Prediction of a cell-type specific mouse mesoconnectome using gene expression data

Nestor Timonidis · Rembrandt Bakker · Paul Tiesinga

Abstract Reconstructing brain connectivity at sufficient resolution for computational models designed to study the biophysical mechanisms underlying cognitive processes is extremely challenging. For such a purpose, a mesoconnectome that includes laminar and cell-type specificity would be a major step forward. We analysed the ability of gene expression patterns to predict cell-type and laminar specific projection patterns and analyzed the biological context of the most predictive groups of genes.

To achieve our goal, we used publicly available volumetric gene expression and connectivity data and pre-processed it for prediction by averaging across brain regions, imputing missing values and rescaling. Afterwards, we predicted the strength of axonal projections and their binary form using expression patterns of individual genes and co-expression patterns of spatial gene modules.

For predicting projection strength, we found that ridge (L2-regularized) regression had the highest cross-validated accuracy with a median r^2 score of 0.54 which corresponded for binarized predictions to a median area under the ROC value of 0.89. Next, we identified 200 spatial gene modules using the dictionary learning and sparse coding approach. We found that these modules yielded predictions of comparable accuracy, with a median r^2 score of 0.51. Finally, a gene ontology enrichment analysis of the most predictive gene groups resulted in significant annotations related to post-

Nestor Timonidis
[1]Neuroinformatics department, Donders Centre for Neuroscience, Radboud University Nijmegen, Heyendaalseweg 135, 6525 AJ Nijmegen, the Netherlands
Tel.: +31-649552777
E-mail: n.timonidis@donders.ru.nl

Rembrandt Bakker
[1]Neuroinformatics department, Donders Centre for Neuroscience, Radboud University Nijmegen, Heyendaalseweg 135, 6525 AJ Nijmegen, the Netherlands
[2]Inst. of Neuroscience and Medicine (INM-6) and Inst. for Advanced Simulation (IAS-6) and JARA BRAIN Inst. I, Jülich Research Centre, Wilhelm-Johnen-Strasse, 52425 Jülich, Germany.

Paul Tiesinga
[1]Neuroinformatics department, Donders Centre for Neuroscience, Radboud University Nijmegen, Heyendaalseweg 135, 6525 AJ Nijmegen, the Netherlands

synaptic function.
Taken together, we have demonstrated a prediction pipeline that can be used to perform multimodal data integration to improve the accuracy of the predicted mesoconnectome and support other neuroscience use cases.

Keywords spatial gene co-expression, connectomics, machine learning, predictive models, mouse brain, axonal projection, gene expression, ontology enrichment analysis, regression, ridge regression, dictionary learning, sparse coding, roc analysis, cellular resolved connectome

1 Introduction

A wiring diagram of the brain (connectome) is a necessary step in advancing modern neuroscience for two main reasons. First, it assists computational neuroscience by providing biologically plausible constraints on brain models and simulations (Choi and Mihalas, 2019). Second, it bridges the gap between experimental data and computational models by providing frameworks exposing its topology and other properties (Sanz-Leon et al., 2013; Ritter et al., 2013; Woodman et al., 2014). Examples of connectome based projects are the Blue Brain or the Virtual Brain that aim to create large scale cellular level models of the human brain (Markram, 2006; Markram et al., 2011; Sanz Leon et al., 2013).

The meso-scale description of the connectome (mesoconnectome) is defined at the level of anatomically distinct sub-areas within each brain region and is typically described by the use of tract-tracing invasive techniques in animal studies, or post mortem dissections in human studies (Kötter, 2007; Sporns et al., 2005; Highley et al., 1999; Lanciego and Wouterlood, 2011). The whole brain coverage provided by these techniques and the ability to delineate layer specific sub-areas make the mesoconnectome neither too coarse grained nor too spatially limited and thus suitable for developing computational models of structural brain connectivity (Oh et al., 2014; Knox et al., 2018; Betzel et al., 2015a,b).

It is difficult with tract-tracing techniques to get good whole brain coverage and they are time consuming (Sporns, 2011). As an alternative to classical neuroanatomy, genomic-based approaches have been used to describe the connectome for a number of reasons (Fornito et al., 2019). First, it is possible to infer connectivity information from genes based on the premise that postsynaptic structures have specific protein profiles and that neurons connected through synapses have highly coupled gene expression patterns (Roy et al., 2018; Sperry, (1963). Second, the recent advances in genome sequencing have resulted in gene expression data being high throughput, relatively cheap and easy to obtain (Shendure and Ji, 2008).

These advantages have led to various studies linking genomic information and structural brain connectivity with computational approaches (Baruch et al., 2008; Kaufman et al., 2006; French and Pavlidis, 2011; French et al., 2011; Wolf et al., 2011). In recent studies, a link has been established between gene expression and the mouse mesoconnectome by building predictive models and associating gene co-expression with network topology and structure (Rubinov et al., 2015; Fulcher and Fornito, 2014;

Ji et al., 2014), resulting in computational frameworks for the mouse mesoconnectome.

Despite the aforementioned advances, research in the field still faces a number of limitations. Examples are the lack of cell-type specificity or synaptic density for describing specific neuronal populations in source and target brain areas that are connected through axonal projections. Such descriptions have been provided at local microcircuit level of the mouse brain but are limited to distinct brain areas such as the primary visual cortex (Lee et al., 2016). Moreover, important cytoarchitectonic features of the connectome such as the number of axonal fibers and the density of axonal arbor ending can not be extracted from models describing it as a binary network of present or absent projections between areas (Ji et al., 2014; Fulcher and Fornito, 2014).

In this work we use computational methods to measure the amount of information about axonal projection patterns present in gene expression patterns of the mouse brain and to associate them with factors related to the functional organization of genes. We have developed a pipeline, which we primarily describe in the methods section, that is available from a number of repositories. This paper is meant to describe the results we obtained with it and serve as a validation. The results section is organized as follows. First, we provide models that predict the strength or presence of axonal projection patterns given gene expression data, and we evaluate their performance on layer and cell class specific projection patterns that cover the whole brain at the mesoscale level. Second, we examine the relationship between the spatial pattern of gene co-expression modules and projection patterns in order to explain the performance of the predictive process. Third, we determine the ontological significance of predictive groups of genes in order to assess the biological relevance and causality of the predictive factors.

2 Methods

We built a computational framework to measure the amount of information about axonal projection patterns present in gene expression patterns of the mouse brain and to associate it with factors related to the functional organization of genes. Here we describe what data we used and how they were pre-processed as well the various steps of the analysis.

2.1 Materials

2.1.1 Allen Mouse Brain Atlas

The gene expression data were obtained from the Allen Mouse Brain Atlas (AMBA) dataset of the Allen Institute for Brain Science (table 1), (Lein et al., 2007). The in situ hybridization (ISH) technique was used to quantify ∼20.000 genes over multiple spatial locations from the brains of C57BL/6J (wild-type) mice which were male and 56-day-old (P56). In this technique, a probe of complementary strand of RNA labelled with fluorescent molecules binds with the RNA of dissected brain tissue.

Given that binding happens in situ, the spatial location of the gene is marked and its expression can be visualized through fluorescence microscopy (Amann and Fuchs, 2008). ISH constitutes a high throughput approach for quantifying expression energies of multiple genes in multiple spatial locations with up to 1 μm resolution (Lein et al., 2007).

In the study that created the AMBA dataset (Lein et al., 2007), mRNA strands were used together with fluorescence microscopy in order to visualize the gene expression energy. The result of this analysis was a set of sagittal and coronal brain slice images containing the expression energy of ∼20000 and ∼3300 individual genes respectively (Lein et al., 2007). In our analysis, the coronal slices were selected because their in plane resolution was higher.

2.1.2 Allen Mouse Brain Connectivity Atlas

The axonal projection data were obtained from the Allen Mouse Brain Connectivity Atlas (AMBCA) dataset. These data were based on the anterograde tract-tracing technique that was used to quantify the strength of axonal projections within the brains of P56 wild-type and transgenic cre-line mice. In anterograde tract-tracing, fluorescent molecules are injected to a source brain area and they reach target brain areas by being transported along the axons and reaching the axonal terminals (Oh et al., 2014; Harris et al., 2018).

The AMBCA dataset is comprised of ∼1400 anterograde tract-tracing experiments, for which the projection density was quantified using two-photon microscopy. There were 14 major transgenic cre-lines used that resulted in the expression of label according to different laminar profiles and different cell classes within each cortical area (Harris et al., 2014, 2018). The cre-lines together with the wild-type data constituted the 15 tract-tracing categories processed in this study and in their raw form consisted of brain slice images containing projection densities of multiple target brain areas at 1 μm resolution (Oh et al., 2014; Harris et al., 2018).

2.1.3 Allen Pre-processing pipeline

The brain slice images were processed using the informatics processing pipeline of the Allen Institute for Brain Science (table 1). Specifically, they were registered and aligned in the same reference space according to the Common Coordinate Framework CCF v3.0 (table 1).

The end product was a 3D volumetric representation of both modalities that was covering the whole mouse brain in voxel form and was provided at 100 μm^3 resolution for the projection data and at 200 μm^3 for the gene expression data. Each resolution referred to the size of voxels in the 3D space and corresponded to a particular total number of voxels in that space defined by x, y and z coordinates. The total number of voxels was 132 x 80 x 114 in the 100 μm^3 resolution and 67 x 41 x 58 in the 200 μm^3 resolution. The last step in the informatics processing pipeline was the unionization process during which the volume of both data modalities was averaged over anatomically distinct brain areas. As a result, 2D arrays were created whose rows corresponded to brain areas and columns corresponded to tracing experiments or genes respectively (Oh et al., 2014).

2.1.4 Data Acquisition

In our predictive workflow we used three sources of neuroanatomical data, namely gene expression, wild-type tracing experiments and cre-line tracing experiments, that were downloaded with the mouse connectivity cache (MCC) API (table 1).
We packaged and pre-processed the data as follows (figure 1). First, a number of experiments corresponding to the expressions of genes or tracing experiments were downloaded from the Allen Institute with the use of MCC. Second, the unionized gene expression experiments were packed in a 2-dimensional array where rows correspond to anatomical brain areas and columns correspond to individual genes. Third, for each wild-type and cre-line tracing experiment, a matrix was created with rows corresponding to brain areas and columns corresponding to individual injections associated with source brain areas. Finally, all tracing-related matrices were assembled into one aggregate data structure together with tracing-related metadata such as the cell-type and laminar specificity of injections, acronyms of source areas and injection coordinates.

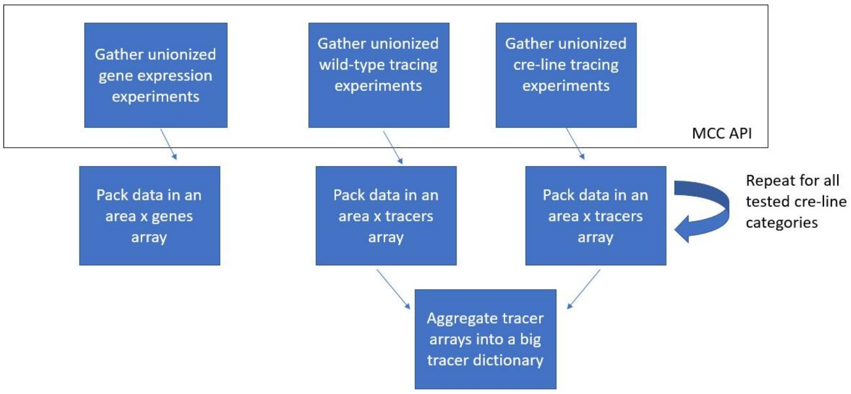

Gather unionized gene expression experiments
Gather unionized wild-type tracing experiments
Gather unionized cre-line tracing experiments
MCC API
Pack data in an area x genes array
Pack data in an area x tracers array
Pack data in an area x tracers array
Repeat for all tested cre-line categories
Aggregate tracer arrays into a big tracer dictionary

Fig. 1

2.2 Procedure

2.2.1 Pre-processing pipeline

We searched for not-a-number (NaN) values in the gene expression and axonal tracing datasets and removed them with a 2-step procedure based on their frequency of occurrence. Supplemental figure 1 documents that brain areas could be clustered in a group with NaN values for less than 10% of the entries, and a group with NaN values for more than 80% of the entries. We removed 610 out of 1038 anatomical brain areas defined by CCF v3.0 because they had a large fraction (> 80 %) of NaNs in either the gene expression or the axonal tracing datasets. The remaining NaN values in the Gene Expression dataset were imputed by taking the median value of the corresponding gene for all non-NaN brain areas (figure 2).

For the tracing data, a sampling-based imputation approach was followed (figure 2). To ensure that zero values would also have a chance of being used for imputing missing values, we stratified projection values per column (that is per tracing experiment) into zero and non-zero values. For each missing value present in a column, one of these groups was chosen with a probability proportional to its fraction in the non-missing data and from the chosen group a random value was drawn to be used for the imputation.

We subsequently rescaled both data modalities to obtain a proper range and distribution for use in the prediction procedure. First, a cube root transformation was applied in order to decrease the skewness of gene expression, since relative changes in expression across genes are considered to be more important than absolute ones and are usually less skewed (Ambrosius, 2007). This transformation was also applied to the axonal tracing data, since absolute changes in projection strength across tracing experiments are considered to be less important than relative ones. The cube root transformation decreased the range of gene expression values from (0 - 70) to (0 - 4) and their skewness from (-2 - 16) to (-3 - 4). This transformation also decreased the range of projection strength values from (0 - 427) to (0 - 7.5) and their skewness from (4 - 20) to (1 - 10) (supplemental figure 3). Second, z-score transformation was applied to both modalities in order to ensure that the regression-based predictive models were trained faster (Friedman et al., 2009). The z-score was obtained by subtracting the mean across areas and normalizing with the corresponding standard deviation (figure 2):

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

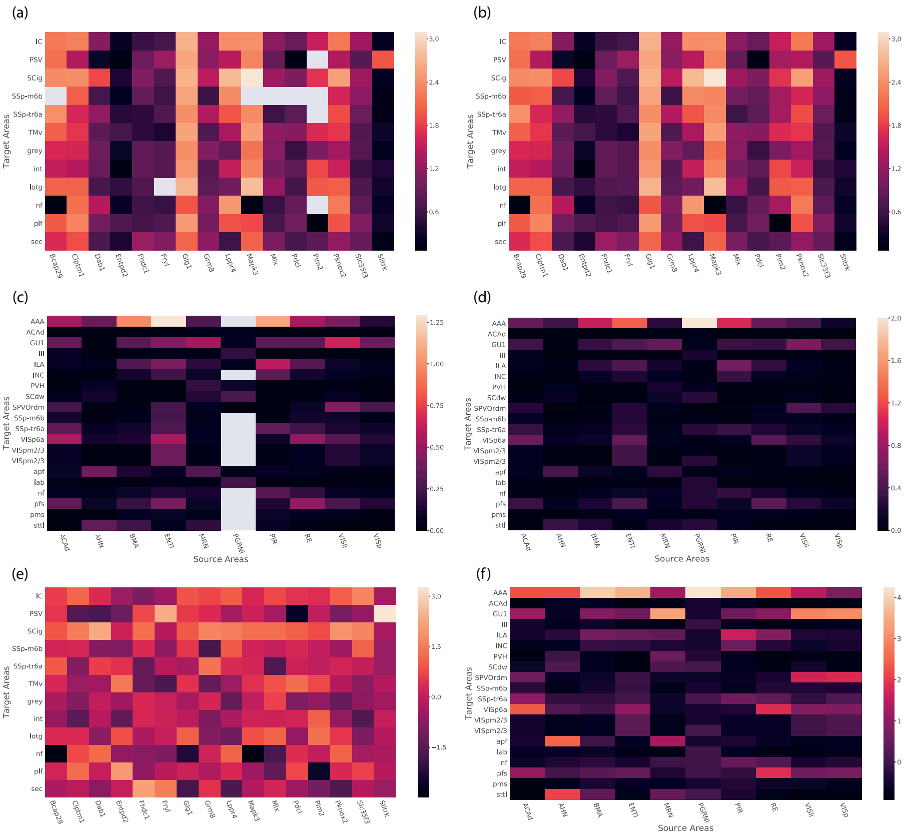


Fig. 2: The gene expression and wild-type axonal projection datasets during different pre-processing steps. (a-b) Subset of the Gene Expression dataset containing sparse NaN values (a) and the same subset after the median imputation (b). (c-d) Subset of the wild-type projection dataset containing sparse NaN values (c) and the same subset after the sampling imputation (d). (e-f) Z-score transformation of the gene expression dataset (e) and the wild-type projection strength dataset (f). The NaN values are shown in gray. The non-NaN values have been cube-root transformed for clarity. In (a-d) the brain areas were chosen to obtain examples with some NaN values present and examples without any NaN values.

2.2.2 *Model construction pipeline*

A separate prediction model was built for each cre-line or wild-type category as follows. First, the gene expression data were trained with either the random forest or ridge regression method. Subsequently, model performance was validated with nested 3-fold cross-validation (Varma and Simon, 2006) and quantified by the r^2 score between the measured and predicted projection patterns. The r^2 score is defined as the fraction of total variance of the measured patterns that can be explained by the predicted ones (Dodge, 2008):

$$r^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \tilde{y})^2}, \tag{2}$$

here y corresponds to a ground truth vector, $\tilde{y}$ corresponds to its mean and f corresponds to the predicted version of the vector. Finally, the predicted projection patterns with their optimal hyperparameter set were extracted as model outputs.
In the paragraphs below we describe the computational methods that we used for building our models.

2.2.3 Ridge Regression

Ridge Regression (also referred to as Tikhonov regularization) is a form of penalized linear regression commonly used in supervised machine learning and regression statistics (Tikhonov and Arsenin, 1977; Friedman et al., 2009). Classical linear regression fits a 2-dimensional array X to a vector y by estimating a coefficient vector w that minimizes the residuals between the actual y and the predicted $\hat{y}$ estimated as: $\hat{y}$ = Xw - b, where b is an intercept term.
The ordinary least squares method is used for optimizing the coefficient vector (Friedman et al., 2009):

$$\hat{w} = \operatorname{argmin}||y - Xw - b||_2^2 \tag{3}$$

With the estimation of $\hat{w}$, new data X_{new} can be given as input to the model for predicting or testing y_{new}:

$$y_{new} = X_{new}\hat{w} + b \tag{4}$$

In cases of high dimensional data, where N > M, the dataset exhibits high variance and thus noise which hinders the generalization performance of the trained model. Ridge regression deals with the problem by constraining the size of the coefficients (Friedman et al., 2009). This is done by adding the l_2 norm of the coefficients, multiplied by a shrinkage parameter λ, to the objective function:

$$\hat{w} = \underset{w \in R}{\operatorname{argmin}}||y - Xw - b||_2^2 + \lambda||w||_2^2 \tag{5}$$

The greater the value of λ the greater the shrinkage of the coefficients towards zero (Friedman et al., 2009). In our analysis we utilize ridge regression in order to fit gene expression data to projection patterns of the tract-tracing data and predict unseen patterns.

2.2.4 Random Forest Regressor

Random Forest Regressor is an ensemble method for performing regression tasks (Dietterich, 2000; Breiman, 2001). The basic premise for ensemble methods is that averaging reduces variance. Ensemble methods deal with high variability of predictive results by utilizing multiple models to fit and predict data, while the final result is derived by a majority voting in classification problems or by averaging in regression problems across all models participating in the ensemble (Dietterich, 2000).
In Random Forest the ensemble is comprised of multiple decision trees (Breiman, 2001). A decision tree is a directed acyclic graph or tree in which non-terminal nodes correspond to the rules of decision splits, edges correspond to each possible decision and terminal nodes correspond to the final decisions. In regression trees the internal

nodes correspond to value intervals and each leaf node corresponds to the average of all training data belonging to the intervals described by its parental nodes. A decision tree is constructed by fitting training data and is evaluated by new testing data that are being assigned to a class or to values based on the decision splits of their features implied by the tree (Tan et al., 2005).

One important property of the Random Forest method is that each decision tree gets assigned a random subset of the dataset, a technique which is also referred to as bagging (Dietterich, 2000). Moreover, each decision tree can use the whole feature set of a dataset or select a random subset of it (Breiman, 2001).

In our analysis Random Forest Regressor is utilized as an ensemble alternative to ridge regression in order to investigate differences in the predictive performance between different methods. Moreover according to literature, it constitutes a robust approach against data overfitting that occurs when the training data error is significantly lower than the testing data error (Breiman, 2001).

2.2.5 Dictionary Decomposition

Parallel to the predictive pipeline, the gene expression data were decomposed into transcriptional networks represented by spatial gene modules and coefficients. The Dictionary Learning and Sparse Coding method was used for decomposition, in which a data array is being represented by a linear combination of sparse but non-orthogonal modules or dictionaries and their coefficients (Mairal et al., 2010; Li et al., 2017). In dictionary learning both the coefficients and dictionaries are obtained by minimizing the deviation from the data under a L1 constraint on the coefficients (atoms) and non-negativity constraints on the elements of both the dictionaries and the coefficients:

$$\begin{aligned}(D,a) &= \operatorname{argmin} \frac{1}{2}||X-Da||_2^2, \\ ||a||_1 \leq \lambda, \quad ||a|| &> 0, \quad D_{i,j} > 0, \quad \forall i,j \in \mathbb{N}\end{aligned} \tag{6}$$

In our analysis the data array corresponded to the gene expression matrix, atoms corresponded to the coefficients of individual genes to each module and dictionaries corresponded to the spatial gene modules of the mouse brain.

There are multiple reasons to perform this gene expression decomposition. First, it allows us to visually inspect various gene co-expression patterns in the mouse brain. Second, it is a way to reduce redundancy since genes belonging to the same co-expression network have a putatively similar function across the brain (Langfelder and Horvath, 2008). Last but not least, it allows us to test multiple hypotheses regarding the effects of gene expression in predicting structural connectivity patterns. Specifically, the modification of atoms or gene coefficients can lead to altered gene expression patterns which can then be given to our models for predicting altered projection patterns.

2.2.6 Internal Model Validation

For an internal evaluation of our predictive models, a technique called nested k-fold cross-validation was applied to the dataset. Before discussing the technique, it is im-

portant to first describe the classical k-fold cross-validation, from which the nested version was developed as a way to reduce bias (Varma and Simon, 2006).
A k-fold cross-validation (also referred to as k-fold CV) is a technique for measuring how well does a supervised machine learning-based model perform on new or unseen data, also referred to as generalization performance (Bishop, 2006).
In the k-fold CV method, the dataset is partitioned into k disjoint subsets of approximately equal size. D corresponds to the dataset and $D_1, D_2, .., D_k$ are its disjoint subsets (Kohavi, 1995).
For i=1,..,k:

1. D_i is used as the testing set and $D \backslash D_i$ is used as the training set.
2. $D \backslash D_i$ constructs a classification/regression model using any relevant algorithm.
3. D_i is tested using the trained model.
4. $score_i$ is estimated as the score of the model for D_i, given any metric of interest (e.g. r^2).

$$score = \frac{\sum_{i=1}^{k} score_i}{k}, \tag{7}$$

After the procedure has been completed for all k-folds, then the total cross-validation score is estimated according to eq. 7, where k is the total number of folds and $score_i$ is the respective score per fold. The final score is the average over all folds (Kohavi, 1995).
Classical cross-validation is biased since both model performance evaluation and hyperparameter optimization can only be tested simultaneously on the same folds, and there is no independent set to test both factors separately. Nested cross-validation deals with the issue by nesting each training fold with internal training and testing folds and applying k-fold cross-validation to them (figure 3), (Varma and Simon, 2006). Internal testing folds are used for validating a hyperparameter set and their average predictive score (i.e. r^2) is the criterion for selecting the most optimal one (figure 3, eq. 2), while external testing folds validate the generalization performance of a model. Since each external testing fold validates a model whose hyperparameter set has been selected from other folds, the aforementioned bias is avoided (figure 3, eq. 7).
Furthermore, the overall stability of the trained models can be tested by comparing the overlap of the hyperparameters selected across all external training folds. If the overlap was more than 80%, we considered the model to be stable and we trained the model on the complete dataset with the most frequently selected hyperparameter set. In this case, new data were being tested on the new complete model If the overlap was between 60% and 80% we considered the model to be moderately instable and we tested new data by averaging their predictions over all folds. If the overlap was less than 60% we considered the model to be unstable and we removed it from our set.

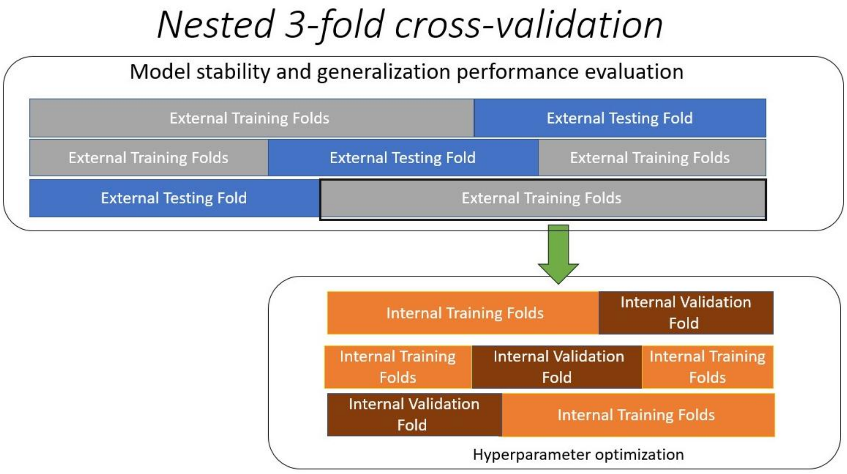


Fig. 3: Schematic describing the structure of the nested cross-validation method.

2.2.7 Post-hoc binarization

We have provided a post-hoc approach to analyze and visualize binary projection patterns in the mouse brain, primarily for facilitating comparison to previous studies (Ji et al., 2014). Since projection strength needed to be binarized, the binarization threshold was found by maximizing the area under the "receiver operating characteristic (ROC)" curve (auROC) value (Fawcett, 2006).

For the ROC analysis, classification scores are converted to binary patterns based on a threshold and the accuracy score between measured and predicted binary patterns is estimated as the ratio between the true positive rate (TPR) and the false positive rate (FPR) (eq. 8).

$$\begin{aligned} TPR &= \frac{TP}{P} = \frac{\text{Positives classified correctly}}{\text{Total number of Positives}} \\ FPR &= \frac{FP}{N} = \frac{\text{Positives classified incorrectly}}{\text{Total number of Negatives}} \\ ROC_{score} &= \frac{TPR}{FPR} \end{aligned} \tag{8}$$

Positives and Negatives correspond to the samples from the positive and negative class respectively. It is important to mention that the terms positive and negative are conventions used to characterize binary classes in classification problems. In our case for instance, the positive and negative classes would correspond to the presence and absence of strong projections from a source area, respectively.

The strength of ROC analysis lies in the application of approximately all possible thresholds in the range 0-1, leading to a curve of approximately all possible TPR and FPR values (figure 4). The auROC is estimated as the integral of the area under the curve that represents the potential quality of classification performance under various thresholds and also reveals the optimal threshold as the point on the curve furthest away from a 45 degree line which represents the performance of a random classification (figure 4). The importance of such a line is that the actual performance quality is visible by comparing the height difference between the line and the curve: the higher the curve from the line is, the more non-random and thus significant the actual performance is considered to be (Fawcett, 2006).

In order to apply ROC analysis on our continuous data, predicted patterns have to be converted to classification scores and a second threshold is needed to convert the measured projection patterns to binary ones. This is achieved by setting up an external threshold set, different from the internal one used in ROC analysis. Moreover, the predicted patterns are transformed to classification scores with the standard logistic sigmoid function: $f(x) = \frac{1}{1+e^{(-x)}}$

Therefore, for each external threshold in the set, the optimal auROC is estimated as the output of ROC analysis between the measured patterns binarized from the threshold and the predicted patterns that are converted to scores. The selected threshold is the one with the maximum optimal auROC value.

The Multi-ROC curve analysis in figure 4 is an example of the optimal threshold selection technique. The data corresponded to a Nr5a1-Cre tracing experiment, expressed in layer 4 and injected in the ventromedial hypothalamic nucleus (VMH). In that example, the external threshold selected was the 76th percentile of the measured data and corresponded to the curve with the maximum auROC value of 0.95. Moreover, the optimal internal threshold of the selected curve was 0.69 and was applied to the predicted data after the sigmoid transformation.

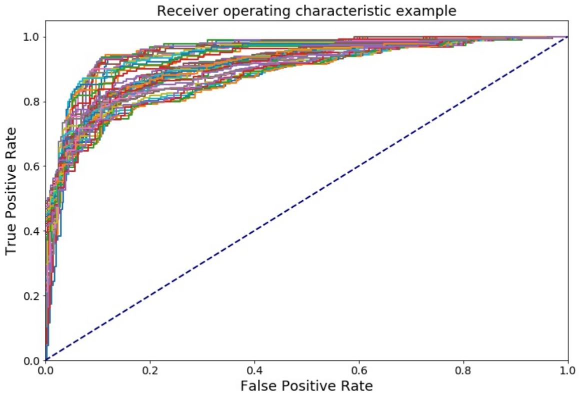


Fig. 4: Multi-ROC curve between measured and predicted projection patterns of the Nr5a1-Cre tracing experiment. The ROC curves correspond to multiple curves induced by applying multiple thresholds to the measured data.

2.2.8 Gene Enrichment Analysis

We used gene ontology (GO) enrichment analysis. This analysis is commonly used in the bioinformatics field for investigating the biological relevance of groups of genes

(Rivals et al., 2007). In particular, the hypergeometric test that utilizes the hypergeometric distribution (Rice, 2007), was applied for estimating the statistical significance of the number of genes with a particular annotation being amongst the most predictive genes in our procedure, relative to the occurrence of genes with this annotation in similarly sized groups drawn randomly from the entire gene set (Rivals et al., 2007).
We integrated GO enrichment analysis with our predictive workflow in a number of steps. First, we selected groups of genes with high coefficient scores in the prediction process (Results section, subsections 3.1 and 3.2) or genes with high coefficient scores in spatial gene modules of interest (Results section, subsection 3.4). Second, the hypergeometric test was applied to each selected gene group. Third, annotations for which the hypergeometric test returned a p-value lower than 0.05, were considered significant and were collected in a table.
The ontology annotations and the gene set for the randomly drawn subsets were taken from the *org.Mm.eg.db* database that was downloaded from the Bioconductor open source bioinformatics software (table 1) and contains genome-wide annotation for the mouse species.

2.2.9 Link to Mouse Connectivity Models

We linked our predictive workflow to the Mouse Connectivity Models (MCM) tool provided by the Allen Institute for Brain Science (Knox et al., 2018). The MCM tool comprises a set of approaches, based on penalized regression, for constructing connectivity matrices on a volumetric scale of 100 μm^3 or on a regionalized scale of structural brain areas. These tools enabled us to integrate the 1397 tract tracing experiments into one connectivity matrix and analyze the differences in projection patterns from different laminar profiles.

Table 1: Hyperlinks for websites, tool descriptions and format descriptions related to our analysis. See main text for details.

Allen Institute	`https://alleninstitute.org/`
CC documentation	`https://allensdk.readthedocs.io/en/latest/connectivity.html`
CCF v3.0	`http://help.brain-map.org/display/mouseconnectivity/Documentation`
MCC use case	`https://alleninstitute.github.io/AllenSDK/_static/examples/nb/mouse_connectivity.html`
NIfTI	`https://nifti.nimh.nih.gov/`
JSON	`https://en.wikipedia.org/wiki/JSON`
SBA	`https://scalablebrainatlas.incf.org/composer-dev/?template=ABA_v3`
Bioconductor software	`http://bioconductor.org/packages/release/data/annotation/html/org.Mm.eg.db.html`
Repository of our Code on the HBP Collaboratory	`https://collab.humanbrainproject.eu/#/collab/8650/nav/65518`
Repository of our Code on Github	`https://github.com/ntimonid/Connectomic-Composition-Predictor-CCP-`

3 Results

We downloaded the unionized label intensities from the Allen Mouse Brain Connectivity Atlas repository (see Methods, download date = 1-07-2018) of viral tracing experiments corresponding to 1397 distinct injection sites, of which the majority (n = 498) was in wild-type subjects and the remainder were made in 14 different cre-lines of transgenic animals. Unionized means that the label intensity is averaged across all voxels belonging to a particular brain region. Here we use the common coordinate framework (CCF v3.0) to assign each voxel to a brain region. In addition, we downloaded the corresponding unionized gene expression data. The data were pre-processed to remove regions with poor quality data, impute missing values and rescale values to an appropriate range for fitting (see Methods).

3.1 Prediction of continuous projection strength based on gene expression patterns

We explored various fitting procedures (see Methods) for predicting the connection strength (label intensity) from the gene expression data. The two supervised learning methods used for fitting the data were random forest and ridge regression, while the performance was measured using the r^2 score which represents the fraction of total variance accounted for by the model. Across all injection sites, irrespective of subject type, ridge regression based predictions yielded a median r^2 of 0.54 with an interquartile range of 0.178. Random forest based predictions yielded a median r^2 score of 0.42, which was lower than the one for the ridge regression based predictions (Figure 5). As an example, the data presented shown in Figure 6 were obtained using nested 3-fold cross-validation of ridge regression.

Variation in performance was analyzed across experiments of different tract-tracing categories. When the performance was partitioned according to transgenic cre-line and wild-type, the performance of wild-type was approximately in the center of the fit range based on transgenic animals. As the number of injections in each transgenic cre-line was much lower than available wild-type data (n = 12 to 125 for transgenic versus n = 498 for wild-type) this variation can be most likely attributed to experimental variability, rather than the specific properties of a transgenic line. Our statistical tests indicated that the difference was not statistically significant (p = 0.004 for 100 random permutations per cre-line, 14000 permutations in total, with the same distribution in set size as the cre-lines).

Predictions of projection patterns with the ridge regression-based models trained on gene expression data were significant. The ridge regression models trained with actual gene expression patterns outperformed in every case surrogate models, that were created by randomly distributing the expression intensity of each gene across areas (for three representative cases see figure 5). This process was repeated 25 times for each cre-line and wild-type tracing experiment. The predictive models that were trained with the surrogate data, also referred to as surrogate models, had a median r^2 score of -0.005 and an interquartile range of 0.007 over all tracing experiments.

All of the ridge regression models outperformed the null models, that were incorporated into the analysis as an additional control (figure 5). The null models predicted unseen projection patterns by averaging values of the seen ones and thus did not account for variability across brain areas. A model was considered inaccurate when it was outperformed by those null models. The null models had a median of -0.003 and an interquartile range of 0.005 over all tracing experiments.

Predictions with low r^2 values can be expected when multiple projection patterns with a noisy subset need to be predicted simultaneously. Specifically, the models were trained to fit multiple tracing experiments belonging to a particular tracing category (i.e. wild-type mice) with the same set of $\sim$3000 genes and the same hyperparameter set. In our data, 10 out of 1397 tracing experiments (0.7%) had a value in the range [0 - 0.2] for ridge regression based models, while the equivalent percentage for random forest based models was 40 out of 1397 (2.8%).

Nevertheless, performance of models with a high r^2 score can be appreciated when the predicted projection patterns are visually compared with the measured ones in the form of brain slices and cortical flatmaps (for an example see figure 6).

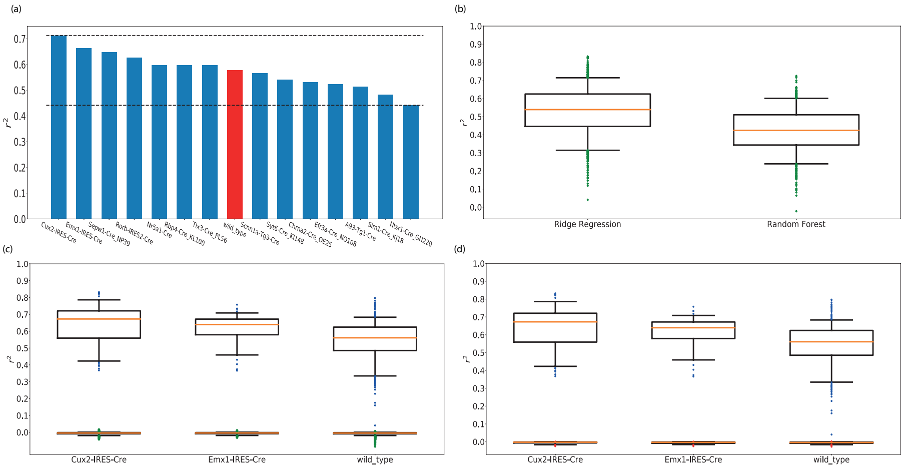


Fig. 5: (a) Prediction scores per tract-tracing category. x-axis: tract-tracing category. y-axis: r^2 values. Results for the wild-type tracing experiments are being highlighted in red in order to be differentiated from the cre-lines. (b) Comparison of ridge regression (left) with random forest (right) based models. y-axis: r^2 scores. The red line is the median, the box encloses the interquartile range and the green dots are outliers which comprised 0.7 % of the injections for ridge regression and 2% for the random forest. (c) Performance comparison of surrogate (bottom panel) and actual models (top panel) for a number of tracing datasets. x-axis: datasets - Cux2-IRES-Cre (left), Exm1-IRES-Cre (middle), wild-type (right). y-axis: r^2 scores. The red line and box are as in (b) while the blue and green dots are outliers for the regression of the actual and surrogate data respectively. (d) Performance comparison of null (bottom panel) and actual models (top panel) for a number of tracing datasets. x-axis: datasets - Cux2-IRES-Cre (left), Exm1-IRES-Cre (middle), wild-type (right). y-axis: r^2 scores. The color conventions are as in panel (b).

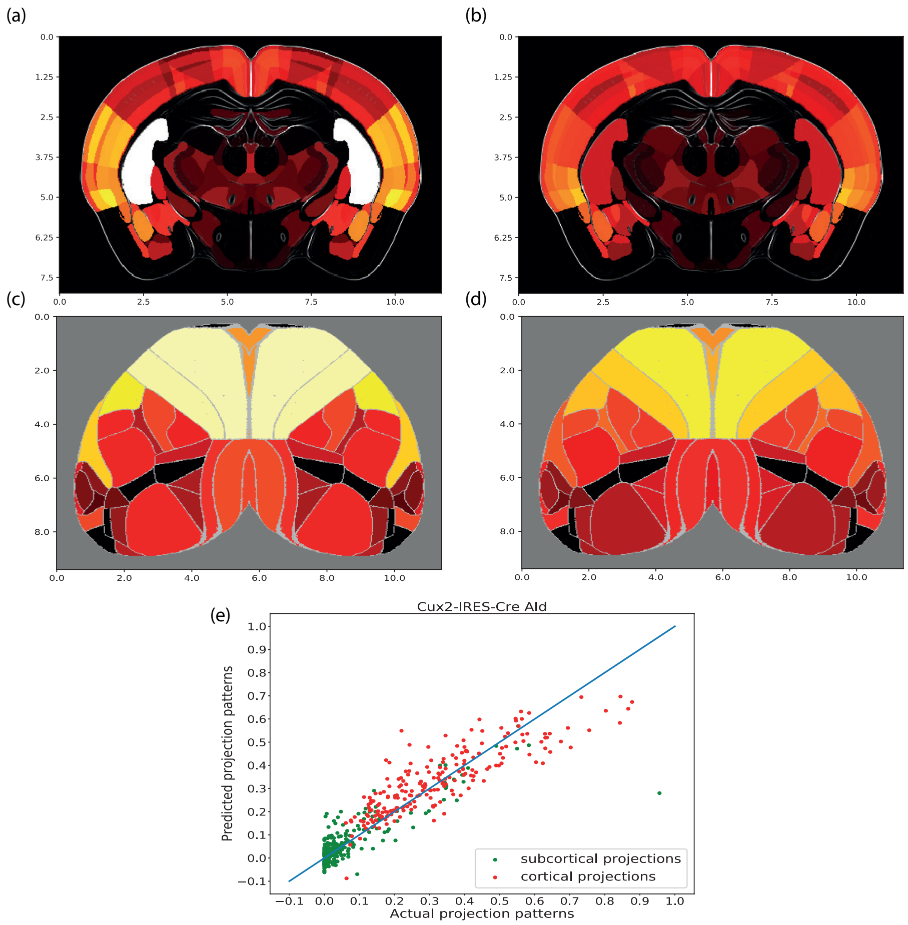


Fig. 6: Subcortical visualizations (a,b), cortical visualizations (c,d) and a prediction performance scatter plot (e) for a Cux2-IRES-Cre tracing experiment which labeled cells in layers 2/3 and was injected in the AId area (agranular insular area, dorsal part). The r^2 score for this experiment was 0.826, which was the highest score across all tracing experiments. (a,c): measured values. (b,d): predictions from gene expression patterns. The subcortical projection patterns were visualized using coronal slices of the projection volume, whereas the cortical projection patterns were projected onto a flatmap and their values have been averaged over all cortical layers. The scaling for both axes is in milimeters. The intensity of each plot was normalized by its maximum value. Cortical areas such as Retrosplenial area dorsal part (RSPd), Anteromedial visual area (VISam), trunk of primary somatosensory area (SSP tr) and Posterior auditory area (AUDpo) exhibit highly similar projection strength between their measured and predicted versions, while on the contrary subcortical similarities exist but are not as strong as the cortical ones. (e) x-axis: measured data. y-axis: predicted data. Green points correspond to subcortical projections, red points to cortical ones and the solid line is the diagonal, for which predicted values are equal to the measured ones. Cortical points are closer to the diagonal compared to subcortical ones, so they are more accurately predicted.

3.2 Binary Predictions

Previous studies have used a binarized version of the mesoconnectome to test the accuracy of their predictive models. In order to compare our performance to these models, we developed an approach to make binary predictions as well (see Methods, figure 4). The accuracy of these predictions was quantified using an ROC analysis with as outcome the area under the ROC curve (auROC).

The median auROC value over all 1397 tracing experiments was 0.89 with a median interquartile range of 0.08 (figure 7). Moreover, performance for wild-type data matched that of the state of the art in binary projection predictions of wild-type experiments with gene expression data, such as in (Ji et al., 2014), where 93% auROC was obtained. The auROC values for all wild-type tracing experiments had a median of 0.93 and an interquartile range of 0.05. Similar values for cre-lines were obtained, which had not been subject to this analysis before (Harris et al., 2018). For instance, the auROC values for Tlx3-Cre_PL56 tracing experiments, labeling cells in layers 2-6, had a median of 0.94 and an interquartile range of 0.03 (figure 7).

Visualization of measured and predicted results, in the form of cortical flatmaps and coronal slices, allows for assessing the quality of predictions in spatial context. An example is the Cux2-IRES-Cre tracing experiment injected in AId area (figure 8), which had an auROC value of 0.98 for binary prediction.

The increased performance of the models on binary predictions compared to continuous ones (figure 8) was due to reduced content of the projection patterns, which can therefore be more easily captured by the gene expression data. However, the resulting connectivity descriptions are on a very coarse-grained level which made the continuous ones more suitable for analytic purposes.

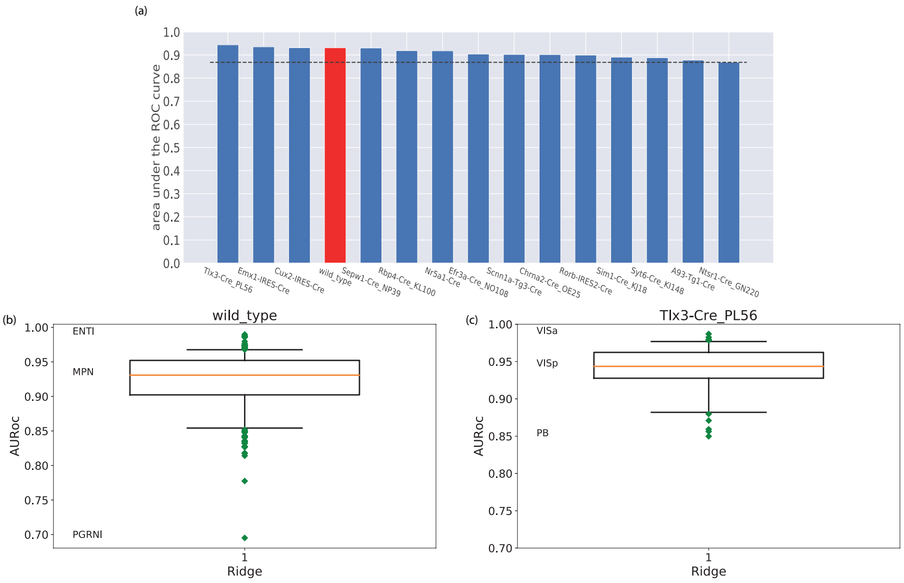


Fig. 7: (a) Median area under the ROC curve per tract-tracing category. x-axis: tract-tracing category. y-axis: auROC values. Results for the wild-type tracing experiments are being highlighted in red in order to be differentiated from the cre-lines. (b-c) Binarized prediction performance for different categories of tracing experiments. (b) wild-type dataset. (c) Tlx3-Cre_PL56 dataset. y-axis: auROC values.

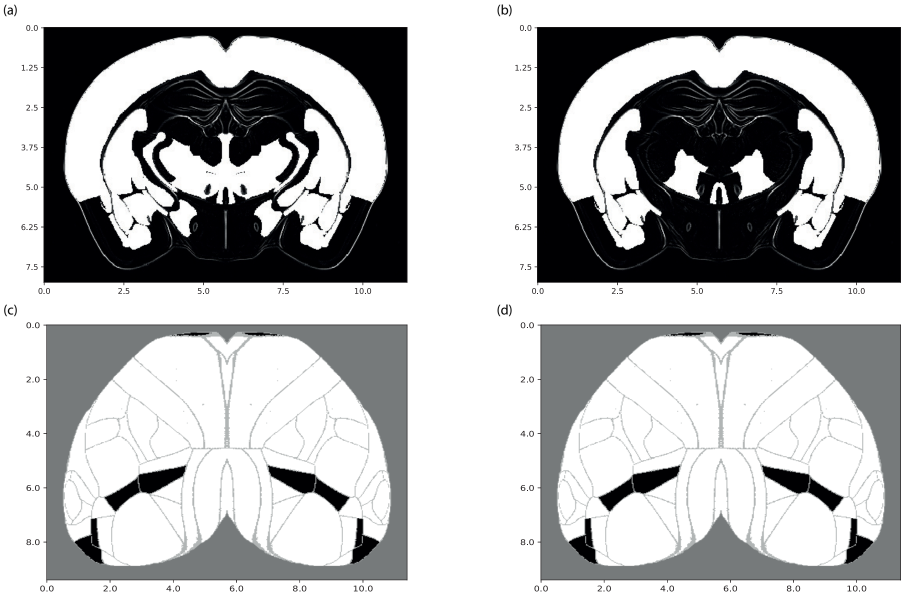


Fig. 8: Subcortical visualizations (a,b) and cortical ones (c,d) of the binarized form for a Cux2-IRES-Cre tracing experiment injected in the AId area (agranular insular area, dorsal part). (a,c): measured values. (b,d): predictions from gene expression patterns. The subcortical projection patterns were visualized using coronal slices of the projection volume (a,b), whereas the cortical projection patterns (c,d), were projected onto a flatmap. The scaling for both axes is in milimeters. White denotes the value 1 (connections present), and black denotes the absence of a projection.

3.3 Creation of a regionalized connectivity array based on laminar specific projection patterns

Our predictive workflow has also incorporated regionalized connectivity models provided by the MCM tool. Specifically, we applied the MCM tool to an aggregate of the measured projection patterns from all cre-lines. The output was a laminar specific regionalized connectivity array between anatomical brain areas, for which both source and target cortical areas were laminar specific.

We investigated the differences in projection patterns across source areas with different laminar profiles. Figure 9 shows an indicative subset of the regionalized array as well as a similarity matrix between source cortical areas with different laminar profiles. The similarity matrix was created by estimating the Spearman's rank coefficient (also referred to as rho) between the different source areas. We clustered the source areas on the similarity matrix based on their laminar profile. In addition, we applied the silhouette score for quantifying the clustering quality, which is a standard measure in clustering analysis with values ranging from -1 to 1 and reflecting the clustering cohesion (Rousseeuw, 1987). The silhouette score was 0.61, which was considered to reflect the cohesive clusters found in figure 9. Therefore, we regarded projection patterns of source areas with the same laminar profile to be more similar compared to those of different profiles.

In order to estimate the significance of that finding, we generated surrogate clusters by randomly distributing the projection densities across source areas 1000 times. In addition, we estimated a p-value based on the number of times that the silhouette score of the surrogate clusters was greater than the score of the actual clusters. The resulting p-value was 0, which indicated that differences in projection patterns from source areas with different laminar profiles were significant.

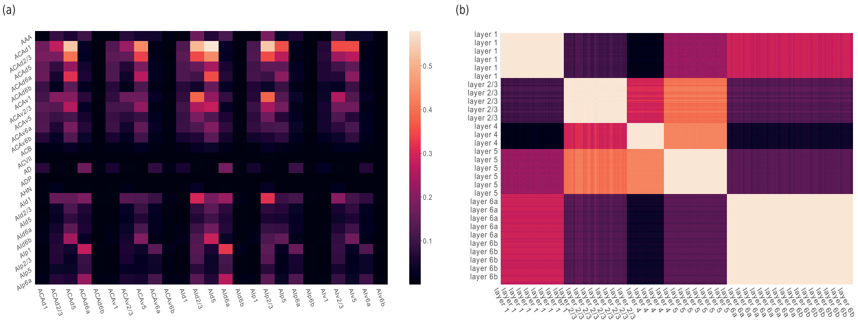


Fig. 9: Heatmaps of a laminar specific regionalized connectivity array. (a) Subset of the array comprised from a selected set of 25 target and 25 source brain areas. x-axis: source brain areas. y-axis: target brain areas. (b) Similarity matrix of source brain areas which are clustered based on their laminar profiles. Both axes correspond to clustered laminar profiles of source brain areas. The similarity matrix was created by taking all pairs of source areas and estimating the Spearman's rho between their projection patterns. All distinct blocks with values greater than 0.9 represent pairs of areas with the same profile. This suggests that groups of areas with the same laminar profile have more similar projection patterns compared to groups of areas with different profiles.

3.4 Gene Module Analysis

Both the projection patterns as well as the gene expression patterns are vectors in an abstract space, for both regionalized as well as volumetric data. They can thus be written as sums of basis vectors. One can ask for a basis that most efficiently represents the variability of gene expression patterns across genes and projection patterns across injections. Therefore, we used the Dictionary Learning and Sparse Coding (DLSC) technique to find overcomplete dictionaries that account for patterns with a small number of basis vectors and small coefficients (see Methods) (Li et al., 2017). These dictionaries define groups of genes with similar spatial patterns (modules), whose spatial profile should be less noisy than the individual genes that contribute to it. Hence, in case that we do not capture the noise of individual gene expression patterns, these modules should form a better basis for predictive approaches.

With the intention of identifying functional groups with a sparse spatial distribution, we set the λ parameter (L1 constraint) to 1.0 (see eq. 6). The dictionary set size was chosen by training models to predict tract-tracing experiments with a different number of spatial modules, and then selecting a model with a high r^2 score (see figure 10). We selected a set of 200 modules despite being second in performance (the median r^2 is 0.51 for 200 modules and 0.52 for 300 modules), since the set of 300 modules was considered to be too large and their difference was considered to be an effect of variability (both interquartile ranges are 0.19 as shown by the vertical lines in figure 10, panel C). The selected dictionary set accounted on average for 10% of variability across genes and had an average spatial footprint of 88% of the brain areas. Therefore, the resulting spatial module matrix was 428 x 200, which was a significant reduction in dimensionality compared to the 428 x 3318 ISH gene expression matrix.

In order to examine the predictive capabilities of the spatial modules, the prediction process was repeated with models trained on the modules instead of genes. We considered an example tracing experiment for which the module based predictive model had the highest r^2 score of 0.79. The tracing experiment was generated by a Cux2-IRES-Cre injection in Retrosplenial area, lateral agranular part (RSPagl). We looked for modules with the highest similarity with the projection pattern, as quantified using the pearson correlation coefficient (r). We selected three modules, labeled as 9, 70 and 88, with a pearson r of 0.51, 0.52 and 0.45 respectively. Each of these modules were non-zero in a mostly non-overlapping group of brain areas, which together cover a part of the experimental projection pattern (see figures 11,12). We analyzed the contribution of their spatial footprint in each area separately, by replacing each nonzero value by 1.0 if present in all three modules and the projection pattern, 0.8 if present in two modules and the pattern, 0.6 if present in one module and the pattern, 0.4 if present in the pattern and absent in all modules and 0.2 if present in the modules but absent in the pattern. As indicated in figure 12, there was a large overlap between the experiment and the modules in cortical areas. Subcortical areas did not have such strong coverage as cortical ones, which might be the reason why predictive performance was not higher in terms of the r^2 score.

Subsequently, we calculated the pearson r between the RSPagl experiment and its prediction by the three modules. We found that this prediction yielded an r^2 score of 0.4 and a pearson r value of 0.64, which was higher than the median pearson r of 0.54 over all tracing experiments. Therefore, these modules were important components to the total prediction, whereas they provided a less accurate prediction as stand-alone predictors (see figure 11,12).

This finding suggests that multiple spatial modules might be needed to reproduce projection density patterns from the mouse cortex (figures 12,10). For the predictive models trained and tested with spatial modules over all tracing experiments, the median r^2 score was 0.51, the interquartile range was 0.19, and the maximum r^2 score was 0.79. Therefore results were slightly lower on average than the corresponding ones for the gene predictions (figure 10). For testing the significance of module based predictions, surrogate models were built as explained in subsection 3.1 and trained with spatial modules instead of genes. All models trained for the 1397 tracing experiments had higher r^2 values than the respective surrogate ones, as indicated by a number of examples in figure 10. Taken together, our findings suggest that spatial gene modules contain a large fraction of the predictive capacity of the much higher dimensional gene set.

Previous studies focused on integrating single-cell RNA sequencing with ISH data in order to provide cell-type densities (Mairal et al., 2010). Furthermore, the neuroexpresso tool has provided access to a large collection of single-cell gene expression data derived from multiple studies (Tasic et al., 2016; Mancarci et al., 2017).

In (Mancarci et al., 2017), the authors collected expression data of ∼11.000 genes from pooled cell-type microarray and single-cell RNA-sequencing studies. For that reason, we tested the capability of the DLSC model to provide meaningful modules by constraining it with the neuroexpresso data. We selected 74 cell-types from their repository (`https://github.com/PavlidisLab/markerGeneProfile`) and 2154 genes that were common between the single-cell and the ISH data, which resulted in a 2154 x 74 array of cell-type specific gene expression. Finally, the cell-type specific and the ISH gene expression arrays were given as inputs to the DLSC model that created a new array of 428 areas x 74 constrained spatial modules.

We trained the prediction models using the constrained spatial modules and evaluated the quality of the resulting projection patterns. The term unconstrained is used to describe the spatial modules from the DLSC model which was not constrained with single-cell RNA sequencing data. The median r^2 score of the constrained models for all tracing experiments was 0.45, which was substantially lower compared to the unconstrained ones (median r^2 of 0.51), and the interquartile range was 0.19. In addition, we created a similarity matrix between the constrained and unconstrained modules and we applied biclustering analysis to it (figure 13). The similarity matrix was created by Spearman's rho, and the biclustering algorithm used was Spectral Biclustering with 3 biclusters (Kluger et al., 2003). Moreover, the ranks of the unconstrained and constrained module arrays were 200 and 74 respectively, which indicates that both arrays were full rank and did not contain redundant modules. This analysis did not result in meaningful biclusters and suggested that there is little relationship between the two types of modules. Nevertheless, the unconstrained modules provided predictions of higher quality than the constrained ones.

Moreover, gene ontology enrichment analysis was applied to the unconstrained spatial modules and the tract-tracing experiments in order to identify significant annotations related to synaptic and neuronal function in the mouse brain (Rivals et al., 2007). The unconstrained modules were preferred over the constrained ones due to providing better quality predictions. For each tracing experiment, we included the most predictive genes whose coefficients exceeded the 99^{th} percentile for that experiment. In the case of modules we included all genes having a non-zero coefficient. The percentage of modules and tracing experiments having at least one significant annotation was 100% and 98% respectively. A tracing experiment was associated with 12 annotations on average (median), while a module was associated with 39 annotations on average.

We observed that annotations related to postsynaptic function were associated with both the RSPagl experiment and module 9 (see figure 14). The jaccard similarity coefficient between the significant annotations of module 9 and modules 70 and 88 was 0.7 and 0.69 respectively, and thus annotations of module 9 were considered to be representative of the three modules.

As a generalization of this observation, the percentage of modules and tracing experiments having at least one annotation related to postsynaptic function was 100% and 70% respectively. Hence, annotations with postsynaptic function was another common denominator between a substantial number of tracing experiments and spatial modules, in addition to strong correlations and predictive capability.

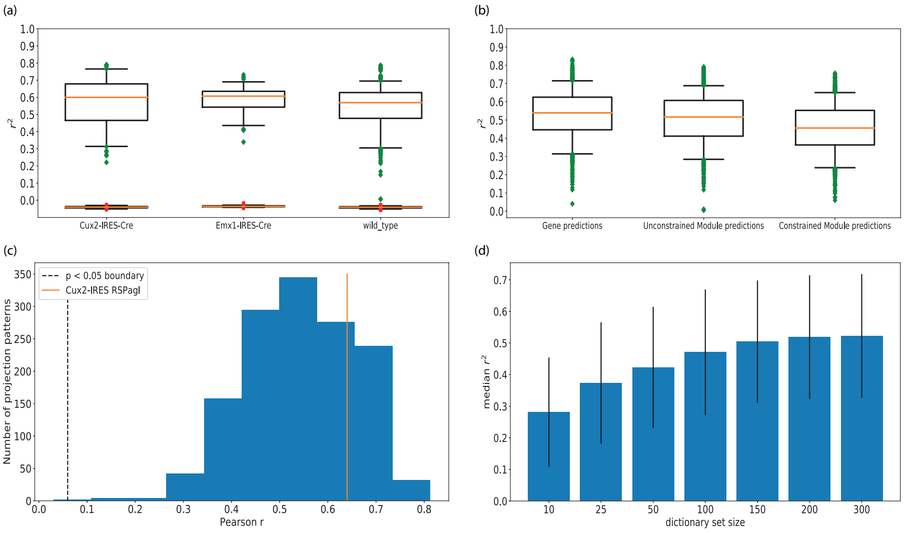


Fig. 10: (a) Performance comparison between surrogate (bottom panel) and actual models (top panel) trained with spatial modules for a number of tracing datasets. x-axis: datasets - Cux2-IRES-Cre (left), Exm1-IRES-Cre (middle), wild-type (right). y-axis: r^2 scores. (b) Comparison of predictive accuracy between models trained using spatial modules and models trained using full gene expression data. x-axis: gene expression based models (left), spatial module based models (middle) and module based models constrained with single-cell RNA sequencing data (right). y-axis: r^2 scores. (c) Histogram of pearson correlation coefficients (r) between all 1397 tracing experiments and their predicted versions. The prediction of each experiment was achieved with its 3 best correlated modules as determined by pearson r. The first vertical line to the left represents the point at which all correlations left from it are not longer statistically significant ($p > 0.05$). The second vertical line to the left corresponds to the pearson r of 0.64 between a Cux2-IRES-Cre experiment injected in the RSPagl area and modules 9, 70 and 88, which is greater than the mean r of 0.54 (see figure 12). A dense distribution of correlations in the range 0.4 - 0.7 indicates that multiple spatial modules correlate with axonal projection patterns. (d) Predictive performance of module-based models with different dictionary set sizes. x-axis: dictionary set size. y-axis: median r^2 score over all tract-tracing experiments. The vertical lines represent the interquartile range across the dictionary sets. The highest peak is for 300 modules with an r^2 score of 0.52.

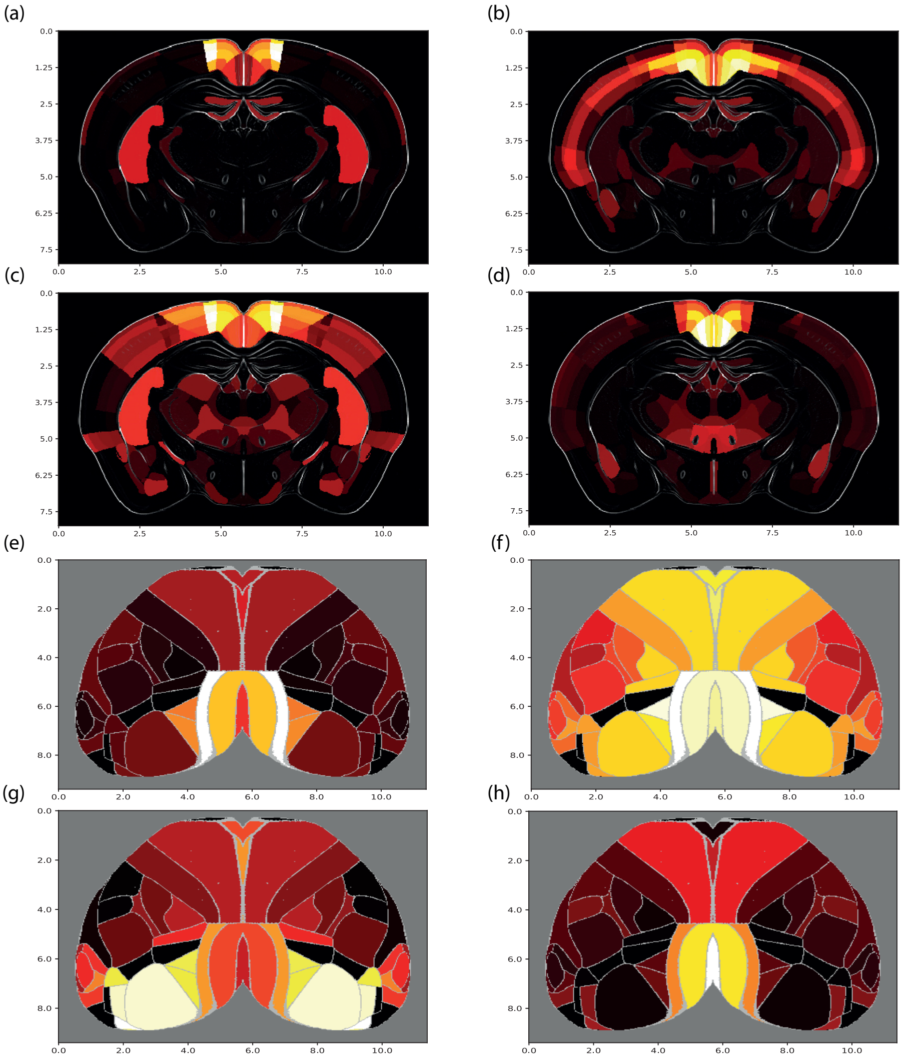


Fig. 11: (a-d) Subcortical and cortical visualizations (a,e) for the Cux2-IRES-Cre RSPagl tracing experiment compared to subcortical and cortical visualizations of spatial gene modules 9 (b,f), 70 (c,g) and 88 (d,h). The subcortical projection patterns were visualized using coronal slices of the projection or module volume (a,b), whereas the cortical projection or module patterns (c,d), were projected onto a flatmap and their values have been averaged over all cortical layers. The intensity of each plot was normalized by its maximum value. The scaling for both axes is in milimeters.

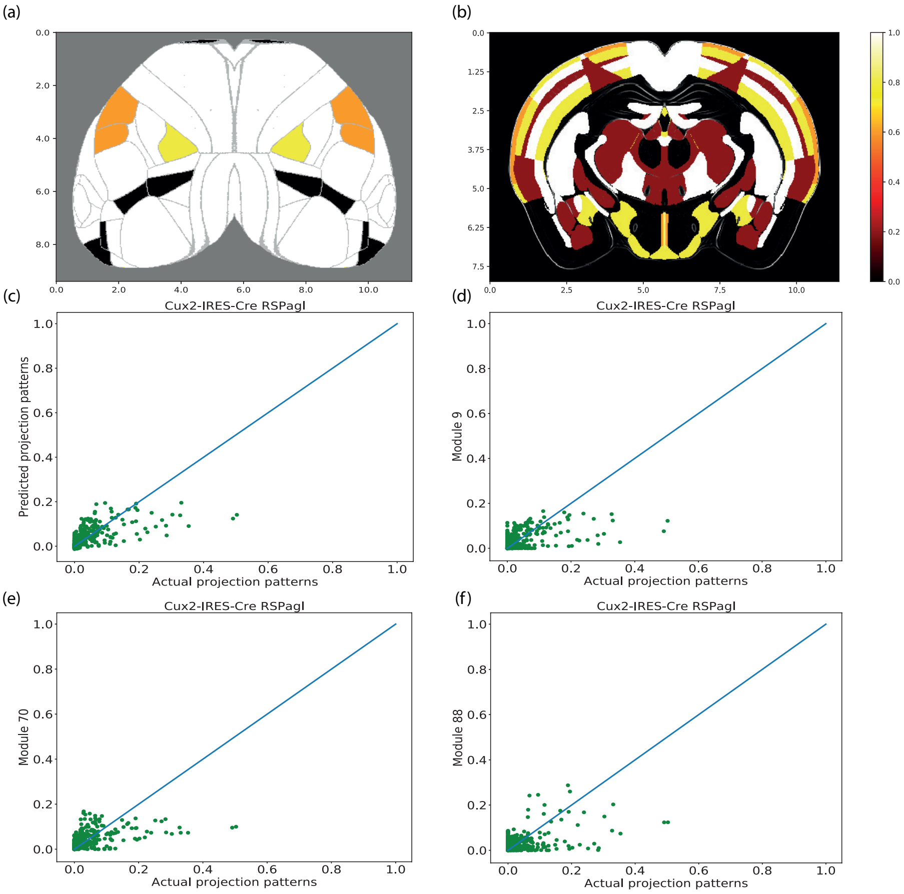


Fig. 12: Cortical (a) and subcortical visualization (b) of a spatial footprint related to a Cux2-IRES-Cre tracing experiment injected in the retrosplenial area, lateral agranular part (RSPagl). The spatial footprint represents the overlap that exists between the RPSagl experiment and modules 9, 70 and 88 with a pearson r of 0.51, 0.52 and 0.45 respectively. Each non-zero value across brain areas is replaced by 1.0 if it was present in all three modules and the projection pattern, 0.8 if present in two modules and the pattern, 0.6 if present in one module and the pattern, 0.4 if present in the pattern and absent in all modules and 0.2 if present in the modules but absent in the pattern. The subcortical projection pattern was visualized using coronal slices of the projection volume, whereas the cortical projection was projected onto a flatmap and its values were averaged over all cortical layers. The scaling for both axes is in milimeters. There is a strong presence of white (1.0), yellow (0.8) and orange (0.6) colors, that suggests a strong overlap between the experiment and the modules and which is also reflected by a r^2 score of 0.4 when the three modules are used for predicting the experiment. (c) Scatter plot between the projection pattern of the same experiment and its prediction by the 3 modules. (d-f) Similar scatter plots between the projection pattern and each module separately (d for module 9, e for module 70 and f for module 88). The solid line in each scatter plot is the diagonal, for which values across axes are equal. The scatter plots suggest that a combination of the three modules scaled by coefficients can lead to a more accurate prediction of the experiment than by the individual modules alone.

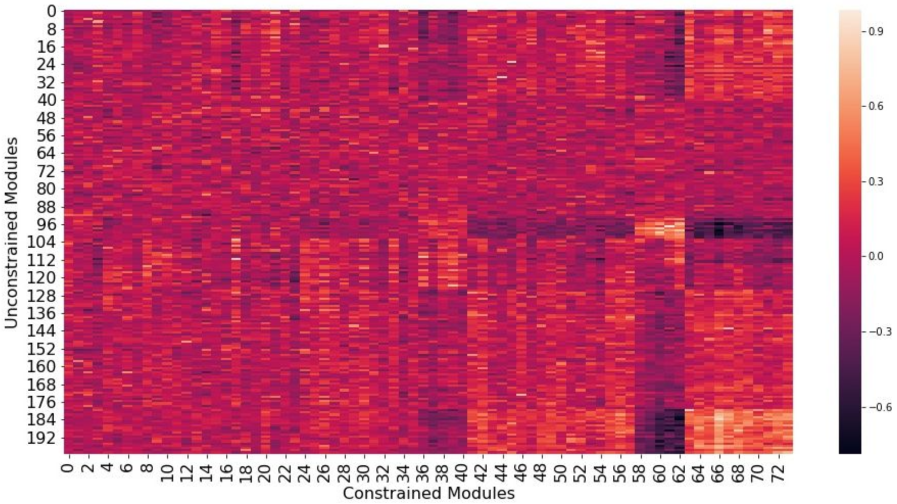
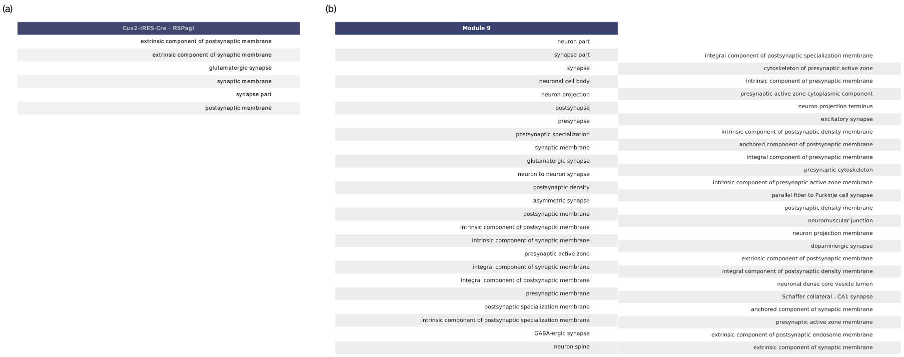


Fig. 13: Heatmap displaying correlations between the constrained and unconstrained spatial gene modules. The correlations were estimated with the Spearman's rank correlation coefficient. 58% out of 6600 correlations in total were considered significant (p $\leq$ 0.05). However, there are no evident correlation patterns between the two types of modules.


(a)

Cux2-IRES-Cre - RSPagl
extrinsic component of postsynaptic membrane
extrinsic component of synaptic membrane
glutamatergic synapse
synaptic membrane
synapse part
postsynaptic membrane

(b)

Module 9	
neuron part	
synapse part	integral component of postsynaptic specialization membrane
synapse	cytoskeleton of presynaptic active zone
neuronal cell body	intrinsic component of presynaptic membrane
neuron projection	presynaptic active zone cytoplasmic component
postsynapse	neuron projection terminus
presynapse	excitatory synapse
postsynaptic specialization	intrinsic component of postsynaptic density membrane
synaptic membrane	anchored component of postsynaptic membrane
glutamatergic synapse	integral component of presynaptic membrane
neuron to neuron synapse	presynaptic cytoskeleton
postsynaptic density	intrinsic component of presynaptic active zone membrane
asymmetric synapse	parallel fiber to Purkinje cell synapse
postsynaptic membrane	postsynaptic density membrane
intrinsic component of postsynaptic membrane	neuromuscular junction
intrinsic component of synaptic membrane	neuron projection membrane
presynaptic active zone	dopaminergic synapse
integral component of synaptic membrane	extrinsic component of postsynaptic membrane
integral component of postsynaptic membrane	integral component of postsynaptic density membrane
presynaptic membrane	neuronal dense core vesicle lumen
postsynaptic specialization membrane	Schaffer collateral - CA1 synapse
intrinsic component of postsynaptic specialization membrane	anchored component of synaptic membrane
GABA-ergic synapse	presynaptic active zone membrane
neuron spine	extrinsic component of postsynaptic endosome membrane
	extrinsic component of synaptic membrane

Fig. 14: An enrichment analysis reveals annotations of neurons and synapses for a spatial module and a tracing experiment. (a) Significant annotations for a Cux2-IRES-Cre experiment injected in the RSPagl area.

4 Discussion

In this study we built a predictive workflow, based on ridge regression and random forest based models, to predict axonal projection patterns in the mouse brain using gene expression data. Using the nested k-fold cross-validation technique, we obtained a median r^2 of 0.54 over 1397 tract-tracing experiments. In order to compare with previous studies (Ji et al., 2014), we developed an approach to make binary predictions and obtained similar performance to previous studies. Furthermore, we analyzed the spatial organization of genes in modules defined in different ways, based on the DLSC method, for determining links between gene expression and axonal projection patterns. We obtained median r^2 scores of 0.51 and 0.45 for the unconstrained and constrained approaches respectively. Finally, we applied gene ontology enrichment analysis to gene groups with high coefficient scores, and a substantial number of the groups were found to be associated with annotations related to postsynaptic function. In the following we will put the performance of the different cases in the context of previous studies, interpret our findings, suggest potential future work and discuss strengths, limitations and other applications of our pipeline.

The results of our study are consistent with the findings from the (Ji et al., 2014) study, specifically since our binary approach yielded a similar performance with a median 93% auROC value on wild-type data. In contrast to this study however, we did not rely on arbitrary thresholds for binarizing each tracing experiment to attain a 50% connectivity. Instead, we provided a data-driven estimation of the most optimal threshold value. In addition, we extended their analysis by including cre-line data that had not been subjected to such an analysis before.

When including both cre-line and wild-type data, we found a median auROC value of 0.89 across all 1397 tracing experiments. The increased performance of the models on binary predictions compared to continuous predictions is presumably due to the reduction of projection pattern related information which can therefore be more easily captured by the gene expression data (figure 8). However, binary connectivity descriptions do not inform the modeler about the strength of a projection. Hence, the continuous predictions are more suitable for analytic purposes. For that reason, we provided richer predictions of the mouse mesoconnectome by incorporating continuous patterns to our analysis (figure 6 for continuous predictions and figure 8 for binary ones).

Overall, our ridge regression models provided significant predictions, since they outperformed in every case the surrogate and the null models. This implies that gene expression contains information related to axonal projection patterns in the mouse brain. Regarding the variability of predictions, our statistical tests indicated that the difference in performance between cre-line and wild-type tracing experiments, quantified as r^2 score, was not statistically significant ($p = 0.004$ for 14000 random permutations). A possible explanation is that both wild-type and cre-line projection patterns fall within the range of predictions that can be covered by the gene expression data. Irrespective of explanation, the results show that the gene expression data contain enough information to also account for the more specific cre-line projection patterns.

The ridge regression models trained with spatial gene co-expression modules rather than expressions of individual genes, also outperformed corresponding surrogate and null models (figure 10). However, we found that such predictions were slightly less accurate on average than the gene expression based ones. Despite that, significant predictions of such models and strong correlations between axonal projections and spatial modules suggest that information related to axonal projections is present in modules of genes instead of being present in individual genes.

When comparing constrained modules with the unconstrained ones, we observed dissimilar patterns and an inferior performance for the constrained one when predicting tracing experiments. Such results suggest a lack of direct relation between spatial modules created exclusively by ISH data and modules that were constrained by single cell RNA sequencing data. A possible explanation is that distinct predictive modules were mixed when including all genes differentially expressed in the 74 cell-types, which suggests that better performance could be reached when selecting a subset from amongst them.

Regarding gene ontology enrichment analysis, a substantial number of tracing experiments (70%) and all unconstrained spatial modules (100%) were statistically associated with postsynaptic function. This may suggest that a potential causal link between axonal projections and gene expression in the mouse brain could be gene co-expression modules with a postsynaptic function and specific spatial footprints. This suggestion is consistent with the findings of (Roy et al., 2018), according to which presynaptic and postsynaptic locations have a particular protein profile. These profiles are partially reflected in gene expression data by locally expressed genes at axonal release sites (Glock et al., 2017; Cajigas et al., 2012; Holt and Schuman, 2013). Nevertheless, the causal links are far from being clear and will thus require further work.

A strength of this study was the inclusion of layer and cell-class specific patterns by including cre-line data to our analysis. To our knowledge, this is the first study that predicts brain-wide and cell-class specific projection patterns from gene expression data. Another advantage of this study was that it went beyond solely providing a predictive workflow, and it focused on discovering links between the two data modalities by analyzing the spatial organizations of genes with the dictionary learning and sparse coding technique (Li et al., 2017) and with gene ontology enrichment analysis (Rivals et al., 2007).

We acknowledge some limitations. First, 0.7% of ridge regression based models had an r^2 score close to zero. This could be attributed to parameters being optimized over all tracing experiments belonging to one cre-line or the wild-type category rather than for each experiment (injection) separately. Therefore, it can be expected that performance will be reduced when multiple projection patterns with a noisy subset need to be predicted simultaneously.

Another explanation could be that including the genetic information of target areas without its relation to source areas has limited capacity in predicting projection patterns. According to (Fulcher and Fornito, 2014), coupled gene expression patterns were shown to be directly linked with the large-scale topology of the mouse mesoconnectome. Furthermore, in (Bleakley et al., 2007) they used the support vector machine algorithm with kernels that coupled the feature vectors of nodes, for inferring the edges of biological networks. As a recommendation for future work, we can adapt this strategy to couple source and target based gene expression patterns and infer their corresponding axonal projections.

Another limitation is that unionization of data leads to information loss, not-a-number values and projection bias because of diversity in sizes of source brain areas. For that reason we will focus our future analyses on the volumetric gene expression and axonal projections data, as to avoid such issues and provide a finer grained predictive pipeline.

Furthermore, ridge regression and random forest based models provided significant predictions of axonal projections from gene expression data, but they are not capable of explicitly modeling the joint distribution between the two data modalities. Such explicit modeling could be advantageous in the case of training models to predict cellular resolved projections since data that could serve as training labels, such as single-neuron axonal reconstruction data, are still limited (Economo et al., 2019; Winnubst et al., 2019). Future directions might include incorporating generative probabilistic models, since models such as the infinite relational model have been successful in capturing the distributions of various connectomes such as the C.elegans connectome and the mouse retina microcircuit (Jonas and Kording, 2015; Ambrosen et al., 2013; Hinne et al., 2014, 2017; Betzel and Bassett, 2017).

Whole brain cellular resolved connections have yet to be described. The capability of our models to provide information for a more faithful reconstruction of the connectome at this resolution will depend on two factors. The first factor will be the ability to incorporate new advances in neuroanatomy and translational neuroscience, such as single-cell RNA sequencing and light sheet fluorescence microscopy (Tasic, 2018; Corsetti et al., 2019; Rolnick and Dyer, 2019).

The second factor will be the ability to mine at a higher spatial resolution from already tested data modalities such as in-situ hybridization based gene expression data. For this factor we will need to adapt additional computational tools for use in our pipeline. One potential tool is spatial point process analysis, which has successfully been used to extract spatially distributed counts of cells and synapses from modalities such as Nissl-stained brain images (LaGrow et al., 2018; Anton-Sanchez et al., 2014).

Translational neuroscientists could benefit from the use of our predictive workflow. A potential use case could include neuroscientists that study the effect of genes in the cognitive processes of the mouse brain. An example gene could be parvalbumin (PV) which according to (Nakazawa et al., 2012) has been linked to schizophrenia. Our workflow can then be used for studying the effect of altering the level and patterns of PV expression on the mesoconnectome and the resulting brain activity, which can in turn by validated by electrophysiological experiments.

Descriptions of potential use-cases for our predictive workflow together with their associated python code can be found online at the HBP Collaboratory and at Github (see table 1). The use cases are intended to be available via the EBRAINS infrastructure, provided as part of the EU-funded Human Brain Project.

Taken together, we have demonstrated a predictive workflow that can further be used to perform multimodal data integration to improve the accuracy of the predicted mouse mesoconnectome using gene expression data and support other neuroscience use cases.

Acknowledgment

Parts of this research have received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under Grant Agreement No

785907 (Human Brain Project SGA2) and the FLAG ERA project FIIND (NWO054-15-104).

References

Amann R, Fuchs BM (2008) Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques. Nature Reviews Microbiology 6:339–348, DOI https://doi.org/10.1038/nrmicro1888

Ambrosen KS, Herlau T, Dyrby T, Schmidt MN, Mørup M (2013) Comparing structural brain connectivity by the infinite relational model. In: Proceedings of the 3rd International Workshop on Pattern Recognition in Neuroimaging (PRNI 2013, pp 50–53, DOI http://dx.doi.org/10.1109/PRNI.2013.22

Ambrosius WT (2007) Topics in Biostatistics, Methods in Molecular Biology, vol 404, 2nd edn. Springer, DOI https://doi.org/10.1007/978-1-59745-530-5

Anton-Sanchez L, Bielza C, Merchán-Pérez A, Rodríguez JR, De Felipe J, Larrañaga P (2014) Three-dimensional distribution of cortical synapses: a replicated point pattern-based analysis. frontiers in Neuroanatomy 8:85, DOI https://doi.org/10.3389/fnana.2014.00085

Baruch L, Itzkovitz S, Golan Mashiach M, Shapiro E, Segal E (2008) Using expression profiles of caenorhabditis elegans neurons to identify genes that mediate synaptic connectivity. PLoS Computational Biology 4:e1000120, DOI https://doi.org/10.1371/journal.pcbi.1000120

Betzel RF, Bassett DS (2017) Generative models for network neuroscience: prospects and promise. R Soc Interface 14(136):20170623, DOI https://doi.org/10.1098/rsif.2017.0623

Betzel RF, Avena-Koenigsberger A, Goñi J, He Y, de Reus MA, Griffa A, Vértes PE, Mišic B, Thirane JP, Hagmann P, van den Heuvel M, Zuo XN, Bullmore ET, Sporns O (2015a) Generative models of the human connectome. Neuroimage 124(A):1054–1064, DOI https://doi.org/10.1016/j.neuroimage.2015.09.041

Betzel RF, Medaglia JD, Bassett DS (2015b) Diversity of meso-scale architecture in human and non-human connectomes. Nature communications 9:346, DOI https://doi.org/10.1038/s41467-017-02681-z

Bishop CM (2006) Pattern Recognition and Machine Learning, 1st edn. Information Science and Statistics, Springer

Bleakley K, Biau G, , Vert JP (2007) Supervised reconstruction of biological networks with local models. Bioninformatics 23:57–65, DOI http://dx.doi.org/10.1093/bioinformatics/btm204

Breiman L (2001) Random forests. Machine Learning 45(1):5–32, DOI https://doi.org/10.1023/A:1010933404324

Cajigas IJ, Tushev G, Will TJ, tom Dieck S, Fuerst N, Schuman EM (2012) The local transcriptome in the synaptic neuropil revealed by deep sequencing and high-resolution imaging. Neuron 3:453–466

Choi H, Mihalas S (2019) Synchronization dependent on spatial structures of a mesoscopic whole-brain network. PLoS computational biology 15(4):e1006978, DOI https://doi.org/10.1371/journal.pcbi.1006978

Corsetti S, Gunn-Moore F, Dholakia K (2019) Light sheet fluorescence microscopy for neuroscience. Journal of Neuroscience Methods 319(1):16–27, DOI https://doi.org/10.1016/j.jneumeth.2018.07.011

Dietterich TG (2000) Ensemble methods in machine learning. In: Proceedings of the First International Workshop on Multiple Classifier Systems, pp 1–15

Dodge Y (2008) The Concise Encyclopedia of Statistics, 1st edn. Springer

Economo MN, Winnubst J, Bas E, Ferreira TA, Chandrashekar J (2019) Single-neuron axonal reconstruction: The search for a wiring diagram of the brain. Journal of Comparative Neurology pp 1–10, DOI https://doi.org/10.1002/cne.24674

Fawcett T (2006) An introduction to roc analysis. Pattern Recognition Letter 27:861–874, DOI https://doi.org/10.1016/j.patrec.2005.10.010

Fornito A, Arnatkevičiūtė A, Fulcher BD (2019) Bridging the gap between connectome and transcriptome. Trends in Cognitive Sciences 23(1):34–50, DOI https://doi.org/10.1016/j.tics.2018.10.005

French L, Pavlidis P (2011) Relationships between gene expression and brain wiring in the adult mouse brain. PLoS Comput Biol 7:e1001049, DOI https://doi.org/10.1371/journal.pcbi.1001049

French L, Tan PPC, Pavlidis P (2011) Large-scale analysis of gene expression and connectivity in the mouse brain: insights through data integration. Frontiers in Neuroinformatics 5:12, DOI https://doi.org/10.3389/fninf.2011.00012

Friedman J, Hastie T, Tibshirani R (2009) The Elements of Statistical Learning. Data Mining, Inference, and Prediction, 2nd edn. Springer Series in Statistics, Springer

Fulcher BD, Fornito A (2014) A transcriptional signature of hub connectivity in the mouse connectome. PNAS 113(5):1435–1440, DOI https://doi.org/10.1073/pnas.1513302113

Glock C, Heumuller M, Schuman EM (2017) mrna transport & local translation in neurons. Current Opinion in Neurobiology 45:169–177

Harris JA, et al. (2014) Anatomical characterization of cre driver mice for neural circuit mapping and manipulation. Front Neural Circuits 8:1–16, DOI https://doi.org/10.3389/fncir.2014.00076

Harris JA, et al. (2018) The organization of intracortical connections by layer and cell class in the mouse brain. bioRxiv DOI https://doi.org/10.1101/292961

Highley JR, Esiri MM, McDonald B, Cortina-Borja M, Herron BM, Crow TJ (1999) The size and fibre composition of the corpus callosum with respect to gender and schizophrenia: A post-mortem study. Brain 122(1):99–110, DOI https://doi.org/10.1093/brain/122.1.99

Hinne M, Ambrogioni L, Janssen RJ, Heskes T, van Gerven MAJ (2014) Structurally-informed bayesian functional connectivity analysis. NeuroImage 86:294–305, DOI https://doi.org/10.1016/j.neuroimage.2013.09.075

Hinne M, Meijers A, Bakker R, Tiesinga PHE, Mørup M, van Gerven MAJ (2017) The missing link: Predicting connectomes from noisy and partially observed tract tracing data. PLoS Comput Biol 13(4):e1005478, DOI https://doi.org/10.1371/journal.pcbi.1005478

Holt CE, Schuman EM (2013) The central dogma decentralized: New perspectives on rna function and local translation in neurons. Neuron 80:648–657

Ji S, Fakhry A, Deng H (2014) Integrative analysis of the connectivity and gene expression atlases in the mouse brain. Neuroimage 84:245–253, DOI https://doi.org/10.1016/j.neuroimage.2013.08.049

Jonas E, Kording K (2015) Automatic discovery of cell types and microcircuitry from neural connectomics. eLife DOI https://doi.org/10.7554/eLife.04250

Kaufman A, Dror G, Meilijson I, Ruppin E (2006) Gene expression of caenorhabditis elegans neurons carries information on their synaptic connectivity. PLoS Comput Biol 2:e167, DOI https://doi.org/10.1371/journal.pcbi.0020167

Kluger Y, Basri R, Chang JT, Gerstein M (2003) Spectral biclustering of microarray cancer data: Co-clustering genes and conditions. Genome Research 13:703–716

Knox JE, Harris KD, Graddis N, Whitesell JD (2018) High resolution data-driven model of the mouse connectome. network neuroscience. Neuroscience 3(1):217–236, DOI https://doi.org/10.1162/netn_a_00066

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th international joint conference on Artificial intelligence (IJCAI, vol 2, pp 1137–1143

Kötter R (2007) Anatomical Concepts of Brain Connectivity. Handbook of Brain Connectivity, Springer

LaGrow TJ, Moore MG, Prasad JA, Davenport MA, Dyer EL (2018) Approximating cellular densities from high-resolution neuroanatomical imaging data. In: 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC, DOI https://doi.org/10.1109/EMBC.2018.8512220

Lanciego JL, Wouterlood FG (2011) A half century of experimental neuroanatomical tracing. Journal of Chemical Neuroanatomy 42(3):157–183, DOI https://doi.org/10.1016/j.jchemneu.2011.07.001

Langfelder P, Horvath S (2008) Wgcna: an r package for weighted correlation network analysis. BMC Bioinformatics 9:559, DOI https://doi.org/10.1186/1471-2105-9-559

Lee WCA, Bonin V, Reed M, Graham BJ, Hood G, Glattfelder K, Reid RC (2016) Anatomy and function of an excitatory network in the visual cortex. Nature 532(1):370–374, DOI https://doi.org/10.1038/nature17192

Lein ES, et al. (2007) Genome-wide atlas of gene expression in the adult mouse brain. Nature 445:168–176, DOI https://doi.org/10.1038/nature05453

Li Y, Chen H, Jiang X, Li X, Lv J, Peng H, Tsien J, Liu T (2017) Discover mouse gene coexpression landscapes using dictionary learning and sparse coding. Brain Structure and Function 222(9):4253–4270, DOI https://doi.org/10.1007/s00429-017-1460-9

Mairal J, Bach F, Ponce J, Sapiro G (2010) Online learning for matrix factorization and sparse coding. Journal of Machine Learning Research 11:19–60

Mancarci BO, Toker L, Tripathy S, Li B, Rocco B, Sibille E, Pavlidis P (2017) Cross-laboratory analysis of brain cell type transcriptomes with applications to interpretation of bulk tissue data. eNeuro 4(6):0212–0217

Markram H (2006) The blue brain project. Nature Reviews Neuroscience 7:153–160, DOI https://doi.org/10.1038/nrn1848

Markram H, Meier K, Lippert T, Grillner S, Frackowiak R, Dehaene S (2011) Introducing the human brain project. Procedia Computer Science 7:39–42, DOI https://doi.org/10.1016/j.procs.2011.12.015

Nakazawa K, Zsiros V, Jiang Z, Nakao K, Kolata S, Zhang S, Belforte JE (2012) Gabaergic interneuron origin of schizophrenia pathophysiology. Bioinformatics 62(3):1574–1583, DOI https://doi.org/10.1016/j.neuropharm.2011.01.022

Oh SW, et al. (2014) A mesoscale connectome of the mouse brain. Nature 508:207–214, DOI https://doi.org/10.1038/nature13186

Rice JA (2007) Mathematical Statistics and Data Analysis, 3rd edn. Mathematics of Computation, Duxbury Press

Ritter P, Schirner M, McIntosh AR, Jirsa VK (2013) The virtual brain integrates computational modeling and multimodal neuroimaging. Brain Connectivity 3(2):121–145, DOI https://doi.org/10.1089/brain.2012.0120

Rivals I, Personnaz L, Taing L, Potier MC (2007) Enrichment or depletion of a go category within a class of genes: which test? Bioinformatics 23(4):401–407, DOI https://doi.org/10.1186/1471-2105-7-91

Rolnick D, Dyer EL (2019) Generative models and abstractions for large-scale neuroanatomy datasets. Current Opinion in Neurobiology 55:112–120, DOI https://doi.org/10.1016/j.conb.2019.02.005

Rousseeuw P (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20:53–65, DOI 10.1016/0377-0427(87)90125-7

Roy M, Sorokina O, McLean C, Tapia-González S, DeFelipe J, Armstrong JD, Grant S (2018) Regional diversity in the postsynaptic proteome of the mouse brain. Proteomes 6(3):31, DOI https://doi.org/10.3390/proteomes6030031

Rubinov M, Ypma RJF, Watson C, Bullmore ET (2015) Wiring cost and topological participation of the mouse brain connectome. PNAS 112(32):10032–10037, DOI https://doi.org/10.1073/pnas.1420315112

Sanz Leon P, Knock SA, Woodman MM, Domide L, Mersmann J, McIntosh AR, Jirsa V (2013) The virtual brain: a simulator of primate brain network dynamics. Frontiers in Neuroinformatics 7:10, DOI https://doi.org/10.3389/fninf.2013.00010

Sanz-Leon P, Knock SA, Woodman MM, Domide L, Mersmann J, McIntosh AR, Jirsa VK (2013) The virtual brain: a simulator of primate brain network dynamics. Frontiers in Neuroinformatics 7:10, DOI https://doi.org/10.3389/fninf.2013.00010

Shendure J, Ji H (2008) Next-generation dna sequencing. Nature biotechnology 26:1135–1145, DOI https://doi.org/10.1038/nbt1486

Sperry RW ((1963) Chemoaffinity in the orderly growth of nerve fiber patterns and connections. PNAS 50(4):703–710, DOI https://doi.org/10.1073/pnas.50.4.703

Sporns O (2011) Networks of the brain. The MIT Press 412

Sporns O, Tononi G, Kötter R (2005) The human connectome: A structural description of the human brain. PLoS Computational Biology 1(4):e42, DOI https://doi.org/10.1371/journal.pcbi.0010042

Tan PN, Steinbach M, Kumar V (2005) Introduction to Data Mining, 1st edn. Pearson

Tasic B (2018) Single cell transcriptomics in neuroscience: cell classification and beyond. Current Opinion in Neurobiology 50:242–249, DOI https://doi.org/10.1016/j.conb.2018.04.021

Tasic B, et al. (2016) Adult mouse cortical cell taxonomy by single cell transcriptomics. Nature Neuroscience 19(2):335–346, DOI https://doi.org/10.1038/nn.4216

Tikhonov AN, Arsenin VY (1977) Solution of Ill-posed Problems, 1st edn. Mathematics of Computation, Winston & Sons

Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7:91, DOI https://doi.org/10.1186/1471-2105-7-91

Winnubst J, Bas E, Ferreira TA, et al. (2019) Reconstruction of 1,000 projection neurons reveals new cell types and organization of long-range connectivity in the mouse brain. bioRxiv pp 1–10, DOI http://dx.doi.org/10.1101/537233

Wolf L, Goldberg C, Manor N, et al. (2011) Gene expression in the mouse brain is associated with its regional connectivity. PLoS Comput Biol 75:e1002040, DOI https://doi.org/10.1371/journal.pcbi.1002040

Woodman MM, Pezard L, Domide L, Knock S, Sanz Leon P, Mersmann J, McIntosh AR, Jirsa VK (2014) Integrating neuroinformatics tools in the virtual brain. Frontiers in Neuroinformatics 8:36, DOI https://doi.org/10.3389/fninf.2014.00036