivity up to a certain extent. Neural polarity, cell maturation and migration are all critical for determining the brain connectivity, which our model cannot capture. A further issue is that, even though random forest models are more interpretable than other machine learning algorithms, uncovering the true contributions of each cell type in constructing inter-regional connectivity is not as straightforward with RF as with generalized linear statistical models. RF models can also exhibit problems with overfitting: since the RF model is an ensemble of decision trees, a single decision tree is very sensitive to data variations and can therefore easily overfit to noise. However, we note that the issue of overfitting cuts across almost all machine learning methods and is not specific to RF. In this study we have taken great care at various steps to minimize this risk, starting from the basic design of using only the cell-type features from the two connecting regions, and eschewing full brain or neighboring regional features. Also, as mentioned above, we have not explored feature selection to produce a minimal set of informative cell types, either for the 25-type Tasic, *et al.* or the 200-type Zeisel, *et al.* dataset, and therefore it is possible that the model performance demonstrated here could be further enhanced. Finally, we were still limited by the resolution of both the mouse brain connectome and cell type density maps, and therefore did not attempt to separately predict additional features of keen interest, such as cell polarity.

Future directions

The important cell type features we found would potentially help future work using generalized models to predict whole brain connectivity. As mentioned above, prior feature selection on either of the cell type datasets used here may facilitate the development of more predictive models. Additionally, machine learning models that integrate both cellular features and anatomic features can be expected to develop better predictions for certain tasks. Creating cell-to-cell or even voxel-to-voxel level connectivity and benchmarking against known neuronal cell-type-specific signaling pathways would be beneficial for the future research but requires higher-resolution input datasets, which may be accessible with future maps inferred using MISS. Given the conservation of central nervous system properties in mammals, we may also be able to apply these data-driven methods to the human brain.

Conclusions

We have successfully implemented a machine-learning-based, data-driven approach that can reconstruct whole-brain connectivity from regional cell type information. Our results using a random forest model exhibited very high accuracy in both identifying connected region-pairs and predicting connectivity density, revealing the advantages of nonlinear methods for this specific task. Furthermore, feature importance analysis implicated non-neural cell types as particularly vital in recreating inter-regional connectivity. The present work is, to our knowledge, the first application of machine learning methods to the task of predicting mammalian brain connectivity with cellular enrichment features as inputs. We anticipate that our results may provide guidelines for future experimental analysis, and our analysis can be extended to examine the dependence of connectivity on cell type distributions in other mammals, including humans.

Methods

Input datasets

We use two primary sources of data: MISS-derived cell type enrichment scores, which are themselves a function of gene expression data and serve as our models' input features, and the Allen Mouse Brain Connectivity Atlas (AMBCA), which serves as our empirical ground truth for training and testing our models.

MISS-derived cell type features

Although the Allen Gene Expression Atlas (AGEA) contains spatially resolved gene expression information for thousands of genes[17], a similar dataset directly mapping a comprehensive set of cell types in the mouse brain has not been produced. The lack of such a cell type atlas has thwarted efforts at quantitatively exploring the dependence of connectivity on cell type composition. However, our lab has recently developed the the Matrix Inversion and Subset Selection (MISS) pipeline[16], which

is capable of deconvolving the spatial gene expression data from the AGEA into cell type densities with cell-type-specific single-cell RNA-seq (scRNAseq) data. First, MISS uses an information-theoretic criterion to identify an informative gene subset, and then it solves a nonnegative least squares problem to infer the densities of each cell type per voxel of the AGEA. These values were then averaged over the 424 regions of the mouse Common Coordinate Framework (CCF) to obtain regional densities that could be used as input features in our machine learning approach. The primary scRNAseq dataset comprised of 25 cell types, which were sampled from the primary visual cortex, the anterior lateral motor cortex, and the dorsal lateral geniculate complex, uses scRNAseq data made available by the Allen Institute for Brain Science (AIBS)[18,20]. A second dataset of 200 cell types was derived from scRNAseq data collected for the Mouse Brain Atlas (mousebrain.org)[19]. These cell type densities are min-max normalized to avoid the bias from the cell types' own artificial scales to create our cell type enrichment features, ensuring that each regional cell type value falls in the range $[0,1]$. For further methodological details on the MISS algorithm, refer to the Supplement and the original publication by Mezias, *et al.*[16].

Mouse connectivity

We use the AMBCA as the source of the mouse connectome we reconstruct from cell type features (http://connectivity.brain-map.org), which was assembled using viral tracing[2]. Briefly, the authors injected enhanced green fluorescent protein (EGFP)-expressing adeno-associated viral vectors to trace axonal projections from defined regions, which were then imaged using high-throughput serial two-photon tomography. For full methodological details, please refer to Oh, *et al.*[2].

The resulting mesoscale connectome, C, is represented as a 426×426 matrix, with $C(i,j)$ representing the total connectivity from region i to region j. The left and right ansiform lobules were removed due to the missing values in the AGEA, leaving a 424×424 connectome. We also removed self-connectivity, since we are primarily interested in reconstructing inter-regional connectivity. We followed the instructions provided by the Allen Institute for Brain Science (AIBS) to compute the connectivity density from the total connectivity by dividing the values by the outputting regions' volumes (http://connectivity.brain-map.org). To transform skewed data to approximately conform to normality, we log-transformed the resulting connectivity densities.

For the classification task, we processed raw connectivity data to convert the raw continuous variables to binary values (1 and 0). Work by Ji, *et al.* set an artificial threshold to make their two outcomes' portion balanced[14]. We decided against thresholding because we wanted to maintain the sparsity of the brain connectome (only 36% non-zero values). The resulting data imbalance certainly made our prediction task harder but we believe this is necessary for capturing the real biology germane to the brain connectome. After removing the self-connectivity as described above, there remained in total 179,352 data samples, pertaining to all connected region-pairs. For the regression task (predicting connectivity density), we first removed all unconnected region-pairs, leaving 64,566 samples.

Machine learning methods

We implemented several machine learning methods for predicting brain connectivity. We divided our ML prediction tasks by separately predicting the absence or presence of a connection and the connectivity density between any given region pair. For all methods, we randomly split the connectivity data and performed a 10-fold cross-validation, assigning 90 percent of the data points to the training set while leaving 10 percent for testing in each iteration. All reported evaluations were performed on the test set, the average of which were calculated for 10 iterations.

Input features

For both the prediction tasks, we used as input features the 25-long cell type enrichment vectors from both the source and target regions, resulting in 50 total feature. We also explored additional feature engineering methods to generate more informative features from these 50 features, but the prediction accuracy did not increase significantly (data not shown); therefore, in this study we report results based on only the cell type features described above. For the independent Zeisel, *et al.* dataset, we again apply as input features the 200-long cell type enrichment vectors from both the source and target regions, resulting in 400 features in total. Other settings were kept identical to the analysis of the main Tasic, *et al.* dataset.

Random forest

The main findings reported in this study were obtained from random forest models for both the classification task and the regression task[51–53]. This model generates a number of decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control overfitting. For each task, 300 random trees were generated and evaluated during each of 10 iterations. Since the random forest model is well-known to suffer from a risk of overfitting[53,54], we employed several means to reduce this risk. First, employing lower-dimensional cell type features in comparison to prior work that uses a much larger number of gene expression features helps to mitigate overfitting issues. Our model design also excluded the use of higher-order features - e.g., from multiple brain regions - this also serves to reduce overfitting risk. We further set the maximal number of features for each tree at 20 and the maximal depth for each tree at 15. The specific random forest model we implemented was the one contained in the Python package `Scikit-learn` *v*0.20.2[55].

Other ML models

In addition to random , we implemented and tested several common machine learning algorithms including linear models such as ridge[56] and LASSO[57]. We also implemented support vector machines (SVMs) with a Radial (RBF) kernel for the task[58,59]. SVMs are suitable for generating classification hyperplanes such that the margins between the hyperplane and the nearest instances of the classified sample categories are maximized. In doing so, they allow for achieving global optimal solutions and hence aid in the generalizations of the resultant classifiers. The rationale for adopting SVM was primarily due to the overfitting issue noted above for the random forest model. Ridge regression, LASSO, and SVM were also implemented using the `Scikit-learn` *v*0.20.2 code library. Other models including DecisionTree, GrandientBoosting, ExtraTrees and KNeighbors are implemented by Scikit-learn.

Neural network models

The above models are all conventional ML methods that excel at lower-dimensional and small sample size scenarios, which are suitable for the current task. However, it is possible that modern neural-network-based models might perform better - an empirical question we subsequently attempted to explore using shallow and deep learning models. We first implemented the most common and practical feedforward artificial neural network, the multilayer perceptron (MLP). We first constructed the common multilayer perceptron model, for both classifier and regressor tasks using `Scikit-learn` *v*0.20.2. The sizes of each MLP are as follows: 1st hidden layer size, 256; 2nd hidden layer size, 64; 3rd hidden layer size: 256. In each, activation leaky RELU (the rectified linear activation function) was implemented. This model did not achieve results comparable to the above classical ML models (see **S. Data Tables 2** and **3**). It is possible that these results might be poor due to our choice of the MLP model in Scikit-learn, which is by design simplified and does not admit more advanced algorithmic choices. To address this aspect we also built more advanced neural network models using a Pytorch[60]-based multi-layer perceptron. The network structure is as follows: number of input features, 50; number of neurons in each layer, 512-64-16-4-1 (the final prediction value). A stochastic gradient descent algorithm that sought to minimize the mean squared error (MSE) loss was used for model training and optimization. The drop-out ratio was set to 0.5 and batch normalization was performed.

Model performance evaluation

As mentioned above, all the model evaluation results in this paper are reported for the testing dataset only, after 10-fold cross-validation. For the classification task, precision and recall metrics are reported:

$$precision = \frac{TP}{TP+FP}$$

$$recall = \frac{TP}{TP+FN}$$

where TP is true positives, TN is true negatives and FP is false positives.

For the regression task, Root Mean Square Error (RMSE) was used to evaluate the quality of predictions:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(true - predicted)^2}{N}}$$

R-squared score, also known as the the coefficient of determination, is defined as the proportion of the variation in the dependent variable that is predictable from the independent variable(s):

$$R^2 = 1 - \frac{RSS}{TSS}$$

where RSS is sum of square residuals and TSS is total sum of squares

3D brain visualization

We used Brainframe, an in-house MATLAB package developed at the Raj laboratory, to generate the 3D mouse brains, the distribution of gene expression and cell-type patterns within, and the brain connectome. The cell-type maps shown here, which were first presented by Mezias, *et al.*[16], are rendered per-voxel level after applying a threshold for clarity. For the neuronal supertypes from the Zeisel, *et al.*, we perform principal component analysis on the cell types that comprise each supertype and then plot the first principal component scores of each voxel after min-max normalization. We render connectivity as a sphere-and-arrow plot, where each sphere and arrow is color-coded by major region group and connections are thresholded by density for clarity. For more details, view the Brainframe documentation (https://github.com/Raj-Lab-UCSF/Brainframe).

Regional distance matrix calculation

To calculate the distance between each region-pair in the mouse CCF[2], we first determine the center of mass of each region by averaging across the x, y, and z coordinates of all voxels with that region label. We then use the pairwise Euclidean distance between these regional centers of mass as a proxy for the lengths of the white matter tracts connecting them. We note that these distances based on center of mass are not a perfect analog for fiber length, as projections can take circuitous routes to connect regions that may be otherwise close in spatial proximity. However, these fiber lengths are challenging to determine from the available data from the AIBS, and for most region-pairs Euclidean center-of-mass distance is a reasonable approximation.

Feature interpretation from random forest models

To decompose the random forest model and calculate the importance of each input feature, we used `Scikit-learn` v0.20.2 Python package[21]. For each decision tree t, Scikit-learn calculates a node's importance, assuming only two child nodes per parent node (binary tree):

$$NI_j = w_j I_j - w_{left(j)} I_{left(j)} - w_{right(j)} I_{right(j)},$$

where NI_j is the node importance of node j, w_j is the weighted number of samples reaching node j, I_j is the impurity value of node j, and the subscripts $left(j)$ and $right(j)$ indicate the left and right child nodes of node j, respectively.

Impurity for a node, I_j, is calculated differently depending on whether the task is regression or classification. The regression task impurity is defined as the variance reduction across instances:

$$I = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2,$$

where y_i is label/value for an instance, N is the number of instances and $\bar{y}$ is the mean across instances.

For the classification task, we used the Gini impurity:

$$I = \sum_{i=1}^{C} f_i(1 - f_i),$$

where f_i is frequency of label i at a node and C is the number of unique labels (e.g., $C = 2$ for the binary classification task).

The importance for the i^{th} input feature on a decision tree, FI_i, is then calculated from node importance as:

$$FI_{i(t)} = \frac{\sum_{j=1}^{N_i} NI_{j(t)}}{\sum_{k=1}^{N} NI_{k(t)}},$$

where N_i is the number of nodes using feature i, N is the total number of nodes in the tree, and t is individual decision tree index.

These raw FI_i values are then normalized such that the sum of feature importance across all features is 1:

$$\overline{FI}_{i(t)} = \frac{FI_{i(t)}}{\sum_{k=1}^{M} FI_{k(t)}},$$

where M is the total number of input features and t is individual decision tree.

The final feature importance at the Random Forest level is its average over all the trees:

$$\overline{FI}_i = \frac{\sum_{t=1}^{T} \overline{FI}_{i(t)}}{T},$$

where T is the total number of trees and t is individual decision tree index. For more operation details, please refer to `Scikit-learn` v0.20.2 Python package[21].

We note that there are two region-level input features for each cell type in the dataset, one for the receiving region and one for the outgoing region. The reported feature importance values for each cell supertype are averaged across their corresponding input cell-type features. Distributions of feature importance represent the results of 10-fold cross-validation.

References

1. Sporns, O., Tononi, G. & Kötter, R. PLoS Comput BiolThe human connectome: A structural description of the human brain. *PLoS Comput. Biol* **1**, e42 (2005).
2. Oh, S. W. *et al.* A mesoscale connectome of the mouse brain. *Nature* **508**, 207–214, DOI: 10.1038/nature13186 (2014).
3. Bullmore, E. T. & Bassett, D. S. Annu Rev Clin PsycholBrain graphs: graphical models of the human brain connectome. *Annu. Rev Clin Psychol* **7**, 113–140 (2011).
4. Zeng, H. Curr Opin NeurobiolMesoscale connectomics. *Curr Opin Neurobiol* **50**, 154–162 (2018).
5. French, L. & Pavlidis, P. Relationships between gene expression and brain wiring in the adult rodent brain. *PLoS Comput. Biol.* **7**, e1001049, DOI: 10.1371/journal.pcbi.1001049 (2011).
6. Tan, P. P. C., French, L. & Pavlidis, P. Neuron-enriched gene expression patterns are regionally anti-correlated with oligodendrocyte-enriched patterns in the adult mouse and human brain. *Front. Neurosci.* **7**, DOI: 10.3389/fnins.2013.00005 (2013).
7. Henriksen, S., Pang, R. & Wronkiewicz, M. A simple generative model of the mouse mesoscale connectome. *eLife* **5**, DOI: 10.7554/elife.12366 (2016).
8. Fulcher, B. D. & Fornito, A. A transcriptional signature of hub connectivity in the mouse connectome. *Proc. Natl. Acad. Sci.* **113**, 1435–1440 (2016).
9. Reimann, M. W. *et al.* A null model of the mouse whole-neocortex micro-connectome. *Nat. Commun.* **10**, DOI: 10.1038/s41467-019-11630-x (2019).
10. Goel, P., Kuceyeski, A., Locastro, E. & Raj, A. Spatial patterns of genome-wide expression profiles reflect anatomic and fiber connectivity architecture of healthy human brain. *Hum. Brain Mapp.* **35**, 4204–4218, DOI: 10.1002/hbm.22471 (2014).
11. Vértes, P. E. *et al.* Gene transcription profiles associated with inter-modular hubs and connection distance in human functional magnetic resonance imaging networks. *Philos. Transactions Royal Soc. B: Biol. Sci.* **371**, 20150362 (2016).
12. Diez, I. & Sepulcre, J. Neurogenetic profiles delineate large-scale connectivity dynamics of the human brain. *Nat. communications* **9**, 1–10 (2018).
13. Shen, X. *et al.* Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *Nat. Protoc.* **12**, 506–518, DOI: 10.1038/nprot.2016.178 (2017).
14. Ji, S., Fakhry, A. & Deng, H. Integrative analysis of the connectivity and gene expression atlases in the mouse brain. *NeuroImage* **84**, 245–253 (2014).
15. Huang, L. *et al.* BRICseq bridges brain-wide interregional connectivity to neural activity and gene expression in single animals. *Cell* **182**, 177–188.e27, DOI: 10.1016/j.cell.2020.05.029 (2020).
16. Mezias, C., Torok, J., Maia, P. D., Markley, E. & Raj, A. Matrix inversion and subset selection (miss): A pipeline for mapping of diverse cell types across the murine brain. *Proc. Natl. Acad. Sci.* **119**, e2111786119 (2022).
17. Lein, E. S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176, DOI: 10.1038/nature05453 (2007).
18. Tasic, B. *et al.* Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78, DOI: 10.1038/s41586-018-0654-5 (2018).
19. Zeisel, A. *et al.* Molecular architecture of the mouse nervous system. *Cell* **174**, 999–1014.e22, DOI: 10.1016/j.cell.2018.06.021 (2018).
20. AIBS. Allen Cell Types Database - Technical White Paper: Transcriptomics (2018).
21. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
22. Fakhry, A. & Ji, S. High-resolution prediction of mouse brain connectivity using gene expression patterns. *Methods* **73**, 71–78, DOI: 10.1016/j.ymeth.2014.07.011 (2015).
23. Fornito, A., Arnatkevičiūtė, A. & Fulcher, B. D. Bridging the gap between connectome and transcriptome. *Trends Cogn. Sci.* **23**, 34–50, DOI: 10.1016/j.tics.2018.10.005 (2019).
24. Anaissi, A., Goyal, M., Catchpoole, D. R., Braytee, A. & Kennedy, P. J. Ensemble feature learning of genomic data using support vector machine. *PLOS ONE* **11**, e0157330, DOI: 10.1371/journal.pone.0157330 (2016).

25. PFEIFFER, S., WARRINGTON, A. & BANSAL, R. The oligodendrocyte and its many cellular processes. *Trends Cell Biol.* **3**, 191–197, DOI: 10.1016/0962-8924(93)90213-k (1993).

26. Emery, B. Regulation of oligodendrocyte differentiation and myelination. *Science* **330**, 779–782, DOI: 10.1126/science.1190927 (2010).

27. Abbott, N. J., Patabendige, A. A., Dolman, D. E., Yusof, S. R. & Begley, D. J. Structure and function of the blood–brain barrier. *Neurobiol. Dis.* **37**, 13–25, DOI: 10.1016/j.nbd.2009.07.030 (2010).

28. Langen, U. H., Ayloo, S. & Gu, C. Development and cell biology of the blood-brain barrier. *Annu. Rev. Cell Dev. Biol.* **35**, 591–613, DOI: 10.1146/annurev-cellbio-100617-062608 (2019).

29. Chow, B. W. & Gu, C. The molecular constituents of the blood–brain barrier. *Trends Neurosci.* **38**, 598–608, DOI: 10.1016/j.tins.2015.08.003 (2015).

30. Daneman, R. & Prat, A. The blood–brain barrier. *Cold Spring Harb. Perspectives Biol.* **7**, a020412, DOI: 10.1101/cshperspect.a020412 (2015).

31. Ballabh, P., Braun, A. & Nedergaard, M. The blood–brain barrier: an overview. *Neurobiol. Dis.* **16**, 1–13, DOI: 10.1016/j.nbd.2003.12.016 (2004).

32. Reemst, K., Noctor, S. C., Lucassen, P. J. & Hol, E. M. The Indispensable Roles of Microglia and Astrocytes during Brain Development. *Front. Hum. Neurosci.* **10**, 566, DOI: 10.3389/fnhum.2016.00566 (2016).

33. Morimoto, K. & Nakajima, K. Role of the Immune System in the Development of the Central Nervous System. *Front. Neurosci.* **13**, 916, DOI: 10.3389/fnins.2019.00916 (2019).

34. Hagberg, A., Swart, P. & S Chult, D. Exploring network structure, dynamics, and function using networkx. Tech. Rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States) (2008).

35. Ouyang, M., Kang, H., Detre, J. A., Roberts, T. P. & Huang, H. Short-range connections in the developmental connectome during typical and atypical brain maturation. *Neurosci. & Biobehav. Rev.* **83**, 109–122, DOI: 10.1016/j.neubiorev.2017.10.007 (2017).

36. Naze, S., Proix, T., Atasoy, S. & Kozloski, J. R. Robustness of connectome harmonics to local gray matter and long-range white matter connectivity changes. *NeuroImage* **224**, 117364, DOI: 10.1016/j.neuroimage.2020.117364 (2021).

37. Stephan, A. H., Barres, B. A. & Stevens, B. The complement system: An unexpected role in synaptic pruning during development and disease. *Annu. Rev. Neurosci.* **35**, 369–389, DOI: 10.1146/annurev-neuro-061010-113810 (2012).

38. Neniskyte, U. & Gross, C. T. Errant gardeners: glial-cell-dependent synaptic pruning and neurodevelopmental disorders. *Nat. Rev. Neurosci.* **18**, 658–670, DOI: 10.1038/nrn.2017.110 (2017).

39. Liu, Y.-J. *et al.* Microglia elimination increases neural circuit connectivity and activity in adult mouse cortex. *The J. Neurosci.* **41**, 1274–1287, DOI: 10.1523/jneurosci.2140-20.2020 (2020).

40. Eroglu, C. & Barres, B. A. Regulation of synaptic connectivity by glia. *Nature* **468**, 223–231, DOI: 10.1038/nature09612 (2010).

41. Kawamura, A. *et al.* Chd8 mutation in oligodendrocytes alters microstructure and functional connectivity in the mouse brain. *Mol. Brain* **13**, DOI: 10.1186/s13041-020-00699-x (2020).

42. Wang, F. *et al.* Enhancing oligodendrocyte myelination rescues synaptic loss and improves functional recovery after chronic hypoxia. *Neuron* **99**, 689–701.e5, DOI: 10.1016/j.neuron.2018.07.017 (2018).

43. Buchanan, J. *et al.* Oligodendrocyte precursor cells prune axons in the mouse neocortex. *bioRxiv* DOI: 10.1101/2021.05.29.446047 (2021).

44. Cauli, B. Revisiting the role of neurons in neurovascular coupling. *Front. Neuroenergetics* **2**, DOI: 10.3389/fnene.2010.00009 (2010).

45. Chow, B. W. *et al.* Caveolae in CNS arterioles mediate neurovascular coupling. *Nature* **579**, 106–110, DOI: 10.1038/s41586-020-2026-1 (2020).

46. Kaplan, L., Chow, B. W. & Gu, C. Neuronal regulation of the blood–brain barrier and neurovascular coupling. *Nat. Rev. Neurosci.* **21**, 416–432, DOI: 10.1038/s41583-020-0322-2 (2020).

47. Jafari, A., de Lima Xavier, L., Bernstein, J. D., Simonyan, K. & Bleier, B. S. Association of sinonasal inflammation with functional brain connectivity. *JAMA Otolaryngol. & Neck Surg.* **147**, 534–543 (2021).

48. Allen, N. J. & Eroglu, C. Cell biology of astrocyte-synapse interactions. *Neuron* **96**, 697–708 (2017).

49. Arnatkeviciute, A. *et al.* Genetic influences on hub connectivity of the human connectome. *Nat. Commun.* **12**, 4237, DOI: 10.1038/s41467-021-24306-2 (2021).

50. Chuhma, N., Tanaka, K. F., Hen, R. & Rayport, S. Functional Connectome of the Striatal Medium Spiny Neuron. *J. Neurosci.* **31**, 1183–1192, DOI: 10.1523/JNEUROSCI.3833-10.2011 (2011).

51. Ho, T. K. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis machine intelligence* **20**, 832–844 (1998).

52. Breiman, L. Random forests. *Mach. learning* **45**, 5–32 (2001).

53. Qi, Y. Random forest for bioinformatics. In *Ensemble machine learning*, 307–323 (Springer, 2012).

54. Segal, M. R. Machine learning benchmarks and random forest regression. *UCSF: Cent. for Bioinforma. Mol. Biostat.* (2004).

55. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. machine Learn. research* **12**, 2825–2830 (2011).

56. Hoerl, A. E. & Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).

57. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc. Ser. B (Methodological)* **58**, 267–288 (1996).

58. Boser, B. E., Guyon, I. M. & Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, 144–152 (1992).

59. Chang, C.-C. & Lin, C.-J. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems technology (TIST)* **2**, 1–27 (2011).

60. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. In Wallach, H. *et al.* (eds.) *Advances in Neural Information Processing Systems 32*, 8024–8035 (Curran Associates, Inc., 2019).

611 # Figures

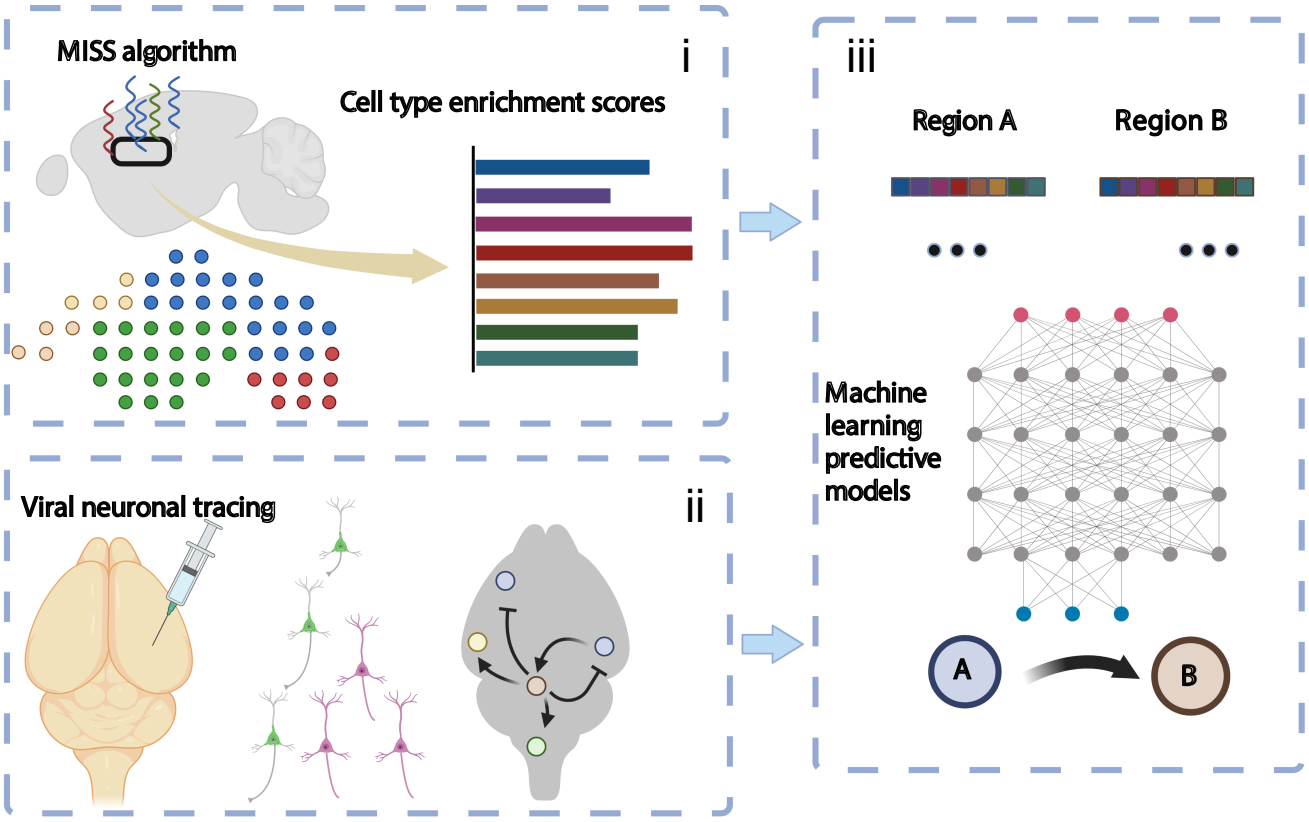

Figure 1. Study design. *Top left*: The spatial quantification of cell type enrichment was computed with the computational pipeline MISS[16] from publicly available gene expression data. *Bottom Left*: The brain connectivity graph was measured by Allen Mouse Brain Connectivity Atlas (AMBCA) using viral neuronal tracing techniques. *Right*: Machine learning algorithms were then implemented to predict the connectivity between each two regions.

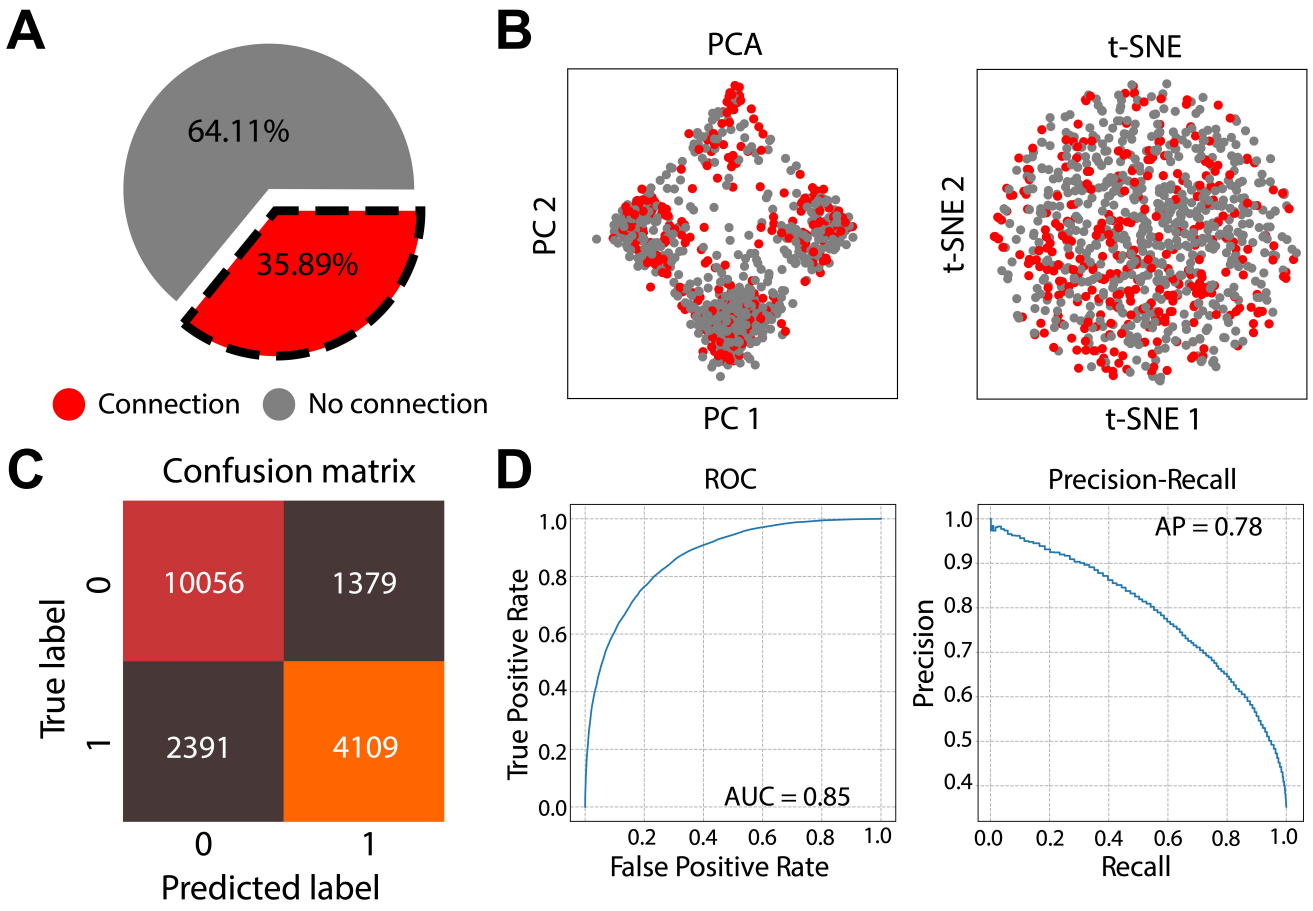

Figure 2. Machine learning applied to regional cell type distributions was able to predict the existence of connectivity
A. Pie chart of the distribution of the two classes in the AMBCA, where a region pair is deemed to be connected if its connectivity density is greater than 0. **B**. Common dimension reduction methods on the cell type spatial quantification for individual inter-regional connectivity. *Left*: principal component analysis (PCA) of the cell type spatial quantification array. *Right*: t-distributed stochastic neighbor embedding (t-SNE) of the cell type spatial quantification array. Neither method shows distinct clusters based on the presence or absence of connectivity. **C**. The confusion matrix of the binary prediction outcome with random forest methods. (0: Connectivity not detected 1: Connectivity detected). **D**. Performance evaluation of the classifier model using ten-fold cross-validation. *Left*: The receiver operating characteristic curve. AUC = 0.85. *Right*: The precision recall curve. AP = 0.77

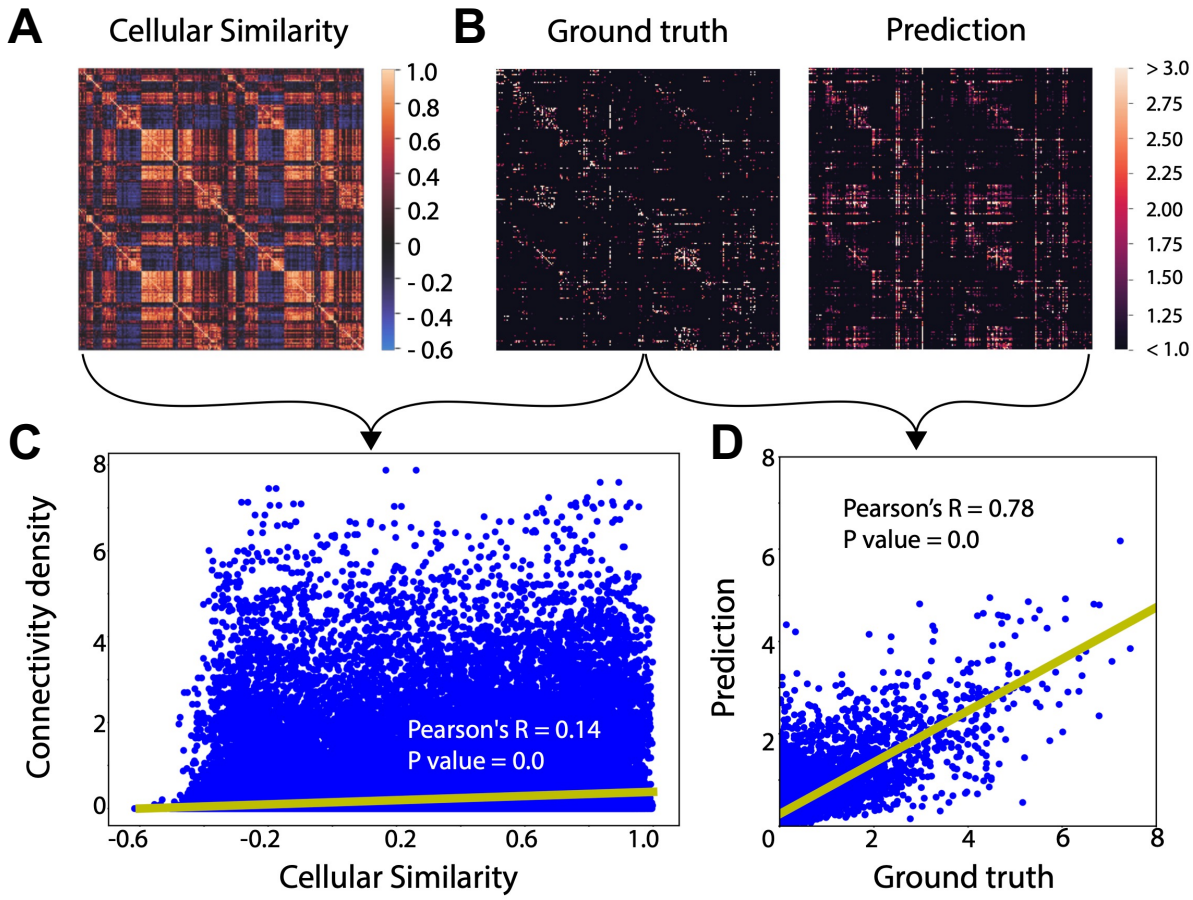

Figure 3. Machine learning model was able to predict connectivity strength across the entire mesoscale mouse connectome. **A**. Cellular similarity matrix (quantified using Pearson correlation) of spatial cell type enrichment quantification across brain regions. **B**. *Left*: Brain connectivity matrix (log2-transformed). *Right*: RF prediction without splitting the training and test set. The depicted matrices' rows and columns represent individual regions, and the connectivity between regions is denoted by the matrix entries. The random forest model was able to qualitatively reconstruct the whole brain connectome. **C**. Scatter plot of pairwise cellular similarity (as depicted in **A**) between two regions' cell type distribution vectors versus the log-transformed connectivity strength between the two regions (as depicted in **B**, left), and the fitted linear regression curve (Pearson's R = 0.14, Spearman's $\rho = 0.08$, $p = 0.0$). **D**. Scatter plot showing the correlation between the ground truth connectivity strength between all regions pairs with non-zero connectivity and their predicted values for connectivity using cell types as predictors in the RF model (test set only), along with the fitted linear regression curve (Pearson's R = 0.78).

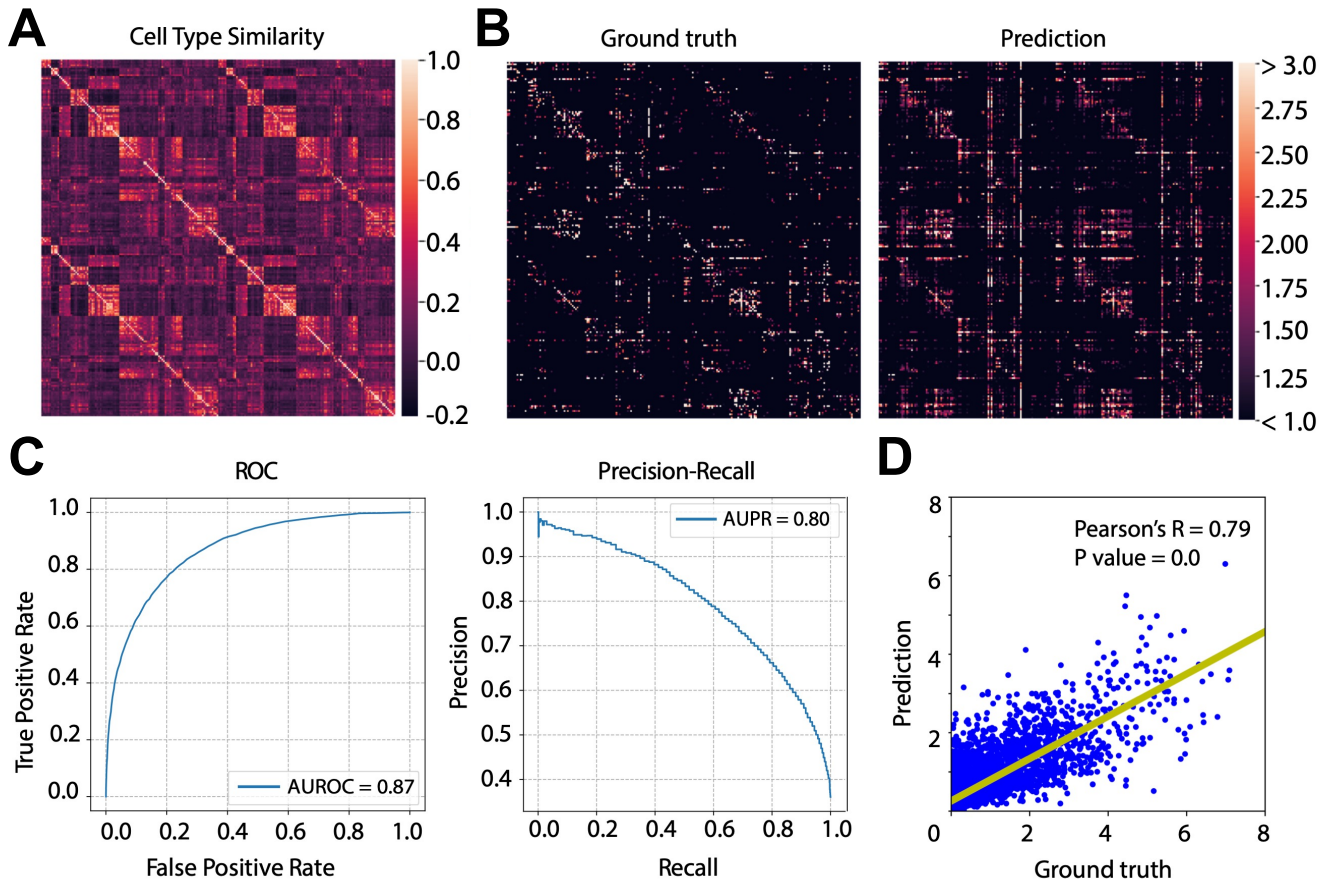

Figure 4. Prediction of brain anatomic connectivity using the independent data set from Zeisel, *et al.* **A.** Cellular similarity matrix (quantified by Pearson's R) of spatial cell type enrichment quantification across brain regions. **B.** Heatmap renderings of inter-regional connectivity of ground truth (*Left*), and cell-type-based RF prediction (*Right*). **C.** Predicting the existence of connectivity. Receiver operating characteristic (ROC) curves *left* and precision-recall curves *right* show almost identical performance to the main results from the Tasic, *et al.* dataset. **D.** Scatter plot showing the correlation between the ground truth connectivity strength between all regions pairs with non-zero connectivity and their predicted values for connectivity using cell types as predictors in the RF model, along with the fitted linear regression curve (Pearson's R = 0.79).

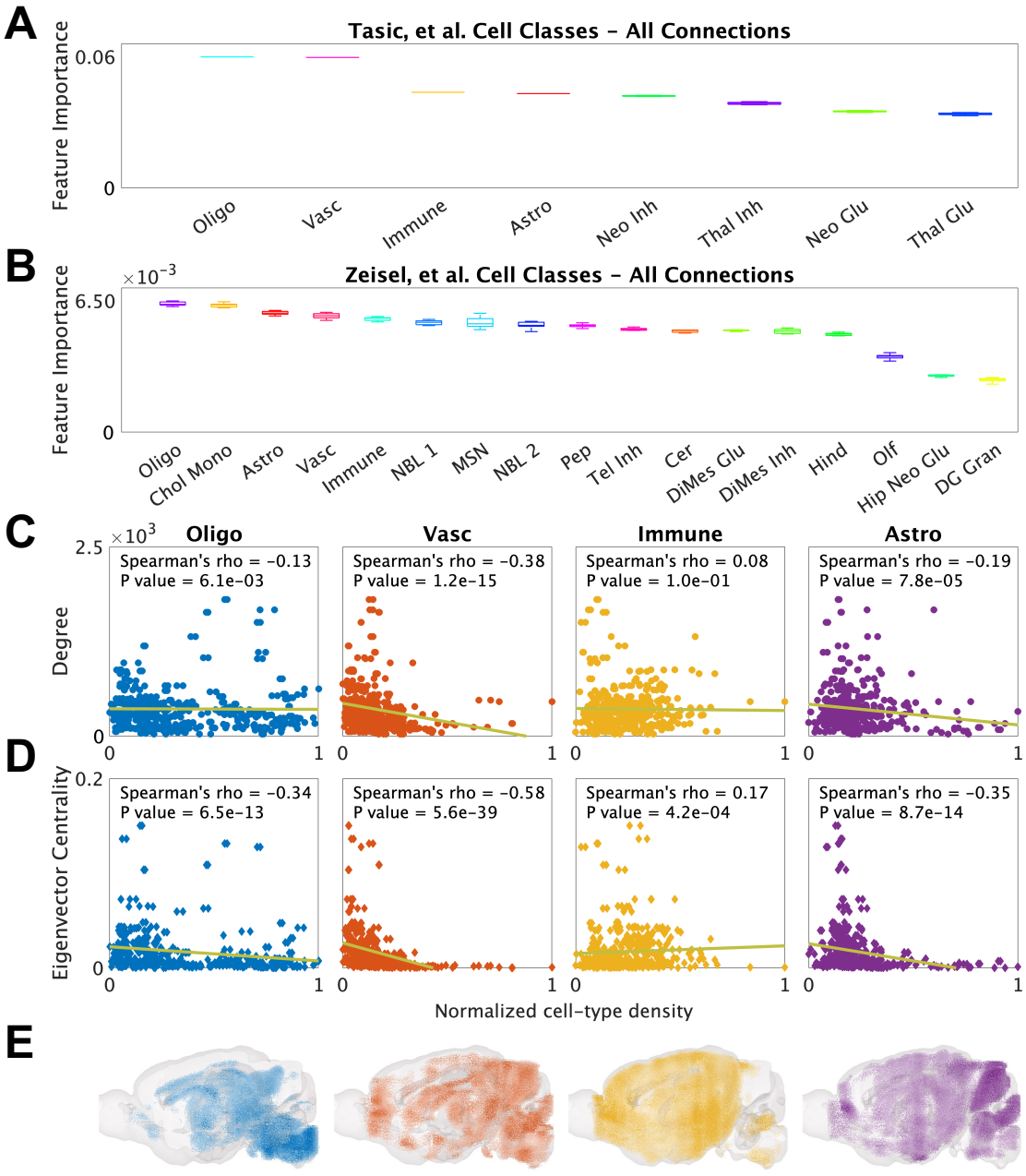

Figure 5. Interrogating the individual contributions of cell types. A. Bar chart showing the feature importance values of all cell types in the random forest model for the Tasic, *et al.* cell-type classes, where each class represents an average of feature importance values across the individual cell-type features comprising that class. **B**. Bar chart showing the feature importance values of all cell types in the random forest model for the Zeisel, *et al.* cell-type classes. **C,D**.For the four cell types with the highest feature importance (from left to right: oligodendrocytes, vascular cells, immune cells including microglia, and astrocytes), we show scatter plots depicting the relationship between scaled cell type quantification of each region and its corresponding centrality measures of the connectivity matrix: degree (sum of in- and out-degree) **C** and eigenvector centrality **D**, along with their corresponding Spearman ρ values. Two regions that were outliers were omitted from these plots for better visualization. **E**. Sagittal views of cell type densities at the voxel level as inferred by MISS for the corresponding Tasic, *et al.* cell-type classes. Please refer to **S. Data Tables 4-7** for the full cell type names and description.

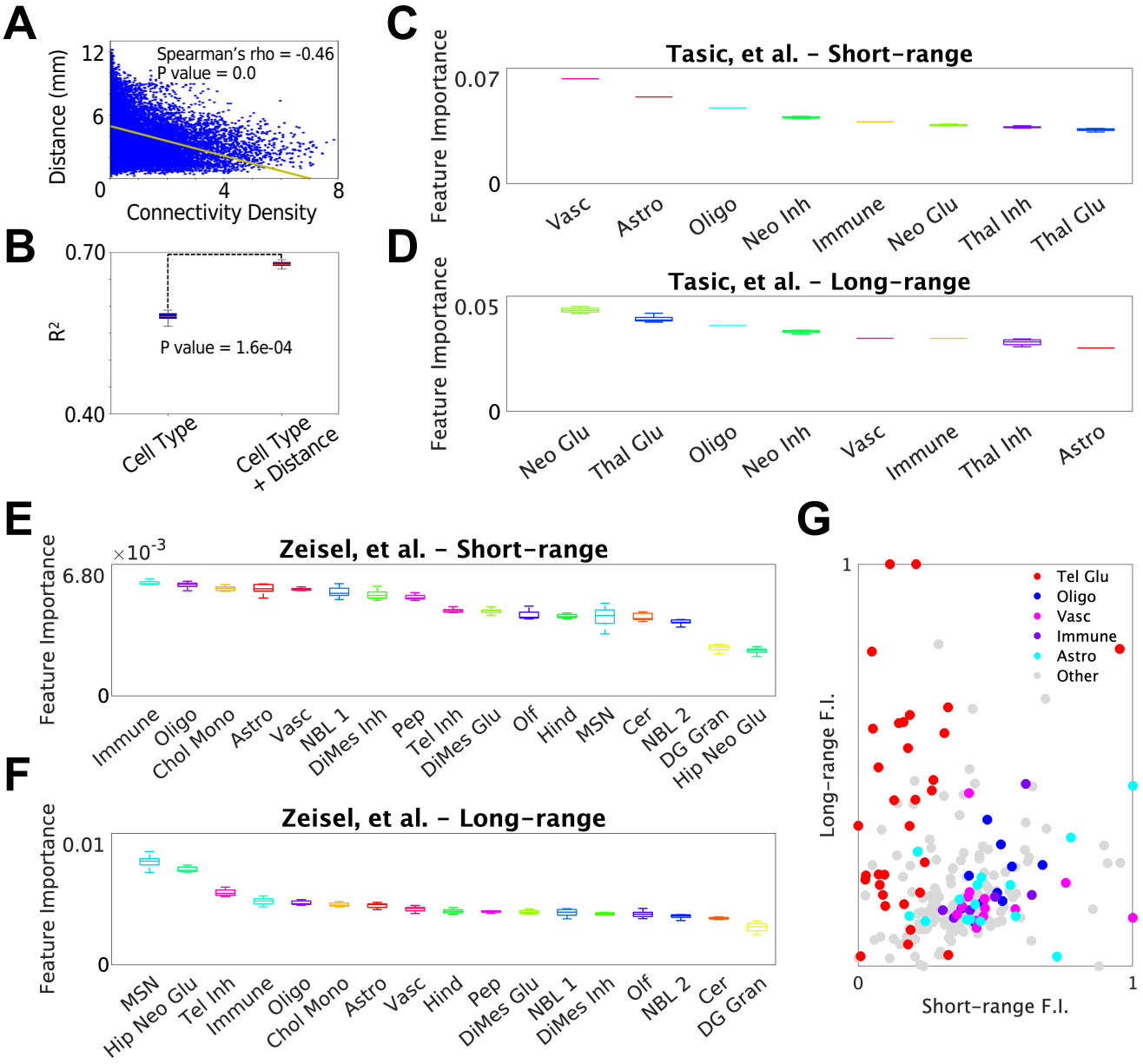

Figure 6. Most important cell type contributors vary depending on inter-regional distance. A. Scatter plot of inter-regional distance and connectivity, showing that distance has only a weak correlation with connection strength. **B**. Box plot of R^2 values following ten-fold cross-validation. *Left*: ML regression using only cell type features; *Right*: ML regression using both cell type and inter-regional distance features. **C**. Box plots showing the importance of cell-type classes in the random forest model for the lower 25th quartile of connections by distance for the Tasic, *et al.* dataset, where each class represents an average of feature importance values across the individual cell-type features comprising that class. **D**. Box plots showing the importance of cell-type features in the random forest model for the upper 75th quartile of connections by distance for the Tasic, *et al.* dataset. **E**. Box plots showing the importance of cell-type classes in the random forest model for the lower 25th quartile of connections by distance for the Zeisel, *et al.* dataset. **F**. Box plots showing the importance of cell-type features in the random forest model for the upper 75th quartile of connections by distance for the Zeisel, *et al.* dataset. **G**. Scatter plot of long-range versus short-range feature importance for all of the individual cell types within both datasets. We highlight in color the most important cell types: telencephalic glutamatergic neurons (Tel Glu; a combination of the Tasic, *et al.* Neo Glu and Zeisel, *et al.* Hip Neo Glu cell supertypes), oligodendrocyte subtypes (Oligo), vascular cell types (Vasc), immune cell subtypes (Immune), and astrocyte subtypes (Astro). Please refer to **S. Data Tables 4-7** for the full cell type names and descriptions.

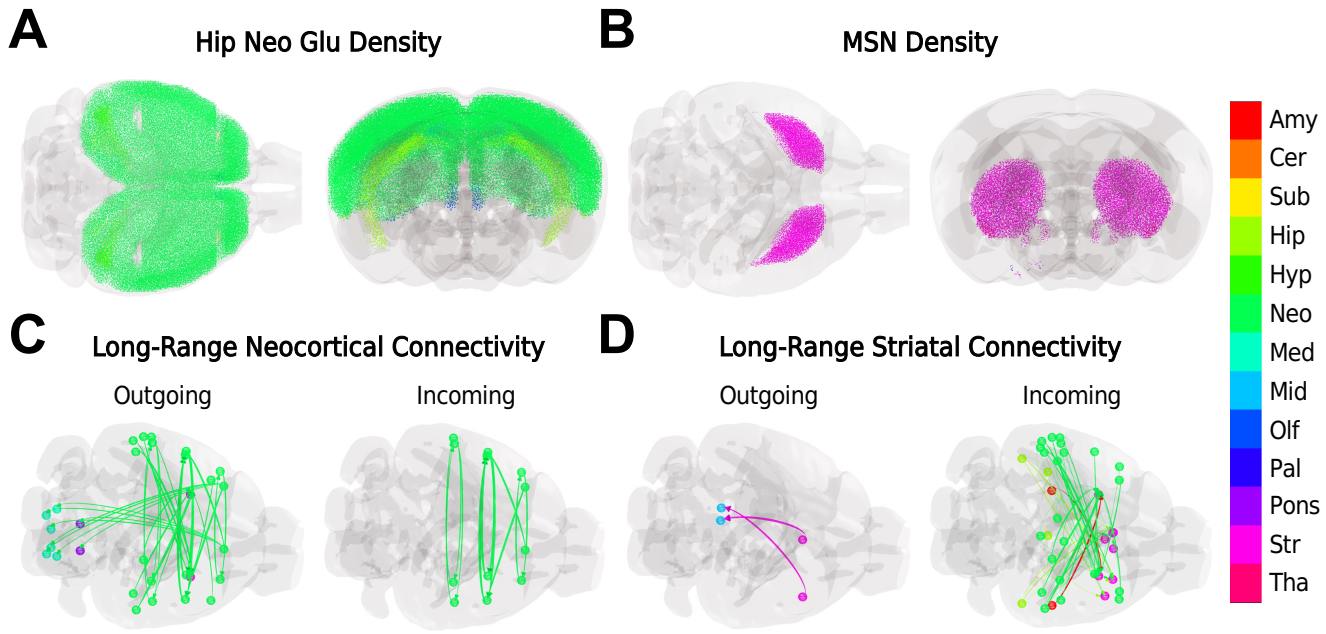

Figure 7. Distribution of top contributors to long-range connectivity (from **Zeisel, *et al.* data). A**. Glass-brain representations of the first principal component of Hipp Neo Glu neuronal distributions (number of types = 24). **B**. Glass-brain representations of the first principal component of MSN neuronal distributions (number of types = 6). **C**. Glass-brain representations of the long-range connectivity from (*Left*) and to (*Right*) neocortical regions. For clarity, only the upper 95th percentile of connections by connectivity density are depicted. **D**. Glass-brain representations of the long-range connectivity from (*Left*) and to (*Right*) striatal regions. For clarity, only the upper 50th percentile of connections by connectivity density are depicted. The colors correspond to the following major region groups: Amy – amygdala; Cer – cerebellum; Sub – cortical subplate; Hip – hippocampus; Hyp – hypothalamus; Neo – neocortex; Med – medulla; Mid – midbrain; Olf – olfactory; Pal – pallidum; Pons – pons; Str – striatum; Tha – thalamus.

Tables

	Regression		Classification	
	R^2 score	Pearson correlation	Accuracy	AUC score
Tasic, *et al.* scRNAseq dataset[18,20]	0.587	0.770	0.856	0.791
Zeisel, *et al.* scRNAseq dataset[19]	0.604	0.784	0.864	0.798

Table 1. Results summary for random forest classification