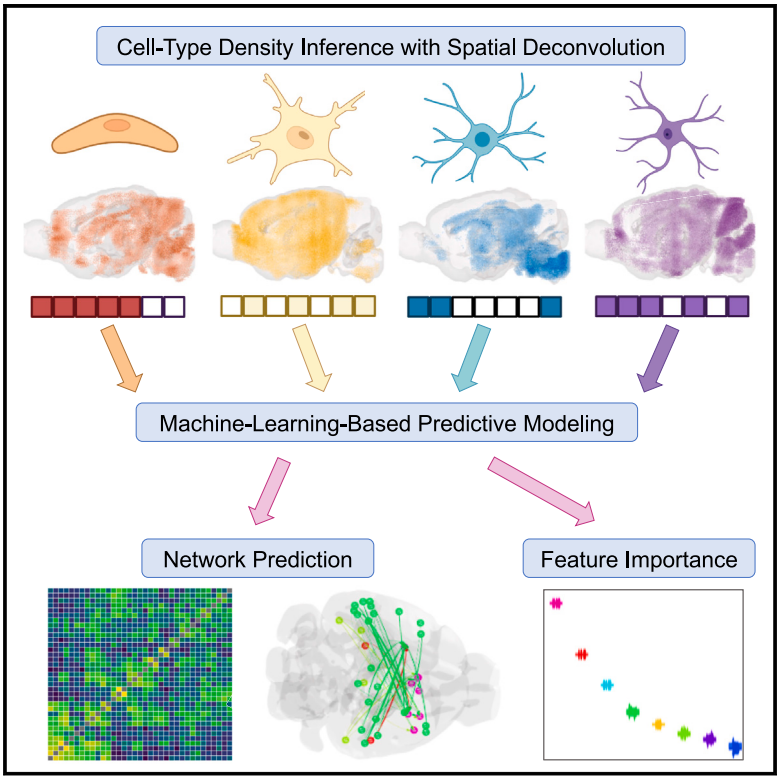
Cell Reports

Article

Spatial cell-type enrichment predicts mouse brain connectivity

Graphical abstract



Authors

Shenghuan Sun, Justin Torok, Christopher Mezias, Daren Ma, Ashish Raj

Correspondence

ashish.raj@ucsf.edu

In brief

Sun et al. use machine learning to predict whole-brain connectivity from cell-type distributions with high accuracy. They find that oligodendrocytes contribute the most information to predicting the entire connectome, while medium spiny neurons and telencephalic glutamatergic neurons disproportionately contribute to the prediction of long-range connectivity.

Highlights

- Whole-brain connectivity is accurately predicted from neural cell types with machine learning
- Feature importance analysis identifies key contributors to connectivity prediction
- Oligodendrocytes are the most important cell type for predicting connectivity overall

Sun et al., 2023, Cell Reports *42*, 113258
October 31, 2023 © 2023
https://doi.org/10.1016/j.celrep.2023.113258

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

Spatial cell-type enrichment predicts mouse brain connectivity

Shenghuan Sun,[1,3] Justin Torok,[1,3] Christopher Mezias,[2] Daren Ma,[1] and Ashish Raj[1,4,*]
[1]Department of Radiology, University of California, San Francisco, San Francisco, CA, USA
[2]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
[3]These authors contributed equally
[4]Lead contact
*Correspondence: ashish.raj@ucsf.edu
https://doi.org/10.1016/j.celrep.2023.113258

SUMMARY

A fundamental neuroscience topic is the link between the brain's molecular, cellular, and cytoarchitectonic properties and structural connectivity. Recent studies relate inter-regional connectivity to gene expression, but the relationship to regional cell-type distributions remains understudied. Here, we utilize whole-brain mapping of neuronal and non-neuronal subtypes via the matrix inversion and subset selection algorithm to model inter-regional connectivity as a function of regional cell-type composition with machine learning. We deployed random forest algorithms for predicting connectivity from cell-type densities, demonstrating surprisingly strong prediction accuracy of cell types in general, and particular non-neuronal cells such as oligodendrocytes. We found evidence of a strong distance dependency in the cell connectivity relationship, with layer-specific excitatory neurons contributing the most for long-range connectivity, while vascular and astroglia were salient for short-range connections. Our results demonstrate a link between cell types and connectivity, providing a roadmap for examining this relationship in other species, including humans.

INTRODUCTION

The structural connectome, which represents the density of physical projections between 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 (CNS).[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. There is a strong interest in gaining a rigorous measure of how gene expression and cell-type composition of brain regions relate to connectivity,[5,6] which can deepen our understanding of how brain circuits mature during the development of the CNS and how they are disrupted in neurodegenerative diseases, among other areas of inquiry.

While the correlation between regional gene expression and connectivity is well established in mice[5,7–9] and humans,[10–12] 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 unseen data.[13] As an alternative approach, Ji et al.[14] applied random forest (RF) methods to predict the presence or absence of brain connectivity from gene expression with high accuracy, but did not attempt to predict the amount of connectivity density. Other 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 implement an algorithm to produce regional cell-type enrichment from spatially resolved gene expression data following a specialized method we have recently developed called matrix inversion and subset selection (MISS).[16] This method is essentially a

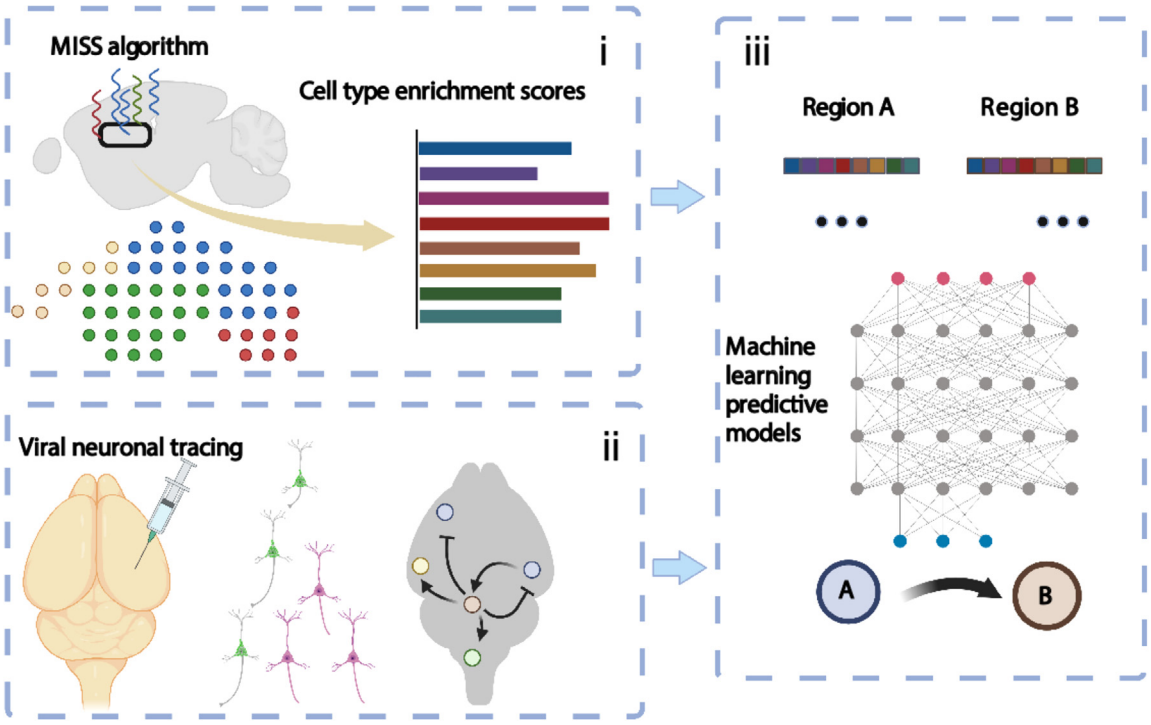


Figure 1. Study design

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 using viral neuronal tracing techniques. Right: machine learning algorithms were then implemented to predict the connectivity between each two regions.

cell-type deconvolution algorithm that was shown to faithfully reproduce cell-type distributions in the mouse brain using the Allen Gene Expression Atlas (AGEA)[17] and publicly available single-cell RNA sequencing (scRNA-seq) data.[18,19] Then, using inferred cell-type enrichment distributions as input features, we applied a number of machine learning methods to reconstruct the mesoscale mouse structural connectome from AMBCA.[2] Among all the models tested, the 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 on the connectivity prediction task, indicating that our approach is not an artifact of a particular input feature set. Our results quantitatively demonstrate that regional cell-type distributions can explain most of the variance in inter-regional connectivity.

To uncover the individual actors in this process, we undertook a thorough feature importance (FI) analysis, with both confirmatory and surprising outcomes. Strikingly, oligodendrocytes were implicated as the most important cell-type feature for recreating 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, another surprising finding. In addition, we identified a strong distance dependency in the cell connectivity relationship, with layer-specific excitatory and medium spiny neurons (MSNs) contributing most for predicting long-range connectivity, while non-neuronal cells were more 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 suggest a 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 used previously computed regional densities for 200 neuronal and non-neuronal cell types from publicly available scRNA-seq data from Zeisel et al.[19] and *in situ* hybridization data from the Allen Institute for Brain Science [17] using the MISS algorithm[16] (Figure 1i). For confirmatory analyses, we also utilized the densities

of 25 cell types from the Tasic et al. scRNA-seq dataset.[16,18,20] 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 STAR Methods). The connectivity data we attempted to reconstruct were derived from the AMBCA (http://connectivity.brain-map.org),[2] which we normalized by volume of the source region, resulting in a 424×424 matrix of normalized connection strengths (Figure 1ii; see also STAR Methods). Our choice of normalization is motivated by the observation by Abdelnour et al. and others that connectome degree is correlated with region volume[21]; therefore, we marginalized out the effect of source region volume before all analyses. As we were only interested in connectivity between disparate regions and not self-connectivity, we set all diagonal entries of the connectivity matrix to zero. 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 1iii). We also note that we considered the enrichment scores within regions sending out connections ("source") and within regions receiving connections ("target") as separate features, resulting in models with 400 total features for the Zeisel et al. dataset and 50 total features for the Tasic et al. dataset.

Predicting the existence or absence of connectivity

We first addressed whether regional cell-type enrichment features can be used to predict the existence or absence of connectivity between any given pair of regions, 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 STAR Methods). Figure S1A shows the proportions of zero and non-zero values within the ABMCA, indicating that the mouse brain connectome is approximately 64% sparse. To perform this binary classification task, we began with common unsupervised clustering methods principal-component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE). Neither approach could distinguish region pairs that form connections from those that do not (Figures 2A and 2B, respectively). However, the RF algorithm produced excellent classification results (Figures 2C and S2A; Table 1; Data S2).[22] The confusion matrix in Figure S2A shows that the RF model predicted the existence of connectivity between pairs of regions with an accuracy of 0.80 for the Zeisel et al. dataset (see also Data S2). AUROC (area under the receiving operator characteristic) and AUPR (area under precision-recall curve) values for RF were 0.87 and 0.80, respectively (Figure 2C). Thus, regional cell-type enrichment profiles can predict the presence of connectivity, paralleling prior findings based on gene expression.[14]

Predicting connectivity density

We next turned to the task of predicting the connectivity density,[2,15,23,24] which we define to be a measure proportional to the number of axonal tracts per unit of source region volume between any region pair. We first examined whether region pairs with similar cell-type compositions were likely to be more densely connected. Figures 2D and 2E (left panels) depict heat maps of the ipsilateral regional cross-correlation matrix with respect to cell-type enrichment scores and the mouse connectome, respectively (see also Figure S3). While there is a degree of visual similarity, the two measures are only weakly correlated (Pearson's R = 0.32; Figure 2F). 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 is related to, but insufficiently predictive of, the whole-brain connectome.

Given that the connectivity density distribution is mostly comprised of very small values with a number of prominent outliers (Figure S1B), we hypothesized that nonlinear predictive models would be more appropriate. Similar to the binary classification task, we found that the RF model recreated connectivity from cell-type enrichment with a high degree of accuracy (adjusted R^2 = 0.60, root mean-square deviation = 0.60, 10-fold cross-validation; Table 1; Data S3). Excellent visual similarity between the connectivity predicted by RF using cell types and the ground truth can be observed in matrix heatmaps (Figure 2E, right) and scatterplots (Figure 2G), with a Pearson's correlation of 0.79.

To more thoroughly explore the significance of these results, we constructed five collections of randomly generated null models, each of which had the same number of input features as the Zeisel et al. dataset (i.e., 200 each for source and target region). Figure 2H displays distributions of R^2 values from each type of null distribution representing 500 random model instances, and the red vertical line indicates the performance of the cell-type-based model (Table 1; see also STAR Methods). As expected, the least informative models incorporate no gene-expression information. The purely random models (purple curve), which involved assigning regional cell-type enrichment scores from a uniform random distribution, were completely uninformative. When these regional values were randomly assigned using distributions whose means depended on the anatomical parcel to which each region belonged (green curve; see also STAR Methods), the predictions improve markedly, reflecting key biology of the anatomical relationships between regions, but remain poor. Performance further improves when scrambling the values of the AGEA before applying MISS on the highly informative MRx3 gene subset (yellow curve; see STAR Methods and Mezias et al.[16] for details), but it is much lower than the true cell-type distributions. We also explored the performance of gene expression directly with two different sampling methods: (1) randomly selecting 200 genes within the 4,083-gene AGEA (red curve) and (2) randomly selecting 200 genes within the 1,360-gene MRx3 subset (blue curve). The model using cell-type features significantly outperforms those using fully random gene sampling, indicating that cell types contain key information for predicting connectivity that is not uniformly reflected in the expression of individual genes. We achieved comparable prediction accuracy using subsets of informative MRx3 genes and cell types; given that MRx3 specifically selects genes based on how well they discriminate between cell types transcriptomically,[16] the agreement between these two types of input features is expected.

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

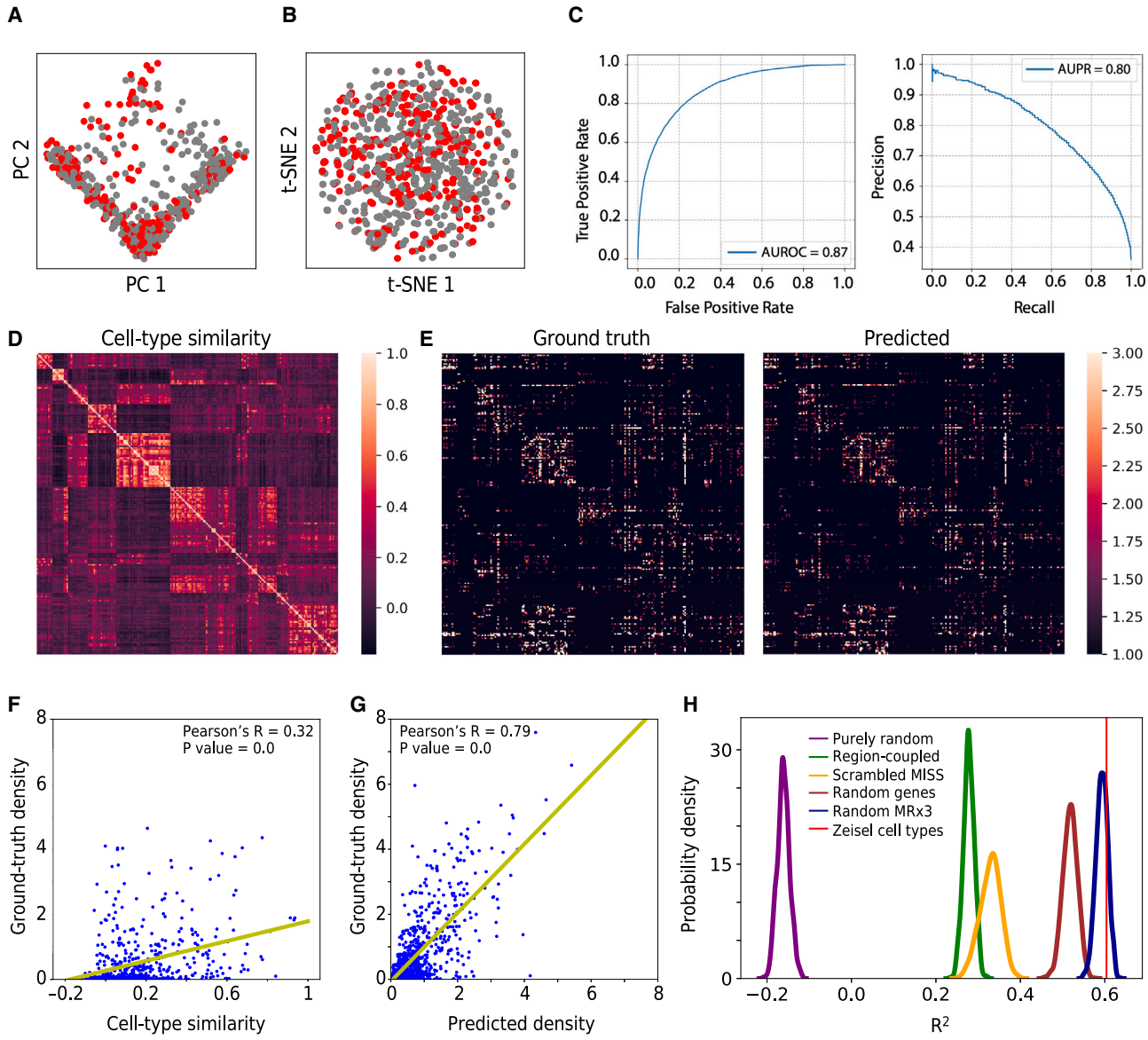


Figure 2. Machine learning applied to regional cell-type distributions predicts both the existence of connectivity and connectivity density

(A) Principal-component analysis of the cell-type spatial quantification array.

(B) 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) Performance evaluation of the classifier model using 10-fold cross-validation. Left: the receiver operating characteristic curve (AUROC = 0.87). Right: the precision recall curve (AUPR = 0.80).

(D) Cellular similarity matrix (quantified using Pearson correlation) of spatial cell-type enrichment quantification across brain regions, ipsilateral only.

(E) 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.

(F) Scatterplot of pairwise cellular similarity (as depicted in D) between two regions' cell-type distribution vectors versus the log-transformed connectivity strength between the two regions (as depicted in E, left), and the fitted linear regression curve (Pearson's R = 0.32, p = 0.0).

(G) Scatterplot 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 = 0.79, p = 0.0).

(H) Distributions of R^2 values from null models using five types of inputs in the figure below, each with the same number of features as the Zeisel et al. dataset (i.e., 200): purely random white noise (purple); region-coupled white noise (green); cell-type "distributions" obtained from MISS after scrambling the regional gene expression values in the AGEA (yellow); randomly selected genes from the 4083 gene AGEA (red); and randomly selected genes from the 1,360-gene "high-information" subset used to infer the Zeisel et al. cell-type distributions in MISS (blue).[16] The red vertical line indicates the performance of the cell-type-based model presented in the manuscript. Each distribution represents 500 random model instances.

Table 1. Random forest model performance

Dataset	Model task	Connectome subset	Accuracy	R^2
Zeisel et al.[19]	kernel density estimate	all	0.864	–
	kernel density estimate	all	–	0.604
	kernel density estimate	short range	–	0.614
	regression	long range	–	0.577
	regression	neocortical to/from other	–	0.585
Tasic et al.[18,20]	classification	all	0.856	–
	regression	all	–	0.587
	regression	short range	–	0.608
	regression	long range	–	0.581

Summary of results for the different random forest models explored, broken up by cell-type dataset (Zeisel et al. or Tasic et al.), model task (classification to predict binary connectivity or regression to predict connectivity density), and connectome subset (portion of the connectivity matrix being predicted). We report accuracy for classification models and mean R^2 values for the regression models. See also Data S2–S6.

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 an RF algorithm to predict connectivity density in the AMBCA without first removing unconnected region pairs (Figure S4). 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 was 0.42. We also achieved strong performance (R^2 = 0.50) when we split our training and test sets by source region rather than purely randomly (Data S4). Several other common machine learning algorithms were implemented to reconstruct both the binary connectome and predict connectivity density, which, however, fail to achieve superior performance over RF (Data S2 and S3).

Confirmation with an independent cell-type dataset

We tested whether the RF 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 scRNA-seq dataset from Tasic et al., which sampled 25 cell types within the mouse neocortex and thalamus.[18,20] A natural question to ask is whether the lack of sampling outside of the neocortex and thalamus may bias the whole-brain predictions of cell-type density from this dataset. To address this concern, we have previously shown that the prediction error within unsampled regions is comparable with that within sampled regions (reproduced from Mezias et al. in Figure S5).[16] t-SNE and PCA were also unable to separate region pairs that share a connection from those that do not using the Tasic et al. dataset (Figures S6A and S6B). But, when we used this less-expansive set of cell types, we were still able to produce an accurate recreation of the binarized connectome (AUROC = 0.85, AUPR = 0.78; Table 1; Figure S6C—only a modest decrease from the 200-type Zeisel et al.-derived results [Figure 2C; Table 1]). The matrix of cell-type similarity is, again, only weakly correlated to the connectome (Pearson's R = 0.21, p = 0.0; Figures S6D–S6F and S7).

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; Figure S8), which we expected given the reliability of the MISS algorithm. The machine learning models were similarly successful in predicting the connectome with the Tasic et al. dataset, only modestly underperforming relative to the 200-type Zeisel et al. dataset (Table 1; Figure S6H; Data S5 and S6). Notably, only RF was able to perform both the classification and regression tasks successfully (Data S5 and S6), reinforcing that RF is uniquely suited to this problem.

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 that give outputs whose dependencies are difficult to discern, RF models are amenable to FI analysis[22,25] (see also STAR Methods). FI can be thought of as a measure of how much information is contributed by a given feature relative to all other features. Therefore, for each RF model we determined the importance of each cell-type feature, and grouped them by "supertype" as determined by their scRNA-seq-based taxonomies. Please refer to Data S7–S10 for the list of cell-type names and the supertypes to which they belong. We show these as boxplots for the Zeisel et al. connectivity density RF model in Figures 3A–3C, where each data point represents the average FI score for each cell type across the 10 cross-validation test sets. We considered the salience of each cell type in terms of its source region (Figure 3A) and target region (Figure 3B) FI, as well as its overall salience as an average of source and target region FI scores (Figure 3C). The corresponding results for the Tasic et al. connectivity density RF model (Figures S6E and S6G) and the classification RF models (Figures 2C and S6C) are shown in Figures S9A–S9C and S10, respectively. Overall, we found that that oligodendrocytes were the most important contributors to both binary connectivity and connectivity density prediction at the whole-connectome level for both the Zeisel et al. and Tasic et al. datasets (Figures 3C, S9C, and S10). On a more granular level, the source region cell-type FI scores strongly resembled the averaged values, with oligodendrocytes again having the highest scores in both the Zeisel et al. and Tasic et al. datasets (Figures 3A and S9A). However, when considering only the target regions' cell-type compositions, we found that a number of neuronal cell types had higher FI scores than oligodendrocytes, with MSNs being a notable outlier for Zeisel et al. (Figure 3B). We found qualitatively similar results when we retrained the RF model to predict the connectivity densities from neocortical to non-neocortical regions and vice versa (Figure S11). We elaborate upon the implications of the divergence between source and target cell-type compositions in the discussion.

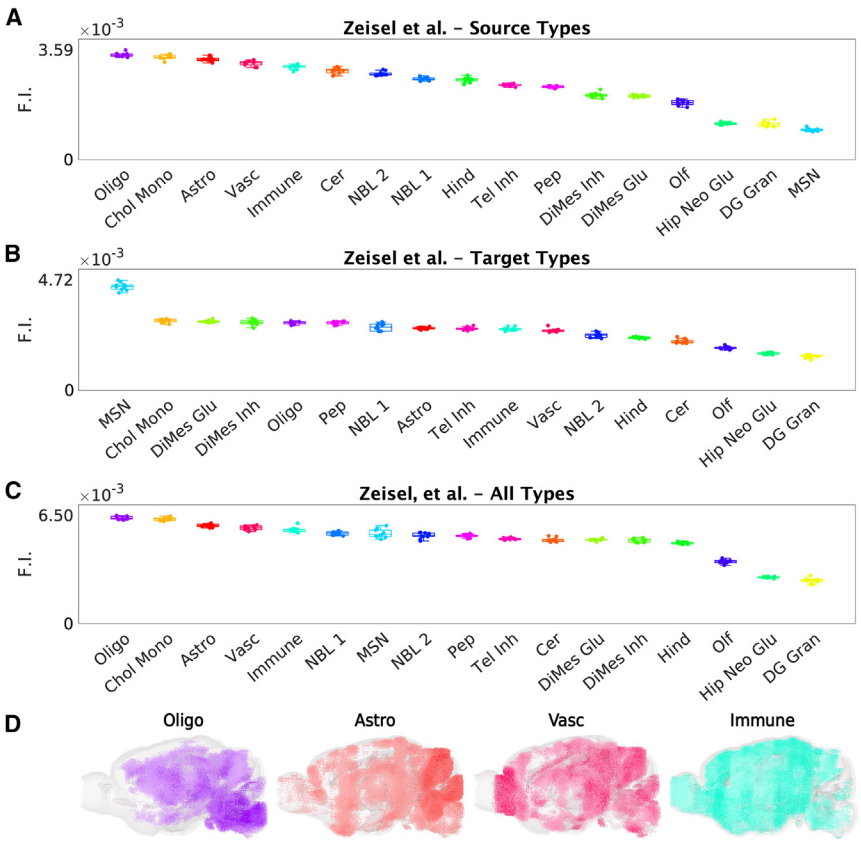


Figure 3. Interrogating the individual contributions of cell types

(A) Boxplots showing the feature importance (FI) values of all source region cell-type features in the random forest model for Zeisel et al. cell-type classes, with the standard error of the mean (SEM) computed as the average across 10-fold cross-validation. Cell types, grouped by supertype.

(B) FI values for target region cell-type features, with the SEM computed as the average across 10-fold cross-validation.

(C) FI values for all cell-type features, with the SEM computed as the average across 10-fold cross-validation.

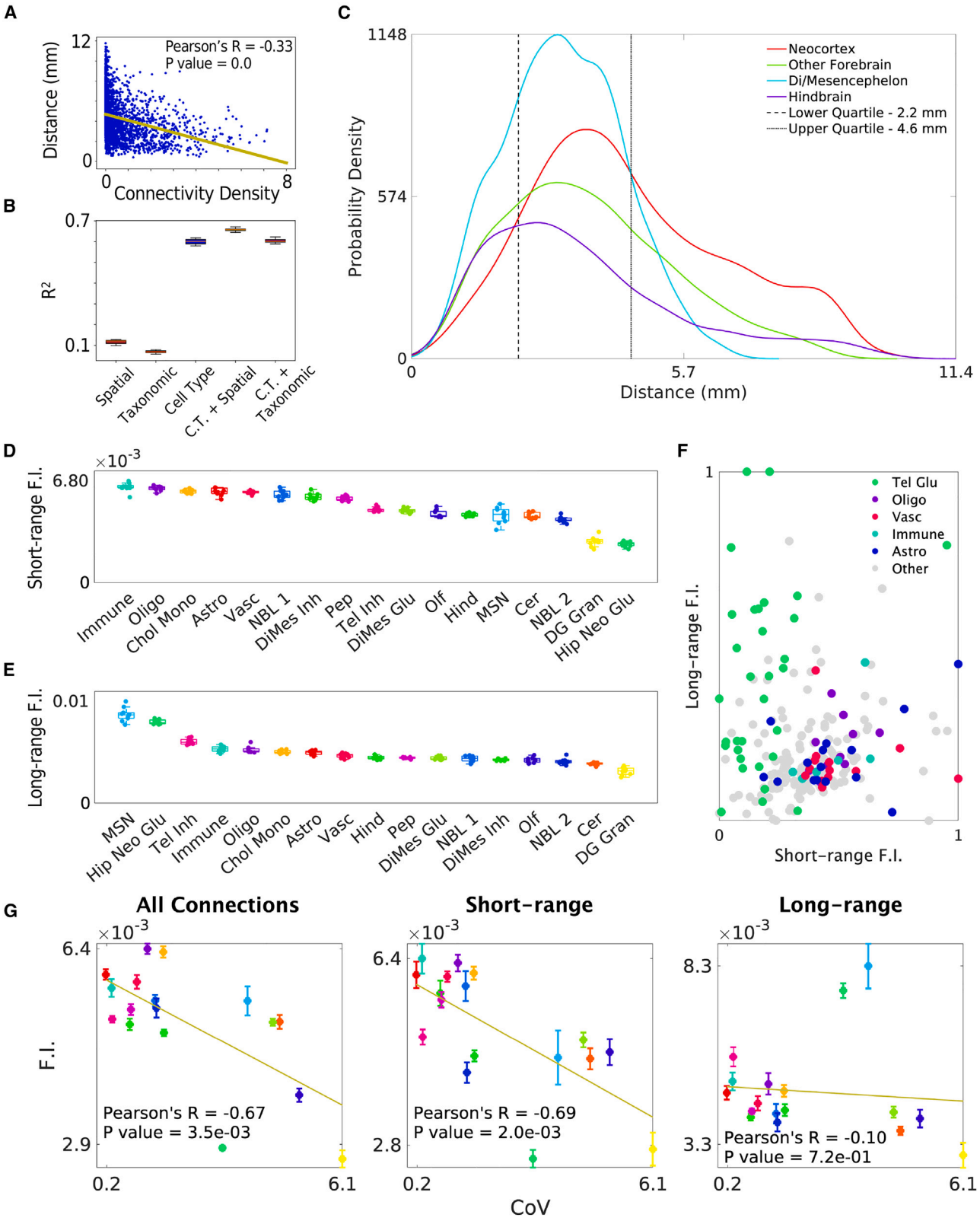
(D) Sagittal views of cell-type densities at the voxel level as inferred by MISS for the corresponding Zeisel et al. cell-type classes. Please refer to Data S7–S10 for the full cell-type names and description.

More generally, the non-neuronal supertypes were more salient in the RF models than neuronal supertypes. We show the voxel-wise distributions of these non-neuronal Zeisel et al. and Tasic et al. supertypes in Figures 3D and S9D, respectively. Overall, the apparent consistency of these FI results between the two independently curated scRNA-seq datasets suggests a true biological connection between these non-neuronal support cells and connectivity at a whole-brain level.

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 (Pearson's $R = -0.33$, $p = 0.0$; Figure 4A), indicating that there is a bias toward short-range connections in the mouse brain. Using spatial distance as a sole predictor of connectivity density produced an RF model with an average R^2 of 0.12, indicating that distance contributes modest but significant information (Figure 4B). Furthermore, including it along with the cell-type distributions produced RF models with higher R^2 values ($\Delta R^2 = 0.09$; Figure 4B). By contrast, using the taxonomic distance matrix as a predictor, where distance is defined in terms of how early each region pair separated anatomically during development,[26] contributed less information than spatial distance and did not provide an improvement over cell-type enrichment scores (Figure 4B; see also STAR Methods). These results indicate that inter-regional spatial distance contributes information that is at least partly independent of that contributed by regional cell-type composition, while the information from taxonomic distance is fully captured by differences in regional cell-type composition. When we looked at the distance dependence of connectivity density within major anatomical region groups, we found that each set of regions generally has a broad distribution of connection lengths (overall interquartile range = [2.2 mm, 4.6 mm]; Figure 4C). However, while each distribution is left-skewed, indicating that shorter-ranged connections predominate, we found that neocortical regions mediate a disproportionate number of the long-range connections in the brain.

Consequently, we were interested in whether there was a distance dependence to cell-type FI, as has been suggested previously.[27,28] We therefore trained the RF algorithm on the upper and lower quartiles of connections by distance separately and determined the FI scores per cell supertype as above (Figures 4D, 4E, and S12–S14). The RF models achieved similar fits regardless of distance bin ($R^2 = 0.61$ and 0.58 for short-range and long-range connectivity, respectively) and performed comparably well with the model of whole-brain connectivity (Table 1). However, clear differences emerged at the level of FI between short-range and long-range connectivity. Although oligodendrocyte distributions from the Zeisel et al. and Tasic et al. datasets were not the strongest contributors to the RF model of short-range connectivity as they were for whole-brain connectivity, they remained among the top features, and generally non-neuronal cells had stronger source- and target-averaged FI scores than neurons, as above (Figures 4D and S12A). In particular, immune cells and vascular cells exhibited the strongest contributions to short-range connectivity for the Zeisel et al. and Tasic et al. datasets, respectively. Of the neuronal



(legend on next page)

supertypes, forebrain glutamatergic neurons (Neo Glu, Thal Glu, Hip Neo Glu) had particularly weak FI 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 (Figures 4E and S12B). As with the target region cell-type FI analysis for Zeisel et al., the supertype with the highest FI score was striatal MSNs, which are unique to that dataset (Figure 4E). We summarize these results in Figure 4F, which shows that, for both the Zeisel et al. and Tasic 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 over-represented 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.

A more nuanced picture emerged when we considered the source and target region cell-type contributions to short- and long-range connectivity prediction separately (Figures S13 and S14). The contributions of individual source and target region non-neuronal cells were variable; as a class, they generally exceeded neuronal supertypes when considering only source region supertypes in predicting short-range connectivity. In other words, consistent with the above analyses, while non-neuronal contributions predominated when considering overall connectivity prediction (Figures 3C and S9C), this was driven predominantly by their source region FI scores and the prediction of shorter-distance connectivity densities. The MSN and telencephalic glutamatergic supertypes also exhibited interesting trends when separating source and target region features. As mentioned above, MSN was the strongest contributor among target region features to overall connectivity prediction (Figure 3B) and among source- and target-averaged features to long-range connectivity prediction (Figure 4E). However, we found that, among only target region features, MSN was in fact was the strongest contributor to both short- and long-range connectivity prediction (Figures S13B and S13D) and did not strongly contribute as a source region feature to long-range connectivity prediction (Figure S13C). For both the Zeisel et al. Hip Neo Glu and Tasic et al. Neo Glu supertypes, there was no effect of separating out source region from target region supertype features, providing similarly weak contributions to short-range connectivity prediction (Figures S13A, S13B, S14A, and S14B) and similarly strong contributions to long-range connectivity prediction (Figures S13C, S13D, and 14C–14D). In summary, while MSNs and telencephalic glutamatergic neurons both disproportionately contributed to predicting long-range connectivity, the contributions between source and target region enrichment scores markedly differed between them.

To further examine the underpinnings of the discrepancy between the supertypes most critical for predicting short-range and long-range FI, we examined whether there was a relationship between how variably distributed the Zeisel et al. supertypes were across the brain and FI. We hypothesized that more spatially homogeneous cell types would contribute less to the RF model's predictiveness. As shown in Figure 4G, we indeed found that, for the RF models predicting all connectivity (left panel) and short-range connectivity (center panel), there was a statistically significant negative association between each supertype's average FI score and its spatial coefficient of variation (CoV), with Pearson's R values of -0.67 ($p = 3.5 \times 10^{-3}$) and -0.69 ($p = 2.0 \times 10^{-3}$), respectively. However, while the association trended negative for the long-range RF model (Figure 4G, right panel), it was weak and not statistically significant (Pearson's $R = -0.10$, $p = 0.72$). The two outliers with especially high long-range FI scores, MSN and Hip Neo Glu, have intermediate CoV values (Figure 4G, right panel), which agrees with their distributions being highly specific to a relatively large set of regions (Figures 5A and 5B). Therefore, we conclude that the contributions of supertypes to long-range connectivity density predictions in particular cannot be simply explained by spatial heterogeneity.

Neuronal contributions to long-range connectivity

To explore some of the relationships between cell-type distributions and connectivity qualitatively, we show the distributions of

Figure 4. Most important cell-type contributors vary depending on inter-regional distance

(A) Scatterplot of inter-regional distance and connectivity, showing that distance has a weak correlation with connection strength.

(B) Boxplots of R^2 values following 10-fold cross-validation using different combinations of input features. From left to right: spatial distance matrix, taxonomic distance matrix, cell-type enrichment scores, cell-type enrichment scores with spatial distance, cell-type enrichment scores with taxonomic distance.

(C) Kernel density estimate plot of the probability of two regions being connected as a function of inter-regional distance. The individual lines represent the subregions comprising the neocortex (red), the combination of subregions within the amygdala, cortical subplate, hippocampal formation, olfactory bulb, pallidum, and striatum (green), the combination of subregions within the hypothalamus, thalamus, and midbrain (cyan), and the combination of subregions within the cerebellum, pons, and medulla (purple). Interquartile range is shown with the black dashed and dotted lines.

(D) Boxplots 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, with the SEM computed as the average across 10-fold cross-validation.

(E) Boxplots 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, with the SEM computed as the average across 10-fold cross-validation.

(F) Scatterplot of long-range versus short-range 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).

(G) Scatterplots of the Zeisel et al. supertype for all connections (left), short-range connections only (center), and long-range connections only (right), regressed against the regional coefficient of variation (CoV) of the cell-supertype densities. There is a strongly negative and statistically significant negative relationship between FI and CoV for all connections and short-range connections, but not long-range connections. Please refer to Data S7–S10 for the full cell type names and descriptions.

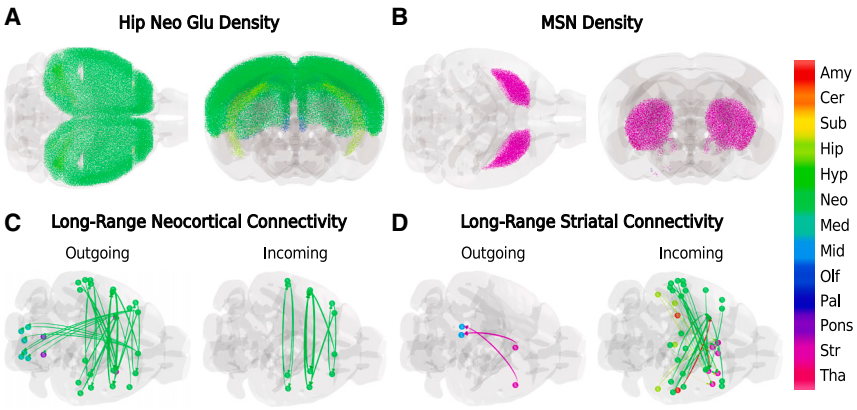


Figure 5. Distribution of top contributors to long-range connectivity (from Zeisel et al. data)

(A) Glass-brain representations of the first-principal component of Hip 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 (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 (right) striatal regions. For clarity, only the upper 50th percentile of connections by connectivity density is 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.

Hip Neo Glu and MSN, the two supertypes from the Zeisel et al. dataset with the highest average FI for predicting long-range connectivity (Figures 5A and 5B). The Hip Neo Glu supertype comprises 24 individual cell types, all of which are excitatory and located within neocortical and hippocampal regions, and the MSN supertype comprises 6 types of striatal MSN.[19] 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 supertypes, we also show the strongest long-range connections to and from the neocortex (Figure 5C) and the striatum (Figure 5D). More specifically, for the neocortex, these include projections to hindbrain nuclei and contralateral neocortical-neocortical connections (Figure 5C). 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 5D). 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 practical applications of data-driven machine learning models for reconstructing whole-brain inter-regional connectivity using spatial cell-type enrichment distributions. We split this reconstruction into two tasks: a classification task to predict the existence and absence of connections between each region pair, and a regression task to predict the values of connectivity density between all connected region pairs. We find that, using the comprehensively sampled Zeisel et al. cell-type distributions,[16,19] RF models are able to perform both tasks with a high degree of accuracy (Figure 2; Table 1), which we replicate using the smaller Tasic et al. dataset (Figure S6; Table 1).[16,18,20] Post hoc FI analyses implicate oligodendrocytes as especially critical in correctly recreating the whole-brain connectome (Figures 3 and S9). We further consider inter-regional distance as an important predictor of the density of brain connectivity (Figure 4). When FI is evaluated separately for short-range versus long-range connections, we find that MSNs and telencephalic glutamatergic neurons appear to be far more important for recreating long-range connectivity than for short-range connectivity, while non-neuronal cell types are 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 machine learning prediction tasks by separately predicting the absence or presence of a connection and the connectivity density between any given region pair for two reasons. First, the connectivity data are quite sparse (36% nonzero region pairs), which can significantly impact 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] 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 (Figure S4).

Model performance is replicated across two different scRNA-seq datasets

We were able to replicate the results of our primary dataset—the 200-type Zeisel et al. dataset[19]—using a separate, 25-type dataset from Tasic et al.[18,20] (Figures 2 and S6; see also STAR 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. This observation is supported by the results from the RF models using subsets of genes (Figure 2H, red and blue curves), whose performance also did not exceed that of either cell-type model. 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 representing only a subset of mouse neuronal diversity, provide close to maximal information content for reconstructing brain connectivity. For example, the four non-neuronal supertypes (astrocytes, oligodendrocytes, immune cells, and vascular cells) from the two datasets are qualitatively very similar in spatial distribution (Figures 3D and S9D) and consistently have higher FI scores than most neuronal supertypes (Figures 3C and S9C). Furthermore, for the more regionally specific long-range connections (Figure 4C), both datasets have robust supertypes of the telencephalic glutamatergic neurons that were especially important in reconstructing the long-range connectome (Figures 4E and S13B). Nevertheless, that we were able to create models with high predictive accuracy with two sets of cell-type enrichment scores coming from independently sampled scRNA-seq 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. 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 this 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 and Pavlidis, 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 RF 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

We propose here that cell type features are a valuable alternative to gene expression 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, particularly for the confirmatory Tasic et al. dataset, was considered a means of avoiding these pitfalls. Compared with the thousands of gene features used in prior studies,[5,14] the sets of 25 and 200 cell types should form a more parsimonious input feature space. That being said, tremendous effort has been invested in obtaining gene profiles of cell-type-specific marker genes as well as genes involved in processes related to the formation and maintenance of projections between brain regions. Our work both complements those efforts and also shows that we can obtain cellular signatures from genes using the MISS algorithm that are not necessarily single-cell markers but nevertheless contribute significant information for predicting connectivity.

Oligodendrocytes are disproportionately associated with whole-brain connectivity patterns

Our analysis demonstrated the importance of oligodendrocyte cell types in recreating the whole-brain connectome. Oligodendrocytes are the most predictive feature in the RF model for both datasets (Figures 3A–3C and S9A–S9C), and are also among the highly predictive features when analyzing the FI for the classification task (Figure S10). Biologically, oligodendrocytes produce the myelin sheath insulating neuronal axons.[29,30] They help protect the vulnerable axons from parenchymal chemokines and cytokines, and ensure the fast and efficient movement of action potentials.[29–31] Dysfunction of oligodendrocytes can interfere with normal micro-structure and functional connectivity in the mouse brain.[32] Oligodendrocyte myelination was also shown in previous work to be able to regulate the loss of synapses.[33] Moreover, recent work from Buchanan et al. showed that oligodendrocyte precursor cells can prune axons in the mouse neocortex.[34] 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 for the Zeisel et al. dataset (Figures 4D and 4E). 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

Non-neuronal cell types also had high FI and we highlight them below. Brain vascular cells compose the blood-brain barrier, which protects the vulnerable CNS, and they interact with the CNS for supporting neuronal cells with nutrients, energy, and oxygen.[35–39] Their breakdown is strongly correlated with brain connectivity disruption and cognitive defects.[35,39] Brain endothelial cells are involved in the process of neurovascular coupling,[40,41] whereby local neural activity stimulates subsequent blood flow changes in the corresponding downstream locations.[41,42] 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 with neuronal cell types. Previous studies have indicated that there is an association between inflammation and functional brain connectivity.[43,44] 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.[45]

Furthermore, we found that non-neuronal cell types contribute disproportionately to predicting short-range connectivity. Of these, immune cells were the most important supertype for the Zeisel et al. dataset and vascular cells were the most important supertype for the Tasic et al. dataset, although all non-neuronal supertypes tended to have higher FI scores than most of the neuronal supertypes (Figures 4D and S12A). There are multiple reasons why these non-neuronal cells have higher FI 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 (Figures 3A–3C and S9A–S9C). However, given that FI 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; Figures 4E and S12B). The distance dependence of cell-type contributions to connectivity is an important line of inquiry for future studies.

Neuronal subtypes differentially mediate long-range connectivity

In addition to oligodendrocytes, we found that telencephalic glutamatergic neurons and striatal MSNs were among the most salient classes of cell types, but only for predicting long-range connections (Figures 4E and S12B). The former are well known to project to remote locations within and outside of the neocortex (Figures 5A and 5C), 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 Zeisel et al. and Tasic et al. datasets (Neo Glu and Hip Neo Glu, respectively) are also among the least important features for predicting short-range connectivity (Figures 4D and S12A), suggesting that these neurons predominantly engage in long-range connections. Similarly, the high FI of MSNs is concordant with their function, as these are long-range-projecting, inhibitory neurons. MSNs 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.[46] When we look at the FI of individual cell types between the two datasets, we see a similar pattern as we do at the supertype level (Figure 4G). In particular, telencephalic glutamatergic neurons contribute weakly to predicting short-range connectivity and are overrepresented among types with high FI 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.

One interesting difference in the ways in which these two classes of cell types contribute to connectivity density prediction emerges when examining the contributions of source and target region cell-type features separately. The Neo Glu and Hip Neo Glu supertypes were disproportionately informative for predicting long-range connectivity when considering either source or target (Figures S13C, S13D, and 14C–14D), whereas the target region MSN supertype had more relative importance for predicting both short- and long-range connectivity and was only moderately informative as a source region feature (Figure S13). As shown in Figure 5A, the distribution of Hip Neo Glu is entirely telencephalic; these regions are involved in a disproportionate fraction of long-range connections (Figure 4C), the strongest of which tended to be contralateral and intra-neocortical (Figure 5C). That source and target region FI values were both high for Hip Neo Glu reflects the intra-cortical nature of these connections. By contrast, the striatum, and caudoputamen in particular, have many more incoming long-range connections than outgoing long-range connections (Figure 5D), and therefore there should be a large difference between source and target region FI values for MSNs, which we observed (Figures S13B and S13D). 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 FI analysis will benefit from further experimental work to elucidate in more detail the biological roles of the identified cell types with respect to connectivity.

Future directions

One extension of the current method would be to apply feature selection on either of the cell-type datasets used here, which may facilitate the development of more predictive models. In addition, machine learning models that integrate both cellular features and anatomic/morphological features can be expected to improve current predictions. 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 future research but will require higher-resolution data. Given the conservation of CNS properties in mammals, we may also be able to apply these data-driven methods to the human brain.

Limitations of the study

The primary limitation of the current work is that cell-type enrichment does not accommodate other factors critical for determining brain connectivity, including neural polarity, cell maturation, and migration. Furthermore, despite their ability to produce FI, RF models are less interpretable than generalized linear models. RF models, an ensemble of decision trees, can also suffer from overfitting, since any constituent decision tree may be sensitive to data variations and 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.

Several caveats are worth mentioning in regard to the input features used here. First, we used the coronal series of the AGEA, which contains far fewer unique genes (4,083) than the sagittal series and has a neuron and hippocampal bias[17] for the MISS pipeline and the null models of Figure 2H. The coronal series, however, has a superior spatial resolution of 200 μm; ultimately, we decided that higher accuracy in regional quantification of gene expression was more important than the limitations inherent to the gene set. Similarly, our choice to use cell densities inferred using the MISS algorithm was motivated by its comprehensive spatial coverage. Within the MISS pipeline, we apply a gene selection algorithm called MRx3 to filter out thousands of uninformative genes for the purpose of reconstructing cell-type densities,[16] so having a more expansive gene set may not necessarily lead to significantly better predictions. However, we note that several promising technologies are emerging that have demonstrated single-cell-level resolution of brain tissue, such as STARmap,[47] osmFISH,[48] and merFISH.[49,50] While the spatial resolution and direct transcriptomic mapping of cell types using these methods is impressive, they have not yet been scaled up beyond single regions. More recent work using BARseq mapped approximately 1.2 million individual cells within the mouse forebrain and labeled them using 107 marker genes[51]; however, the authors biased their sampling toward neocortical glutamatergic neurons and so this dataset lacks the breadth of transcriptomic diversity captured within the Zeisel et al. dataset used here. Therefore, for exploring the architecture of the whole-brain connectome at a mesoscopic scale as it relates to cell-type distributions, we chose to use MISS for its breadth of spatial coverage and amount of cell-type diversity. Many questions about whole-brain microarchitecture, which would require mapping cell types and projections at a single-cell level to answer, remain the subject of future work in this area.

Conclusions

We report a data-driven approach that successfully predicts whole-brain connectivity from regional cell-type information in the mouse brain. We report quantitative evidence of the vital importance of interareal distance and non-neural cell types in recreating connectivity, especially of oligodendrocytes and other non-neuronal cell types. Our results may provide guidelines for future experimental analysis, and can be extended to other mammals, including humans.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - MISS-derived cell type features
  - Mouse connectivity
  - Machine learning methods
  - Cell type input features
  - Null model input features
  - Random forest
  - Other ML models
  - Neural network models
  - Model performance evaluation
  - 3D brain visualization
  - Inter-regional distance matrix calculation
  - Feature interpretation from random forest models
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2023.113258.

ACKNOWLEDGMENTS

This work was supported by the following NIH grants: R01NS092802, RF1AG062196, and R01AG072753.

AUTHOR CONTRIBUTIONS

A.R. and S.S. conceived of the presented idea of the study. A.R. developed the theory and main experiment design. S.S. performed the early machine learning experiments and generated early results with the help of D.M. C.M. and J.T. verified the analytical methods. J.T. further introduced MISS as an important analytical tool that expanded the scope of the current research. S.S. performed data collection and built the machine learning pipeline and J.T. implemented the statistical analysis and contributed in feature selection. All authors discussed the results and contributed to the final manuscript. J.T. and S.S. took lead in the manuscript writing and figure generation. D.M., C.M., and A.R. polished the manuscript; each of them has provided important advice based on their domain knowledge. A.R. supervised the project's progress and provided necessary guidance on general writing and submission.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: December 12, 2022
Revised: June 7, 2023
Accepted: September 28, 2023
Published: October 19, 2023

REFERENCES

1. Sporns, O., Tononi, G., and Kötter, R. (2005). PLoS Comput BiolThe human connectome: A structural description of the human brain. PLoS Comput. Biol. *1*, e42.
2. Oh, S.W., Harris, J.A., Ng, L., Winslow, B., Cain, N., Mihalas, S., Wang, Q., Lau, C., Kuan, L., Henry, A.M., et al. (2014). A mesoscale connectome of the mouse brain. Nature *508*, 207–214. https://doi.org/10.1038/nature13186.
3. Bullmore, E.T., and Bassett, D.S. (2011). Annu Rev Clin PsycholBrain graphs: graphical models of the human brain connectome. Annu. Rev. Clin. Psychol. *7*, 113–140.
4. Zeng, H. (2018). Mesoscale connectomics. Curr Opin NeurobiolMesoscale connectomics. *Curr Opin Neurobiol 50*, 154–162.
5. French, L., and Pavlidis, P. (2011). Relationships between gene expression and brain wiring in the adult rodent brain. PLoS Comput. Biol. *7*, e1001049. https://doi.org/10.1371/journal.pcbi.1001049.
6. Tan, P.P.C., French, L., and Pavlidis, P. (2013). Neuron-enriched gene expression patterns are regionally anti-correlated with oligodendrocyte-enriched patterns in the adult mouse and human brain. Front. Neurosci. *7*, 5. https://doi.org/10.3389/fnins.2013.00005.
7. Henriksen, S., Pang, R., and Wronkiewicz, M. (2016). A simple generative model of the mouse mesoscale connectome. Elife *5*, e12366. https://doi.org/10.7554/elife.12366.
8. Fulcher, B.D., and Fornito, A. (2016). A transcriptional signature of hub connectivity in the mouse connectome. Proc. Natl. Acad. Sci. USA *113*, 1435–1440.
9. Reimann, M.W., Gevaert, M., Shi, Y., Lu, H., Markram, H., and Muller, E. (2019). A null model of the mouse whole-neocortex micro-connectome. Nat. Commun. *10*, 3903. https://doi.org/10.1038/s41467-019-11630-x.
10. Goel, P., Kuceyeski, A., Locastro, E., and Raj, A. (2014). Spatial patterns of genome-wide expression profiles reflect anatomic and fiber connectivity architecture of healthy human brain. Human Brain Mapping *35*, 4204–4218. https://doi.org/10.1002/hbm.22471.
11. Vértes, P.E., Rittman, T., Whitaker, K.J., Romero-Garcia, R., Váša, F., Kitzbichler, M.G., Wagstyl, K., Fonagy, P., Dolan, R.J., Jones, P.B., et al. (2016). Gene transcription profiles associated with inter-modular hubs and connection distance in human functional magnetic resonance imaging networks. Philos. Trans. R. Soc. Lond. B Biol. Sci. *371*, 20150362.
12. Diez, I., and Sepulcre, J. (2018). Neurogenetic profiles delineate large-scale connectivity dynamics of the human brain. Nature communications *9*.
13. Shen, X., Finn, E.S., Scheinost, D., Rosenberg, M.D., Chun, M.M., Papademetris, X., and Constable, R.T. (2017). Using connectome-based predictive modeling to predict individual behavior from brain connectivity. Nat. Protoc. *12*, 506–518. https://doi.org/10.1038/nprot.2016.178.
14. Ji, S., Fakhry, A., and Deng, H. (2014). Integrative analysis of the connectivity and gene expression atlases in the mouse brain. Neuroimage *84*, 245–253.
15. Huang, L., Kebschull, J.M., Fürth, D., Musall, S., Kaufman, M.T., Churchland, A.K., and Zador, A.M. (2020). BRICseq bridges brain-wide interregional connectivity to neural activity and gene expression in single animals. Cell *182*, 177–188.e27. https://doi.org/10.1016/j.cell.2020.05.029.
16. Mezias, C., Torok, J., Maia, P.D., Markley, E., and Raj, A. (2022). Matrix inversion and subset selection (miss): A pipeline for mapping of diverse cell types across the murine brain. Proc. Natl. Acad. Sci. USA *119*, e2111786119.
17. Lein, E.S., Hawrylycz, M.J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A.F., Boguski, M.S., Brockway, K.S., Byrnes, E.J., et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. Nature *445*, 168–176. https://doi.org/10.1038/nature05453.
18. Tasic, B., Yao, Z., Graybuck, L.T., Smith, K.A., Nguyen, T.N., Bertagnolli, D., Goldy, J., Garren, E., Economo, M.N., Viswanathan, S., et al. (2018). Shared and distinct transcriptomic cell types across neocortical areas. Nature *563*, 72–78. https://doi.org/10.1038/s41586-018-0654-5.
19. Zeisel, A., Hochgerner, H., Lönnerberg, P., Johnsson, A., Memic, F., van der Zwan, J., Häring, M., Braun, E., Borm, L.E., La Manno, G., et al. (2018). Molecular architecture of the mouse nervous system. Cell *174*, 999–1014.e22. https://doi.org/10.1016/j.cell.2018.06.021.
20. Allen, A.I.B.S. (2018). Cell Types Database - Technical White Paper: Transcriptomics.
21. Abdelnour, F., Voss, H.U., and Raj, A. (2014). Network diffusion accurately models the relationship between structural and functional brain connectivity networks. Neuroimage *90*, 335–347. https://doi.org/10.1016/j.neuroimage.2013.12.039.
22. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.
23. Fakhry, A., and Ji, S. (2015). High-resolution prediction of mouse brain connectivity using gene expression patterns. Methods *73*, 71–78. https://doi.org/10.1016/j.ymeth.2014.07.011.
24. Fornito, A., Arnatkevičiūtė, A., and Fulcher, B.D. (2019). Bridging the gap between connectome and transcriptome. Trends Cognit. Sci. *23*, 34–50. https://doi.org/10.1016/j.tics.2018.10.005.
25. Anaissi, A., Goyal, M., Catchpoole, D.R., Braytee, A., and Kennedy, P.J. (2016). Ensemble feature learning of genomic data using support vector machine. PLoS One *11*, e0157330. https://doi.org/10.1371/journal.pone.0157330.
26. Allen, A.I.B.S. (2013). Developing Mouse Brain Atlas - Technical White Paper: Reference Atlases for the Allen Developing Mouse Brain Atlas.
27. Ouyang, M., Kang, H., Detre, J.A., Roberts, T.P.L., and Huang, H. (2017). Short-range connections in the developmental connectome during typical and atypical brain maturation. Neurosci. Biobehav. Rev. *83*, 109–122. https://doi.org/10.1016/j.neubiorev.2017.10.007.
28. Naze, S., Proix, T., Atasoy, S., and Kozloski, J.R. (2021). Robustness of connectome harmonics to local gray matter and long-range white matter connectivity changes. Neuroimage *224*, 117364. https://doi.org/10.1016/j.neuroimage.2020.117364.
29. Pfeiffer, S.E., Warrington, A.E., and BANSAL, R. (1993). The oligodendrocyte and its many cellular processes. Trends Cell Biol. *3*, 191–197. https://doi.org/10.1016/0962-8924(93)90213-k.
30. Emery, B. (2010). Regulation of oligodendrocyte differentiation and myelination. Science *330*, 779–782. https://doi.org/10.1126/science.1190927.
31. Eroglu, C., and Barres, B.A. (2010). Regulation of synaptic connectivity by glia. Nature *468*, 223–231. https://doi.org/10.1038/nature09612.
32. Kawamura, A., Abe, Y., Seki, F., Katayama, Y., Nishiyama, M., Takata, N., Tanaka, K.F., Okano, H., and Nakayama, K.I. (2020). Chd8 mutation in oligodendrocytes alters microstructure and functional connectivity in the mouse brain. Mol. Brain *13*, 160. https://doi.org/10.1186/s13041-020-00699-x.
33. Wang, F., Yang, Y.J., Yang, N., Chen, X.J., Huang, N.X., Zhang, J., Wu, Y., Liu, Z., Gao, X., Li, T., et al. (2018). Enhancing oligodendrocyte myelination rescues synaptic loss and improves functional recovery after chronic

hypoxia. Neuron *99*, 689–701.e5. https://doi.org/10.1016/j.neuron.2018.07.017.

34. Buchanan, J., Elabbady, L., Collman, F., Jorstad, N.L., Bakken, T.E., Ott, C., Glatzer, J., Bleckert, A.A., Bodor, A.L., Brittan, D., et al. (2021). Oligodendrocyte precursor cells prune axons in the mouse neocortex. Preprint at bioRxiv. https://doi.org/10.1101/2021.05.29.446047.
35. Abbott, N.J., Patabendige, A.A.K., Dolman, D.E.M., Yusof, S.R., and Begley, D.J. (2010). Structure and function of the blood–brain barrier. Neurobiol. Dis. *37*, 13–25. https://doi.org/10.1016/j.nbd.2009.07.030.
36. Langen, U.H., Ayloo, S., and Gu, C. (2019). Development and cell biology of the blood-brain barrier. Annu. Rev. Cell Dev. Biol. *35*, 591–613. https://doi.org/10.1146/annurev-cellbio-100617-062608.
37. Chow, B.W., and Gu, C. (2015). The molecular constituents of the blood–brain barrier. Trends Neurosci. *38*, 598–608. https://doi.org/10.1016/j.tins.2015.08.003.
38. Daneman, R., and Prat, A. (2015). The blood–brain barrier. Cold Spring Harbor Perspect. Biol. *7*, a020412. https://doi.org/10.1101/cshperspect.a020412.
39. Ballabh, P., Braun, A., and Nedergaard, M. (2004). The blood–brain barrier: an overview. Neurobiol. Dis. *16*, 1–13. https://doi.org/10.1016/j.nbd.2003.12.016.
40. Cauli, B., and Hamel, E. (2010). Revisiting the role of neurons in neurovascular coupling. Front. Neuroenergetics *2*, 9. https://doi.org/10.3389/fnene.2010.00009.
41. Chow, B.W., Nuñez, V., Kaplan, L., Granger, A.J., Bistrong, K., Zucker, H.L., Kumar, P., Sabatini, B.L., and Gu, C. (2020). Caveolae in CNS arterioles mediate neurovascular coupling. Nature *579*, 106–110. https://doi.org/10.1038/s41586-020-2026-1.
42. Kaplan, L., Chow, B.W., and Gu, C. (2020). Neuronal regulation of the blood–brain barrier and neurovascular coupling. Nat. Rev. Neurosci. *21*, 416–432. https://doi.org/10.1038/s41583-020-0322-2.
43. Jafari, A., de Lima Xavier, L., Bernstein, J.D., Simonyan, K., and Bleier, B.S. (2021). Association of sinonasal inflammation with functional brain connectivity. JAMA Otolaryngol. Head Neck Surg. *147*, 534–543.
44. Morimoto, K., and Nakajima, K. (2019). Role of the Immune System in the Development of the Central Nervous System. Front. Neurosci. *13*, 916. https://doi.org/10.3389/fnins.2019.00916.
45. Allen, N.J., and Eroglu, C. (2017). Cell biology of astrocyte-synapse interactions. Neuron *96*, 697–708.
46. Chuhma, N., Tanaka, K.F., Hen, R., and Rayport, S. (2011). Functional Connectome of the Striatal Medium Spiny Neuron. J. Neurosci. *31*, 1183–1192. https://doi.org/10.1523/JNEUROSCI.3833-10.2011.
47. Wang, X., Allen, W.E., Wright, M.A., Sylwestrak, E.L., Samusik, N., Vesuna, S., Evans, K., Liu, C., Ramakrishnan, C., Liu, J., et al. (2018). Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science *361*, eaat5691. https://doi.org/10.1126/science.aat5691.
48. Codeluppi, S., Borm, L.E., Zeisel, A., La Manno, G., van Lunteren, J.A., Svensson, C.I., and Linnarsson, S. (2018). Spatial organization of the somatosensory cortex revealed by osmfish. Nat. Methods *15*, 932–935. https://doi.org/10.1038/s41592-018-0175-z.
49. Moffitt, J.R., Bambah-Mukku, D., Eichhorn, S.W., Vaughn, E., Shekhar, K., Perez, J.D., Rubinstein, N.D., Hao, J., Regev, A., Dulac, C., and Zhuang, X. (2018). Molecular, spatial and functional single-cell profiling of the hypothalamic preoptic region. Science *362*, eaau5324. https://doi.org/10.1126/science.aau5324.
50. Zhang, M., Eichhorn, S.W., Zingg, B., Yao, Z., Cotter, K., Zeng, H., Dong, H., and Zhuang, X. (2021). Spatially resolved cell atlas of the mouse primary motor cortex by merfish. Nature *598*, 137–143. https://doi.org/10.1038/s41586-021-03705-x.
51. Chen, X., Fischer, S., Zhang, A., Gillis, J., and Zador, A.M. (2022). Modular cell type organization of cortical areas revealed by in situ sequencing. Preprint at bioRxiv *598*. https://doi.org/10.1101/2022.11.06.515380.
52. Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. *27*, 1226–1238. https://doi.org/10.1109/TPAMI.2005.159.
53. Ho, T.K. (1998). The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. *20*, 832–844.
54. Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32.
55. Qi, Y. (2012). Random forest for bioinformatics. Ensemble Machine Learning, *307–323* (Springer).
56. Segal, M.R. (2004). Machine Learning Benchmarks and Random Forest Regression (UCSF: Center for Bioinformatics and Molecular Biostatistics).
57. Hoerl, A.E., and Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics *12*, 55–67.
58. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. B *58*, 267–288.
59. Boser, B.E., Guyon, I.M., and Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory, 144–152.
60. Chang, C.-C., Yu, S.C., McQuoid, D.R., Messer, D.F., Taylor, W.D., Singh, K., Boyd, B.D., Krishnan, K.R.R., MacFall, J.R., Steffens, D.C., and Payne, M.E. (2011). Libsvm: a library for support vector machines. Psychiatr. Res. *193*, 1–6.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Data and code for: Spatial Cell Type Enrichment Predicts Mouse Brain Connectivity	This study	https://doi.org/10.5061/dryad.t76hdr866
Mouse mesoscale connectome	Oh et al.[2] (https://doi.org/10.1038/nature13186)	Website: https://connectivity.brain-map.org/
Tasic et al. scRNAseq data	Tasic et al.[18] (https://doi.org/10.1038/s41586-018-0654-5)	Website: http://portal.brain-map.org/atlases-and-data/rnaseq
Zeisel et al. scRNAseq data	Zeisel et al.[19] (https://doi.org/10.1016/j.cell.2018.06.021)	Website: http://mousebrain.org/
Allen Gene Expression Atlas	Lein et al.[17] (https://doi.org/10.1038/nature05453)	Website: https://mouse.brain-map.org/agea
MISS-inferred cell densities	Mezias et al.[16] (https://doi.org/10.1073/pnas.2111786119)	Zenodo: DOI: 10.5281/zenodo.8360355
Software and algorithms		
Data and code for: Spatial Cell Type Enrichment Predicts Mouse Brain Connectivity	This study	Dryad: DOI: 10.5061/dryad.t76hdr866
MISS pipeline	Mezias et al.[16] (https://doi.org/10.1073/pnas.2111786119)	Zenodo: DOI: 10.5281/zenodo.8360355
Brainframe	Justin Torok, Christopher Mezias	Zenodo DOI: 10.5281/zenodo.836076

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Ashish Raj (ashish.raj@ucsf.edu).

Materials availability

This study did not generate any new materials.

Data and code availability

- All data files for performing the analyses and generating the figures presented here are publicly available at the DOI provided in the key resources table
- All code for performing the analyses and generating the figures presented here are publicly available at the DOI provided in the key resources table
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

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. These data are available at the DOI listed in the key resources table above.

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,[16,52] 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 used here contains 200 cell types from the Mouse Brain Atlas (mousebrain.org), which were sampled from 12 locations throughout the mouse cortex.[19] A second, confirmatory 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] 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 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, 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.

Cell type input features

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

Null model input features

To benchmark the performance of our 200-feature model using the Zeisel et al. cell types, we five types of ''null'' input features, each of which is 424×1 vector of regional values:

1. *Purely random*: For each of the 200 input features, each regional value is independently sampled from a uniform random distribution.
2. *Region-coupled*: The 424 connectome regions are grouped into 13 major anatomical parcels (Data S11) and each of these parcels is assigned a normal distribution with a standard deviation of 1 and a mean specific to that parcel (i.e., the distribution for parcel 1 has a mean of 1, the distribution for parcel 2 has a mean of 2, etc.). Then, for each of the 200 input features, each regional value is sampled from the distribution corresponding to the parcel within which that region resides.
3. *Scrambled MISS*: 200 pseudo cell-type densities were obtained by performing nonnegative matrix inversion on a spatial gene expression matrix where each gene's regional expression values were scrambled. We used the same 1360 MRx3 genes that were used to infer the true densities of the 200 Zeisel et al. cell types for the inversion problem (see the Supplement and Mezias et al.[16] for further details on the MISS algorithm).

4. *Random genes*: Each of the 200 input features is the min-max normalized regional expression vector of a randomly selected gene from the 4083-gene AGEA.
5. *Random MRx3*: Each of the 200 input features is the min-max normalized regional expression vector of a randomly selected gene from the 1360-gene MRx3 subset that was used to infer the true Zeisel et al. cell-type densities.

We created 500 input feature sets for each of these null model types.

Random forest

The main findings reported in this study were obtained from random forest models for both the classification task and the regression task.[53–55] 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,[55,56] 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[22].

Other ML models

In addition to random, we implemented and tested several common machine learning algorithms including linear models such as ridge[57] and LASSO.[58] We also implemented support vector machines (SVMs) with a Radial (RBF) kernel for the task.[59,60] 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 Data S2 and S3). 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-based multi-layer perceptron. The network structure is as follows: number of input features, 400; 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).

Inter-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. We also manually constructed a taxonomic distance matrix based on the AIBS developmental atlas, where the timing of anatomical splits in the developing mouse brain defines a hierarchical relationship between each pair of regions in the adult mouse brain.[26] In all there were 6 such splits considered, with an emphasis on splits within the forebrain. Each region-pair was then assigned a distance from 0 to 6, which represents the number of branch points separating them in the hierarchical clustering tree (Figure S15; see also Data S12 for the taxonomy derived from the AIBS atlas).

Feature interpretation from random forest models

To decompose the random forest model and calculate the importance of each input feature, we used `Scikit-learn` *v*0.20.2 Python package.[22] 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:

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

where y_i is label/value for an instance, N is the number of instances and $\overline{y}$ 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 F.I. 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 F.I. 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.[22]

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 F.I. values for each cell supertype are averaged across their corresponding input cell-type features. Distributions of F.I. represent the results of 10-fold cross-validation.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analyses were performed using Python and MATLAB programming languages. Machine learning approaches were assessed and averaged over 10-fold cross-validation, with the Standard Error of the Mean (SEM) computed for precision and dispersion. Statistical significance was defined based on p values, and appropriate techniques were used for randomization, stratification, and sample size estimation. Accuracy and AUROC for the classification tasks were obtained directly from the Python implementation of random forest. R^2 and Pearson's R values were obtained using standard linear regression. Two-sample t-tests following Fisher's R-to-Z transformation were used to compare model performance. Preliminary analyses were conducted to ensure data met the assumptions of the chosen statistical methods, addressing any deviations through data transformation or non-parametric alternatives.