Review article

Progress towards a cellularly resolved mouse mesoconnectome is empowered by data fusion and new neuroanatomy techniques

Nestor Timonidis *, Paul H.E. Tiesinga

Neuroinformatics department, Donders Centre for Neuroscience, Radboud University Nijmegen, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands

ARTICLE INFO

Keywords:
Neuroanatomy review
Data fusion
Computational framework
Cell-type specificity
Mouse mesoconnectome
Shared factorisation
Multi-modal clustering
Probabilistic inference
Spatial registration
Spatial transcriptomics and proteomics
Connectomics
Tract-tracing
In situ hybridization
Single-cell RNA sequencing
Morphological reconstructions
Barcode sequencing
Light-sheet microscopy
Diffusion tensor imaging

ABSTRACT

Over the past decade there has been a rapid improvement in techniques for obtaining large-scale cellular level data related to the mouse brain connectome. However, a detailed mapping of cell-type-specific projection patterns is lacking, which would, for instance, allow us to study the role of circuit motifs in cognitive processes. In this work, we review advanced neuroanatomical and data fusion techniques within the context of a proposed Multimodal Connectomic Integration Framework for augmenting the cellularly resolved mouse mesoconnectome.

First, we emphasize the importance of registering data modalities to a common reference atlas. We then review a number of novel experimental techniques that can provide data for characterizing cell-types in the mouse brain. Furthermore, we examine a number of data integration strategies, which involve fine-grained cell-type classification, spatial inference of cell densities, latent variable models for the mesoconnectome and multi-modal factorisation. Finally, we discuss a number of use cases which depend on connectome augmentation techniques, such as model simulations of functional connectivity and generating mechanistic hypotheses for animal disease models.

1. Introduction

Over the last decade there has been a rapid improvement in techniques that measure gene expression and structural connectivity data at a cellular resolution. At the same time, the wiring diagram of the brain or connectome has been characterized for multiple species at varying scales (Swanson and Lichtman, 2016), such as for C. elegans (Ward et al., 1975), Platynereis dumerilii (Randel et al., 2015), Drosophila melanogaster (Chiang et al., 2011), rat (Swanson, 2004), mouse (Oh et al., 2014; Zingg et al., 2014), rhesus monkey (Felleman and Essen, 1991) and human (Sporns et al., 2005). A description of the whole-brain mouse connectome at the meso-scale level of anatomical sub-regions (Oh et al., 2014), also referred to as the mesoconnectome, offers currently the best trade-off between coverage and resolution for a mammalian species. However, connections are not cell-type-specific, which does not allow us to study the global cell-type-specific dynamics of information transmission. Moreover, while the layer-specificity of intra-cortical projections has now been characterized (Harris et al., 2019), the exact laminar position of projecting neurons is still not precisely known.

There is a need for a computational framework to exploit the latest advances in neuroanatomy for augmenting the mouse mesoconnectome using data fusion. Based on the currently available experimental approaches, data repositories and analytic techniques, we consider it feasible to construct a cell-type-specific mouse mesoconnectome via registration to a volumetric atlas followed by data fusion. This allows for posing relevant questions and extracting related information, with a number of examples presented as follows. Firstly, whole-brain mean field approximations can become more biologically plausible by being constrained with densities of specific cellular populations (Freestone et al., 2017; di Volo et al., 2019; Yochum and Modolo, 2020). Secondly, whole-brain models that link structural to functional connectivity (Hlinka and Coombes, 2012; Choi and Mihalas, 2019; Melozzi et al., 2019), can be updated with cell-type-specific projection patterns to compute the correspondingly more spatially-resolved functional

* Corresponding author.
E-mail address: n.timonidis@donders.ru.nl (N. Timonidis).

https://doi.org/10.1016/j.neubiorev.2021.06.016
Received 15 December 2020; Received in revised form 2 April 2021; Accepted 5 June 2021
Available online 10 June 2021

© 2021 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
connectome or activity patterns (Chevée et al., 2018; Kim et al., 2020). Additionally, data obtained from single-cell RNA sequencing, neuronal reconstructions and barcode sequencing approaches can be related to the global projection patterns of the respective cell-types, thus connecting local to global patterns (Gal et al., 2017).

In this work we propose a relevant framework, namely Multimodal Connectomic Integration Framework (MCIF), and we anticipate its development within the next five years. This will make possible to improve several features of the connectome, such as including the number of axonal fibers and the density of axonal arbor endings in long-range projections between cell-type-specific populations. MCIF will be crucial for connectomics research until the development of more precise neuroanatomical techniques that provide direct access at sufficient coverage. Therefore, we aim to present here the state-of-the-art developments, specifically the relevant acquisition and integration strategies.

The outline of this review follows the structure of the MCIF framework, as shown in Fig. 1. The steps of MCIF are being discussed in the order of appearance from top to bottom. We only make an exception for the Atlas registration step and we discuss it first due to its conceptual importance (see Section 2), despite being the middle step. In Section 3 we categorize the modalities of interest based on their properties and we highlight related works. In Section 4 we discuss techniques that have been used in literature to integrate these modalities. Finally, in Section 5 we suggest use cases of the data fusion techniques that could be beneficial to experimental and theoretical neuroscientists.

2. Reference space

Reference atlases serve as a common anatomical framework for datasets that have been acquired in brain space and are available in 2D or 3D versions (Wang et al., 2020). Given a template brain, two dimensional atlases comprise a concatenation of brain images section according to a fixed plane or flatmaps to represent highly convoluted surfaces. In contrast, three dimensional atlases digitally parcellate and annotate the brain directly in 3D space. For that reason, 3D atlases are generally preferred for data integration and related workflows (Wang et al., 2020).

Examples of 2D reference atlases include the Mouse Brain in Stereotaxic Coordinates (MBSC) (Paxinos and Franklin, 2001) and the Allen Reference Atlas (ARA) (Dong, 2008), while a recently established brain flatmap is the adult mouse brain flatmap version 1.0 (MsBF1) (Hahn et al., 2020). An example 3D reference atlas is the Allen Mouse Brain CCF (Table 2), several versions of which have been released over the last 15 years. A significant advancement in brain coverage and parcellation was achieved with each version, starting from 200 μm and reaching up to a 10 μm isotropic resolution. See Table 1 for more details regarding the different CCF versions. The CCF v3.0 template was created by averaging the volumes of 1675 adult mouse brains (Wang et al., 2020). Development of the 3D mouse brain atlas is still a work in progress: CCF v3.0 cannot be assessed in terms of stereotaxic coordinates yet. Moreover, the annotation of specific brain regions is still being fine tuned (Chon et al., 2019; Wang et al., 2020).

The registration of data modalities to a common reference atlas is a crucial step for data fusion and for augmenting the mouse mesocnectome. It is necessary for actions related to the processing of multimodal data, such as visualization, integration, prediction and post-hoc statistical analyses. These analyses can lead to finding meaningful links between connectome-related modalities and contribute to connectome augmentation.

Hence, in this work we consider registration to the CCF v3.0 as the most important step towards the mouse connectome augmentation. In the following sections we shall use the term compatibility to quantify the minimum number of steps needed for a dataset of interest to be registered to CCF v3.0. For instance, maximum compatibility would mean that the dataset has already been registered to this reference atlas. Similarly, the term registration shall mean registration to CCF v3.0.

| Table 1 Description of the different versions of the Allen Mouse Brain Common Coordinate Framework (CCF). |
| --- |
| **Release year** | **Resolution** | **Sample size** | **Hemisphere** |
| CCF v1.0 | 2005 | 200 μm | 200 structures | ipsilateral |
| CCF v2.0 | 2011 | 100 μm | 860 structures | ipsi and contralateral, symmetric |
| CCF v3.0 | 2017 | 10 μm, isotropic | 1324 structures | ipsi and contralateral, symmetric |

Fig. 1. Schematic overview of the Multimodal Connectomic Integration Framework (MCIF) that has been proposed in this review. For simplicity, we partition the key data resources into two groups: cellularly resolved data without tissue preservation and volumetric data that contain information about axonal projection patterns (see Section 3 for a more detailed description of the resources). Each box corresponds to a step in the workflow of MCIF, for which we discuss related computational strategies in separate subsections of Section 3. Boxes “Modality group 1: cellular level” and “Modality group 2: volumetric, axonal projection patterns” correspond to Section 3. The boxes that start from “Cell-type Classification” and end with “Shared Factorisation” correspond to Section 4, with the exception of “Atlas registration/overlay”, which corresponds to Section 2. The boxes “Post-hoc: Activity Simulation” and “Post-hoc: hypotheses generators/next steps” correspond to Section 5. The comments next to each arrow provide the motivation for proceeding to the following step. Given the structure of the workflow, steps “Cell-type classification” and “Spatial cell density inference” are specific to the first modality group (Sections 4.1 and 4.2), while step “Connectivity augmentation” is specific to the second modality group (Section 4.3). The last two steps correspond to analyses that follow data fusion and they can thus be performed in parallel.
Volumetric datasets are highly suitable for registration since it often is a matter of applying the proper affine transformation to the ARA co-ordinates. In contrast, datasets which are not spatially registered will require further work, which shall be discussed in Section 4.

3. Key data sources for connectomics

The relevant data modalities have been used for defining cell-types in the mouse brain based on its transcriptomic, morphological, projection or electrophysiological properties. Potential candidates are thus In situ hybridization (ISH) (Lein et al., 2007), sCRENA-seq (Tasic et al., 2016, 2018), synapto-specific proteomics (Roy et al., 2018; Zhu et al., 2018; Alvarez-Castelao et al., 2019), cell-densities, single-neuron full axonal reconstructions (Winnubst et al., 2019), transgenic cre driver line-based anterograde tracing (Harris et al., 2019), and high-resolution DTI (Calabrese et al., 2015). These techniques which will be discussed in this section, where Figs. 2–4 highlight a selection of schematic examples and Tables 3 and 4 contain quantitative details regarding their implementation.

3.1. Projection and tractography

Projection cell-classes refer to neuronal populations that are being distinguished in terms of their long-range projection patterns and often have a distinct functional role in feedforward and feedback information processing. Typical cortical projection classes are corticothalamic (CT), which mainly target ipsilateral thalamic structures, intratelencephalic (IT), which target ipsilateral and contralateral cortical and striatal structures, and pyramidal tract (PT), which target subcortical structures as well as thalamus, ipsilateral cortex and striatum to a lesser extent (Harris et al., 2019). Typical thalamic projection classes are core, which project to layer 4 cortical neurons, intralaminar, which project inside thalamus, and matrix, which project to layer 1 neurons (Harris et al., 2019).

The typical approach for quantifying cell-class-specific projections is through the use of virus-based anatomical tract-tracing techniques in mouse cre-driver lines (Harris et al., 2014). Cre-driver lines refer to mice expressing cre recombinase in specific genetically defined cell classes, which allows for labeling excitatory neurons in specific cortical layers that belong to a specific cell-class (Harris et al., 2019). Anatomical tract-tracing is an invasive method used in animal studies to trace axonal projections between brain areas (Sorns et al., 2005; Kötter, 2007; Lanciego and Wouterlood, 2011), either in the retrograde or anterograde direction. In anterograde tract-tracing, a source brain area is injected by a virus tagged coding for fluorescent molecules, such as green fluorescent protein or lipophilic dye, which are transported to the target brain areas through the axons and to label their terminals (Chamberlin et al., 1998; Harris et al., 2012). Retrograde tract-tracing follows the reverse direction with the virus being transported from the axonal terminals of the target area across synapses to the source areas (Katz et al., 1984). Usually, the tracing process is visualized using a microscopy technique such as two-photon microscopy (Oh et al., 2014).

An extensive collection of anterograde tracing experiments spanning the mouse mesoconnectome can be found in the Allen Mouse Brain Connectivity Atlas (AMBCA) dataset (https://help.brain-map.org/display/mouseconnectivity/Documentation). This dataset is comprised of ~1600 anterograde tract-tracing experiments performed in 56-days-old (P56) male mice (Oh et al., 2014; Harris et al., 2019). From these experiments, 498 were C57BL/6J (wild-type) and the rest belonged to 46 major cre-driver lines (see Fig. 3a).

Projections were visualized using two-photon microscopy, which produced brain slice images at 1 μm resolution (Oh et al., 2014; Harris et al., 2019). Moreover, all produced images were aligned and registered to the Allen Reference Atlas (Wang et al., 2020), hence generating whole-brain volumetric data available at resolutions ranging from 10 μm³ to 100 μm³. The Mouse Connectivity Cache (MCC) API allows for downloading this data at the aforementioned levels of resolution and at the unionized level, in which the voxels have been averaged across brain areas (Oh et al., 2014).

While AMBCA is considered to be the most complete of tract-tracing studies of the mouse mesoconnectome, some studies have presented alternative interpretations of the underlying data. According to (Oh et al., 2014), the density of inter-areal connections was in the range 35–54%. However, the application of a different statistical model to the same wild-type experiments yielded a 77% inter-areal density instead (Ypma and Bullmore, 2016). Furthermore, a recent study used retrograde labeling to map inter-cortical connections in the mouse brain and found a connection density of 99% (Gamanut et al., 2018). These results suggest that a number of false negatives might be present in the inter-areal connectivity matrix generated by Oh et al. (2014).

Anterograde labeling of a single neuron may only incompletely label the axonal tree leaving out important parts of the projection. This cannot be solved by bulk labeling, as this yields too much labeling in target areas, which makes it impossible to recognize the single-cell projection motifs demonstrated in Han et al. (2018). This imprecision can be addressed with two techniques. First, for projections that involve an intermediate area, such as interareal cortical projections via thalamus, can be tracked using an injection of anterograde and retrograde tracer (Deller et al., 2000; Zingg et al., 2014). This allows the input to the intermediate area to be linked to where the corresponding output is sent. Second, trans-synaptic techniques label cell bodies and thus do not suffer from incomplete axonal labeling. Specific post-synaptic cell-type targeting can be achieved by trans-synaptic tracing using the pseudotyped rabies virus (PRV) (DeFalco et al., 2001; Ekstrand et al., 2008; Wall et al., 2010). This virus has been successfully used in retrograde experiments, by constructing a virus encoding enhanced green
fluorescent protein (EGFP) that can infect directly connected cells while being unable to spread beyond these cells as it can only cross one synapse (Wickersham et al., 2007; Ahrlund-Richter et al., 2019). Moreover, a successful application of trans-synaptic tracing in anterograde experiments was recently achieved using the adeno-associated virus (AAV) coupled with cre-dependent trans-gene expression in the selected target region (Zingg et al., 2017). This targeting approach provides post-synaptic cell-type specificity that is complementary to the pre-synaptic one achieved by PRV and is thus an ideal candidate for future anatomical studies (Harris et al., 2019).

Regardless of the points raised about the connection density inaccuracies across the different tract-tracing datasets, AMBCA spans the whole mesoconnectome up to the cell-class level. For that reason, it can serve as a structural template of cell-type-specific projections. Additional datasets will be needed to augment the mesoconnectome at various points where important information is missing, such as the number of axonal fibers, the densities of axonal endings and the number of pre-synaptic and post-synaptic puncta from-and-to fine-grained cell-types, such as the ones discovered in recent works by electrophysiological and transcriptomic means (Tasic et al., 2018; Gouwens et al., 2019, 2020).

An alternative to anatomical tracing for providing projection volumes, is the Diffusion Tensor Imaging (DTI) method (Le Bihan and Breton, 1985; Taylor and Bushell, 1985; Merboldt et al., 1985). DTI is a non-invasive in vivo technique that measures the diffusion of water molecules in brain tissues for mapping the white matter tracts of the brain (Mueller et al., 2015) (see Fig. 3b). It can also be applied ex vivo. The most recent studies have used state of the art DTI to provide tractography volumes at a microscopic resolution of 43 μm (Calabrese et al., 2015). The major advantage of DTI tractography is that it covers the entire brain of a subject, hence allowing for taking inter-subject variability into account.

However, DTI still generally has low spatial resolution on the millimetre scale, especially in humans, a low signal-to-noise ratio and a lack of directionality in inferred connections. Therefore, tract-tracing data is the state-of-the-art in quantifying meso-scale projection patterns (Fillard et al., 2011; Oh et al., 2014; Chen et al., 2015a).

3.2. Transcriptomic

Transcriptomic-defined cell-types refer to cellular populations that are characterized by specific transcriptional profiles, which is usually quantified by the mRNA levels of protein-coding genes (Poulin et al., 2016). A number of techniques that measure the transcriptional activity in the mouse brain on a cellular level include ISH (Lein et al., 2007), scRNA-seq (Tasic et al., 2016, 2018; Tasic, 2018) and pooled-cell microarray (Okaty et al., 2011; Mancarci et al., 2017) (see Fig. 2a and b, Table 2). From these techniques, ISH and scRNA-seq are capable of assessing the expression of thousands of genes, with the former providing a full brain spatial coverage and the latter providing cell-level resolution.

In the ISH technique, single-stranded RNA probes with
complementary nucleotide sequences are labeled with fluorescent dyes, digoxygenin or quantum dots (Carter and Shieh, 2015). The probes are first attached to the RNA of interest, which is usually followed by applying fluorescence microscopy to visualize the labels (Amann and Fuchs, 2008). Example datasets include the Allen Mouse Brain Atlas (AMBA) and Developing Mouse Brain Atlas (ADMBA), which contain the expression of ~20,000 genes in the adult mouse brain and 2000 genes related to brain development, respectively (Lein et al., 2007; Henry and Hohmann, 2012). A major advantage of incorporating ISH-based data-sets in data integration frameworks is the full brain coverage at appropriate volumetric resolutions, which enable analysing the spatial structure of gene expression in 3D brain space across all major brain regions. Moreover, AMBA and ADMBA have a large gene sample size that allows to determine gene modules that are related to particular brain regions and are significantly enriched in specific biological processes and molecular functions. Lastly, these datasets can be directly overlaid with connectomic and other volumetric datasets, since they have been registered. In the AMBA however, the 50 μm thick sections spaced at 200 μm intervals along the posterior-anterior axis resulted in missing data for multiple genes which would necessitate imputation strategies (Lein et al., 2007).

scRNA-seq is characterized by deep sequencing of the RNA from tissues of individual cells, thus enabling a level of resolution that is higher than hybridization or microarray approaches (Kolodziejczyk et al., 2015). This enables the detection of low level transcripts and a higher dynamical range for gene expression that can be used to define cell-types based on their transcriptional profiles.

While various approaches have been applied for deep single-cell mRNA sequencing in the mouse brain, they share the following steps. First, single-cells are being isolated, followed by the reverse
transcription of RNA to cDNA. Subsequently, cDNA is being amplified and sequenced, while the last step is analysing and validating the newly generated data using dimensionality reduction, clustering or visualization techniques (Poulin et al., 2016). In a recent exemplary study, a combination of deep scRNA-seq and retrograde labeling was used to identify 56 glutamatergic, 61 GABAergic and 16 non-neuronal cell-types in the mouse brain (Tasic et al., 2018). The most important contribution of this study is the characterization of most glutamatergic cell-types as localized in a few specific brain regions and of most GABAergic cell-types as globally distributed across brain areas. This helps us to better understand the molecular programs that underlie local and long-range connectivity in the mouse brain.

An additional novel contribution can be found in Sylwestrak et al. (2016): a combination of the CLARITY tissue-clearing technique (Chung et al., 2013; Ueda et al., 2020) with the EDC cross-linking strategy (Table 2) (Lee et al., 1996) allowed for simultaneously detecting multiple mRNAs and microRNAs in intact tissue, as well as providing their cellular location in a volumetric form (see Fig. 2f). The automatic generation of volumes is an advantage compared to other spatial approaches that produce brain slice images, which will thus require spatial alignment as an additional processing step. This helps us to better understand the molecular programs that underlie local and long-range connectivity in the mouse brain.

Fig. 4. Schematic representation of novel experimental techniques based on barcode tagging and RNA sequencing for identifying the electrophysiological, transcriptomic and projection properties of multiple cell-types simultaneously. (a) MAPseq and BARseq (Kebschull et al., 2016; Chen et al., 2019). (b) SYNseq (Peikon et al., 2017). (c) Sci-RNA-seq (Cao et al., 2017). (d) Div-Seq (Habib et al., 2016). (e) Patch-seq (Cadwell et al., 2016). See Section 3.5 for the description of each technique and Table 2 for the acronym explanations. Panel a is reprinted from Chen et al. (2019), copyright 2019, with permission from Elsevier. Panel b is reprinted from Peikon et al. (2017). Panel c is reprinted from Cao et al. (2017), with permission from AAAS. Panel d is reprinted from Habib et al. (2016), with permission from AAAS. Panel e is reprinted by permission from Springer Nature: Nature Biotechnology (Cadwell et al., 2016), copyright 2015.

Given the above described gap between spatial and single-cell transcriptomic approaches in whole-brain experiments, it is crucial for MCIF to integrate both types of modalities in order to enjoy their benefits and mitigate their flaws, until promising techniques such as the aforementioned ones can provide high throughput whole-mouse-brain data (see Fig. 1). In Section 4.2, we shall discuss a potential strategy to infer the spatial location of cells and provide spatially registered transcriptional signatures of individual cells.
cell-types and testing biophysically plausible models of neuronal connectivity (Chklovskii, 2004; Hill et al., 2012; Economo et al., 2019; Winnubst et al., 2019). Additionally, their morphological reconstruction and transcriptomic data now identify a number of cell-types based on patch-clamp recordings. In Gouwens et al. (2019, 2020), they have demonstrated the integration of multiple types of modalities for finer-grained connectivity (Nandi et al., 2020). The established procedure for neuronal projection mapping of cell-types. In Gouwens et al. (2019, 2020), they demonstrated the integration of multiple types of modalities for finer-grained connectivity (Nandi et al., 2020). The established procedure for neuronal projection mapping of cell-types.

### 3.3. Morphology and electrophysiology

Neuronal morphology describes neurons in terms of the shape of their axons and dendrites, which provide an important constraint on the connectivity (Chklovskii, 2004; Hill et al., 2012; Economo et al., 2019; Reimann et al., 2019). Morphological reconstructions of individual neurons based on these features have been performed for defining cell-types and testing biophysically plausible models of neuronal connectivity (Nandi et al., 2020). The established procedure for neuronal reconstructions is comprised of labeling the selected neurons within the brain, followed by electron or light microscopy imaging of the tissue that contains the labeled neurons, and completed by tracing the arbors of reconstructed neurons in dense cortical circuits (Helmstaedter, 2013). Two example groups from LM techniques are Light sheet fluorescence microscopy (LSFM) and expansion microscopy (ExM) that enabled imaging of whole-brain tissue under analysis (Li et al., 2010; Porrero et al., 2016).

Morphological reconstructions comprise currently the best data source in bridging the gap between the micro-scale world of local microcircuits and the meso-scale world spanning the whole brain with projections of coarse grained cell-classes. A prime example was the capability to identify cell-types with hitherto unknown projection patterns, such as those in the zona incerta and subiculum neurons (Winnubst et al., 2019). Example repositories of reconstructed neurons can be found in the NeuroMorpho.org web archive (Ascoli et al., 2007) and in the MouseLight database (Economo et al., 2016). An exemplary dataset can be found in Winnubst et al. (2019), where they reconstructed the long range axons of ~1100 individual neurons originating in the motor cortex, thalamus, hypothalamus and subiculum (see Fig. 3d).

However, transcriptomics actually defines more distinct cell-types than morphology-based analyses (Tasic et al., 2018), which necessitates the integration of multiple types of modalities for finer-grained projection mapping of cell-types. In Gouwens et al. (2019, 2020), they identified a number of cell-types based on patch-clamp recordings. Additionally, their morphological reconstruction and transcriptomic profiling led to the identification of multiple morphological, morpho-electric and morpho-electric-transcriptomic cell-types. These studies introduced a multimodal unsupervised approach for delineating cell-types by finding common patterns across the molecular, electrophysiological and transcriptomic properties of the cells. The approach is extensible to additional datasets and has been tested in transgenic cre-line data. In Sections 3.5 and 4, novel techniques and methods will be discussed where other types of integration of these modalities have been explored for different species and scales.

### 3.4. Microscopy-based techniques

It has become feasible to access structural data volumetrically at high resolution in all directions through techniques based on high resolution microscopy. The two main branches of high resolution microscopy are Electron microscopy (EM) and Light microscopy (LM) techniques.

EM techniques comprise a popular group of techniques, capable of reaching sub-cellular resolutions and sampling high fraction of the neurons in dense cortical circuits (Helmstaedter, 2013). Two example EM techniques with high sub-cellular resolution are Speckle microscopy (SBE) and FIB-SEM (Knott et al., 2008) (Table 2), which are capable of reaching 25 nm resolution. However, EM techniques require hundreds to thousands of hours for analysis of dense circuits and tens of thousands of hours for their reconstruction. Given that circuit reconstruction presents a complicated data analysis problem, semi-automated and online crowd-sourcing approaches are considered more feasible than of fully automated algorithms (Helmstaedter, 2013). Given these issues, we will focus further on LM techniques, but readers can find more information about EM techniques in Helmstaedter (2013).

An exemplar group from LM techniques is Light sheet fluorescence microscopy (LSFM) combined with tissue clearing methods. LSFM has achieved whole-brain coverage and has been shown to lead to reduced levels of photodamage and photobleaching (Gonsetti et al., 2019). Specifically, when combined with various tissue clearing methods such as CLARITY, LSFM was able to reach a sub-cellular resolution up to 180 nm (Tomer et al., 2014). Two additional useful combinations were with the BABB technique, that enabled imaging of whole-brain

### Table 3

| Coverage                  | Resolution                  | Sample size                      | Can be registered to CCF v3.0 |
|---------------------------|-----------------------------|----------------------------------|------------------------------|
| MAPseq (Kebisch et al., 2016) | Single-region               | 995 neurons from LC, 553 neurons from V1 | Targets: yes. Sources: spatial location inference required |
| SYNseq (Peikon et al., 2017) | Tested in cultured primary hippocampal neurons | ~2 million cells | Yes |
| BARseq (Chen et al., 2019) | Single-region               | ~3500 neurons | Yes |
| BRISeq (Huang et al., 2020) | Mouse cortex                | 6 cortical areas × 12 areas (ipsi and contra) | Yes |
| Div-Seq (Habib et al., 2016) | Hippocampus (mouse)        | 1367 single nuclei | Inference of spatial location required |
| sci-RNA-seq (Cao et al., 2017) | Whole brain                | 50,000 cells | Tests on mice required |
| Patch-seq (Cadwell et al., 2016) | Neocortex (mouse)       | 58 cells | Yes |
| SYNMAP (Zhu et al., 2018) | Whole-brain (mouse)        | ~10^6 synapses | Yes |
| Proteomic Mass Spectrometry (Roy et al., 2018) | 7 major forebrain and hindbrain regions | 1173 proteins | Yes |
| qBrain (Kim et al., 2017) | Whole-brain (mouse)        | 7 GABAergic cell-types | Yes |
| MouseLight (Winnubst et al., 2019) | Whole brain                | ~1100 neurons | Yes |
| Gouwens approach (Gouwens et al., 2019) | Visual cortex (mouse)  | 461 neurons, 17 morphology, 38 electrophysiology and 46 morpho-electric cell-types | Yes |
| ISH (Lein et al., 2007) | Whole brain                | ~20,000 genes | Yes |
| Anatomical tract-tracing (Oh et al., 2014) | Whole brain                | ~1500 injections | Yes |
| LSFM combined with CLARITY (Tomer et al., 2014) | Whole brain                | Sub-cellular | Yes |
| LSFM combined with BABB (Dodd et al., 2007) | Whole brain                | Cellular – level | Yes |
| LSFM combined with ultimate DISCO (Pan et al., 2016) | Whole brain                | Dendritic spine-level | Yes |
| Expansion Microscopy (ExM) × Lattice Light-sheet microscopy (LLSM) (Gao et al., 2019) | Mouse cortex and entire Drosophila brain | ~60 × 60 × 90 - nm | Dendritic spines, synaptic proteins, Dopaminergic neurons | Yes |
GFP-expressing neurons and visualization of details of the hippocampus (Dodt et al., 2007), and with ultimate DISCO that enabled imaging of neuronal connections and vasculature within the entire brain of adult mice at the dendritic-spine-level (Pan et al., 2016) (Table 2).

Recent advances in tissue clearing, such as the iDISCO (Renier et al., 2014), CUBIC (Tainaka et al., 2018), FDISCO (Qi et al., 2019) and MACS techniques (Zhu et al., 2020) (Table 2), have addressed some shortcomings of the previous generation, allowing for clearing of larger volumes, preserving fluorescent labels better and for a longer time, the ability to be combined with multiplex viral labeling and to reduce the time needed to clear the tissue. In addition, specific image processing pipelines, such as clearmap (Renier et al., 2016) and tubemap (Kirst et al., 2020), have been developed for analyzing the obtained data volumes, which were demonstrated with stunning images of networks of arteries, veins and capillaries. These studies allowed for characterizing the effects of sensory-deprivation as well as artificially induced strokes in rodent models.

Expansion Microscopy (ExM) in combination with Lattice Light-sheet microscopy (LLSM), could reach a sub-cellular resolution and has been tested in the mouse neocortex and the entire Drosophila brain (Gao et al., 2019; Ueda et al., 2020). This approach is the most impressive in the LSFM collection yet, because it is ~700 and 1200 times faster than current super-resolution (SR) fluorescence microscopy and EM methods. In the mouse cerebral cortex, it delivered a collection of thousands of dendritic spines and synapses, both of which can serve as useful modalities for augmenting the mesoconnectome (see Fig. 3c).

An advantage of LSFM is that it provides high resolution volumes than can be registered to Allen CCF v3.0 via standard procedures (Perens et al., 2018), which is the highest that has been provided by the Allen Institute at the moment. Large volumes can be imaged within a time-scale of minutes at a micrometer resolution, but the sampling rate is inversely proportional to the imaging depth (Corsetti et al., 2019). Similarly to the Mouselight data, LSFM can help fill the meso-scale gap between anatomy tracing techniques and ultra-high resolution EM pipelines and corroborate cell-type-specific long-range projection patterns (Ueda et al., 2020). In addition, it will be used to image in-vivo the CNS development and provide longitudinal structural data in future works (Corsetti et al., 2019).

### 3.5. Molecular and barcode-based techniques

High-throughput DNA/RNA sequencing techniques can provide high resolution volumetric data. An exemplar group of this category is based on barcode tagging of neuronal populations, which can be extracted by in situ RNA sequencing and tracing of projections (Kebsschull et al., 2016). We could make a distinction between techniques that have been tested exclusively on invertebrates and techniques that have been tested in mice, given that they are often piloted in the former and then adopted for use in the latter.

The techniques tested in mice have achieved a varying degree of brain coverage. SYNMAP is a proteomic method with whole brain coverage applicable to mouse brains (Zhu et al., 2018) (see Fig. 2d). The pipeline of SYNMAP is comprised of synaptic-protein-tagging in knock-in mice, followed by tissue imaging using spinning disk confocal microscopy (Huang et al., 2010), image analysis and data storage. In Zhu et al. (2018), SYNMAP was used to analyse the post-synaptic protein

### Table 4
Feature descriptions of a number of experimental techniques: utility with respect to MCIF, features of the obtained data, time required for corresponding pipeline completion and advantages.

| Technique | Utility | Feature size | Time required | Pros |
|-----------|---------|-------------|---------------|------|
| MAPseq (Kebsschull et al., 2016) | Long-range cell-type specific connection patterns | Neuron-to-area binary matrix | 11 days | Fast, high-throughput, localized, low-cost, long-range |
| SYNseq (Peikon et al., 2017) | Technical obstacles still need to be overcome | Neuron-to-neuron binary matrix | Scale of weeks | Fast, localized, high-throughput, low-cost, Synaptic-connectivity |
| BARseq (Chen et al., 2019) | Long-range cell-type-specific connection patterns | Neuron-to-area binary matrix | Fast, high-throughput, localized, low-cost, Long-range |
| BRICseq (Huang et al., 2020) | Integration of meso-scale connectivity and gene expression per individual subject | Area-to-area connectivity matrix | <4 weeks | Fast, high-throughput, multiplex per individual subject, multi-modal |
| Div-Seq (Habib et al., 2016) | Longitudinal cell-type specific gene expression data for rare cell-types | Single nuclei × gene expression | Single-cell longitudinal data, newborn neurons |
| sci-RNA-seq (Gao et al., 2019) | Single-cell gene expression data for diverse and rare cell types | Gene expression × single-cell | 2 days for for −10,000 cells | Multiplex – low cost (0.035–0.20 $/cell), high throughput, fast |
| Patch-seq (Cadwell et al., 2016) | Access to fine-grained cell-types from multiple modalities | Cells × axonal arborization, cells × action potential amplitude, cells × genes | Integrating electrophysiology, morphology and transcriptomics for single-cell profiling |
| SYNMAP (Zhu et al., 2018) | Population-synapse density integration to the mouse mesoconnectome | Synapse densities to whole-brain regions | Scale of weeks | Compatible with multiple molecular labeling methods |
| Proteomic Mass Spectrometry (Roy et al., 2018) | Synapse density integration to the mouse mesoconnectome | Proteins × brain regions | Access to post-synaptic proteomic data |
| qbrain (Kim et al., 2017) | Cell-type-specific density integration | Single-cell densities | Direct cell-density integration |
| Mouselight (Winnubst et al., 2019) | Single neuron long range projections | Neurons × voxels | −1 week | Whole-brain single-neuron data |
| Gouwens approach (Gouwens et al., 2019) | Integrating patch-clamp recordings with morphological reconstructions | Cells × action potential, cells × morphology features | Cell-type classification using different data sources |
| ISH (Lein et al., 2007) | Spatial transcriptomic landscapes | Gene expression × voxels | |
| Anatomical tract-tracing (Oh et al., 2014) | Meso-scale bulk projection patterns | Injections × voxels | Scale of weeks | |
| Light-sheet microscopy (Corsetti et al., 2019) | Neuron morphology data integration | | |
| Expansion Microscopy (ExM) × Lattice Light-sheet microscopy (LLSM) (Gao et al., 2019) | | | |

---

**Revenue:**

- **Annual:**
  - **Volume:** ~10,000 cells
  - **Cost:** $0.03–0.20 $/cell

**Deployment:**

- **Scale:** Whole-brain in-vivo imaging
  - **Speed:** Varies from 1–100 Hz
  - **Accuracy:** ~100 person hours for Drosophila
- **Cost:** Low cost ($0.03–0.20 $/cell)

**Methodology:**

- **Specialized:** Bridges the gap between neural anatomy and ultra-high resolution-based EM pipelines
Neuroscience and Biobehavioral Reviews 128 (2021) 569–591

N. Timonidis and P.H.E. Tiesinga

577
densities of billion synapses across the whole mouse brain and to stratify the synapses into multiple subtypes. The major advantage of this method is that it provides whole-brain single-synapse data in an industrial fashion comparable to AMBCA and AMBA and it is compatible with registration (Zhu et al., 2018). A full SYNMAP pipeline, comprised of protein marker imaging, analysis and registration, is labour intensive and takes a few weeks to complete. Despite that, it is a valuable resource for integrating densities of various synapse-types with the mouse mesoconnectome.

Proteomic mass spectrometry (PMS) is another technique that has been used to acquire synapse-related data. In Roy et al. (2018), the expression patterns of multiple post-synaptic proteins were isolated from seven major forebrain and hindbrain regions in the mouse brain (see Fig. 2e). While the resolution is at the level of brain areas and is thus lower than aforementioned techniques, a recently developed protocol allows for the identification of cell-type-specific proteomes in mice (Alvarez-Castelao et al., 2019). Post-synaptic proteomic data can augment the mesoconnectome in a number of ways. First, proteomic data can either filter out non-post-synaptic genes or select genes of a particular synapse type. Second, the total post-synaptic expression per-brain-region can be used as a constraint when inferring connectivity patterns from specific synapse-types. Third, cell-type-specific proteomics can directly quantify the metabolic activity of a given cell-type instead of indirectly validating its metabolic context by using (KEGG) enrichment analysis (Ogata et al., 1999). These different analyses can result in more detailed representations of cell-types and steps to achieve this are discussed in Section 4.1.

Besides synaptic and proteomic data, a direct mapping of cell-densities could bypass issues caused by inaccuracies or missing data, such as due to the 200 µm section used in AMBA, as mentioned before (Lein et al., 2007). The qBrain resource developed in Kim et al. (2017), offers an alternative (see Fig. 2c). It comprises of a brain-wide cell-density acquisition pipeline that ranges from labelling of cell-types to imaging to volume registration and cell segmentation. In this study, they mapped the distribution and density of multiple GABAergic cell-types in the entire mouse brain. Data produced by this technique were registered and are thus compatible for data fusion with other registered modalities.

Currently, most of the barcode-tagging-based techniques are limited to single-regions in the mouse brain either because of limitations in data acquisition or analysis. Example approaches that preserve brain tissues are MAPseq (Kebschull et al., 2016; Han et al., 2018), Patch-seq (Cadwell et al., 2016), BARseq (Chen et al., 2019) and BRICseq (Huang et al., 2020) (Table 2 for the acronyms). MAPseq, BARseq and BRICseq are quite similar approaches, with the difference that MAPseq and BRICseq are characterized by in situ sequencing of target neurons, while in BARseq both source and target neurons are sequenced in situ (see Fig. 4a). Additionally, BRICseq enables multiplex barcoding, allowing the tracking of multiple projection patterns per subject and delivering subject-specific data. In Chen et al. (2019), ~ thousands of neurons were sequenced using BARseq, resulting in the identification of long-range and cell-type specific projection patterns. BARseq and MAPseq data can also be combined with two-photon imaging to produce axonal projections that can be registered together with the locations of source and target neurons.

Patch-seq is the first technique that has integrated scRNA-seq, patch-clamp recordings and morphological delineation in a single experiment (see Fig. 4e). In Cadwell et al. (2016), multiple cells in the mouse cortex were analysed according to their electrophysiological, morphological and gene expression properties. This was the first dataset that enabled features such as axonal arborization, action potential amplitude and gene expression to be measured in the same experiment. Brain coordinates are not preserved in Patch-seq, thus spatial inference will be needed for registration. The integration goal of this procedure would be to augment the mesoconnectome with cell-types that have been corroborated using multiple modalities.

Finally, a sequencing-based approach without tissue preservation is Div-Seq (see Fig. 4d). This technique combines scorable single-nucleus RNA-Seq (sNuc-Seq) with the EdU cell-labeling technique to profile single dividing cells (Habib et al., 2016). In the corresponding study, Div-Seq was used for obtaining and analysing the gene expression profiles of single nuclei from newborn neurons in the adult mouse hippocampus. This allowed them to track the lineage of these neurons by analysing their nuclei over multiple time points following division, in multiple mice. It constitutes the first dataset providing single-cell temporally-resolved expression data, analysing rare newborn neurons and identifying cell-types with highly similar transcriptomic profiles.

Two approaches that have not been applied to mice yet but can be of great value for future studies are sci-RNA-seq (single-cell combinatorial indexing RNA sequencing) (Cao et al., 2017) and SYNseq (Peikon et al., 2017), from which the former has achieved whole-brain coverage and the latter has only been tested in cultured neurons (Cao et al., 2017). sci-RNA-seq enables combinatorial analysis by sequencing tissues comprised of multiple cell-types in a single experiment (see Fig. 4c). It is a low cost, high throughput and fast approach that requires two days for constructing expression profiles of thousands of single-cells in a single experiment. This is achieved by assigning barcodes with unique molecular identifiers (UMI) to each analysed cell prior to the RNA sequencing process, thereby allowing for recovering their identity after sequencing. Further studies are needed, since it has only been tested in the L2 larval stage of C. elegans (Cao et al., 2017).

SYNseq has been used to simultaneously sequence and map the connectivity patterns of millions of cultured cells resulting in a single-neuronal connectivity matrix of unprecedented scale (Peikon et al., 2017) (see Fig. 4b). Since the neuronal labeling is performed in situ, the exact locations of the connected neurons can be directly estimated. However, it is still a work in progress with a number of technical obstacles that need to be overcome, such as the inability to recover a significant fraction of the labeled barcode pairs of connected neurons (Peikon et al., 2017). This method could be used for spatial cell density inference and connectivity augmentation (see the following section for these definition of the terms). In order for SYNseq to be applicable, an updated version needs to be successfully applied to mice and yield a sufficient amount of data.

4. Data fusion

We examine and discuss a number of computational strategies for integrating the aforementioned data modalities in order to improve the mouse mesoconnectome. The order of reference for the strategies follows the scaling of the data, hence progressing from augmenting cell-type-specific data to augmenting the entire mesoconnectome. Readers are suggested to consult Tables 5 and 6 for additional quantitative details on the techniques described in this section. Figs. 5–8 provide schematic examples of strategies related to cell-type classification, spatial cell density inference, connectivity corroboration and shared factorisation, respectively.

4.1. Multi-modal cell-type classification

The data reviewed in Section 3 shows that different types of neuronal populations exhibit differences in their projection properties, such as the spatial extent of the projections and preferences in targeting particular brain areas and hemisphere. Therefore, identifying fine-grained cell-types leads to highly specific information routing pathways in the brain, which is an important feature of a cell-type-specific mouse mesoconnectome.

A typical cell-type classification approach involves applying clustering methods to single modalities related to cellular data, with scrRNA-seq being the most frequently used modality (Poulin et al., 2016) (see box Clustering). However, modalities such as morphology, electrophysiology, transcriptomics and embedding circuits reveal information about complementary properties of the cells. (Markram et al., 2004)
highlighted a mapping between morphology-based and electrophysiology-based cell-types in the rat brain and suggested the presence of a template neocortical microcircuit. In recent years these approaches have accelerated because of improvements in omics technologies as well as optogenetic approaches (Fenno et al., 2011; Poulin et al., 2016). In Section 3.3, two studies were discussed where electrophysiology, morphology and transcriptomic data from mice was used to identify cell-types with a precision that could not be achieved using each modality independently (Gouwens et al., 2019, 2020) (see Fig. 5).

For that reason, a cell-type classification strategy should focus on multimodal clustering (see box Clustering) to provide a more precise definition of cell-types by finding shared patterns across these

| Data types | Speed/Cost | Utility |
|------------|------------|---------|
| Linked ICA (Gouwens et al., 2011) | Volumetric data sharing a common dimension | Requires variational inference with 10x speed due to component elimination | Shared Factorisation |
| MCM (Knox et al., 2018) | Volumetric bulk tract-tracing volumes | Requires Markov Chain Monte Carlo inference thus slow/works on small datasets | Connectivity Augmentation |
| Relational Learning Models (Ambrosen et al., 2013) | Relational data (DTI, tract-tracing, single-neuron projections, protein or gene interaction networks) | Requires translinear models whose connectivity is modulated by shared variation | Connectivity Augmentation |
| ConnectID (Klingler et al., 2018) | scRNA-seq and spatial cell data | Spatial cell density patterns in terms of connectivity and target voxels | Connectivity Augmentation |
| SEURAT (Satija et al., 2015) | scRNA-seq and ISH | Spatial cell density patterns in terms of connectivity and target voxels | Connectivity Augmentation |
| Ortiz et al. approach (Ortiz et al., 2020) | scRNA-seq and ISH | Studying connectivity patterns in terms of cell-density dynamics of source and target voxels | Connectivity Augmentation |
| Penalized Least Squares approaches | All data registered to the CCF v3.0 | Studying connectivity patterns in terms of cell-density dynamics of source and target voxels | Connectivity Augmentation |
| Point cell spatial inference (Ero et al., 2018) | ISH and Nissl data | Studying connectivity patterns in terms of cell-density dynamics of source and target voxels | Connectivity Augmentation |
| Constrained Louvain Algorithm (Reimann et al., 2019) | Anatomical tract-tracing and single-neuron projection reconstructions | Studying connectivity patterns in terms of cell-density dynamics of source and target voxels | Connectivity Augmentation |
| TVMB (Melozzi et al., 2017) | DTI or tract-tracing data | Studying connectivity patterns in terms of cell-density dynamics of source and target voxels | Connectivity Augmentation |
| Single-neuron Models (Nandi et al., 2020) | Single-neuron data, e.g., ephys experiments, neuronal morphologies, scRNA-seq | Studying connectivity patterns in terms of cell-density dynamics of source and target voxels | Connectivity Augmentation |
| Spatial point patterns (LaGrow et al., 2018) | ISH data | Studying connectivity patterns in terms of cell-density dynamics of source and target voxels | Connectivity Augmentation |
| Multimodal Stacked Autoencoder (Zhang et al., 2015) | Sensor data | Studying connectivity patterns in terms of cell-density dynamics of source and target voxels | Connectivity Augmentation |
| Sparse Autoencoder and Deep Belief Network (Chen et al., 2017) | Sensor data | Studying connectivity patterns in terms of cell-density dynamics of source and target voxels | Connectivity Augmentation |
| Deep Coupling Autoencoder (Ma et al., 2018) | Sensor data | Studying connectivity patterns in terms of cell-density dynamics of source and target voxels | Connectivity Augmentation |

Table 5 Feature descriptions of a number of computational approaches: tested types of data, speed, cost and utility with respect to neuroanatomical data.

Table 6 Feature descriptions of a number of computational approaches: relationship to MCIF, number of tested species and requirements for being useful to MCIF.

| Data types | Speed/Cost | Utility |
|------------|------------|---------|
| Linked ICA (Gouwens et al., 2011) | Finding sources of shared variation across all modalities of interest | Mice, Humans |
| MCM (Knox et al., 2018) | Integrating anatomical tracing data with laminar resolution into a unified voxel connectivity array | Mice |
| Relational Learning Models (Ambrosen et al., 2013) | Inferring missing information on connections between neurons or brain areas | Any |
| ConnectID (Klingler et al., 2018) | Identify transcriptional modules whose combined activities reflect multiplex projections to Mop, SS and Sub | Mice |
| SEURAT (Satija et al., 2015) | Studying connectivity patterns in terms of cell-density dynamics of source and target voxels | Zebrafish embryos |
| Ortiz et al. approach (Ortiz et al., 2020) | Studying connectivity patterns in terms of cell-density dynamics of source and target voxels | Mice |
| Penalized Least Squares approaches | Finding sources of shared variation across all modalities of interest | Mice |
| Point cell spatial inference (Ero et al., 2018) | Studying connectivity patterns in terms of cell-density dynamics of source and target voxels | Mice |
| Constrained Louvain Algorithm (Reimann et al., 2019) | Studying connectivity patterns in terms of cell-density dynamics of source and target voxels | Mice |
| TVMB (Melozzi et al., 2017) | Enables testing cell-type-specific structural connectivity patterns in-silico by generating a corresponding functional activity | Mice |
| Single-neuron Models (Nandi et al., 2020) | Corroborating cell-type data and testing causal relationships | Mice |
| Spatial point patterns (LaGrow et al., 2018) | Studying connectivity patterns in terms of cell-density dynamics of source and target voxels | Mice |
| Multimodal Stacked Autoencoder | Finding sources of shared variation | Sufficient gene samples per brain slice |

(continued on next page)
Table 6 (continued)

| Method | Relation to the end-goal | Species | Requirement |
|--------|--------------------------|---------|-------------|
| Autoencoder (Zhang et al., 2015) | across all modalities of interest | Not tested in neuroanatomical data yet | Sharing a common dimension |
| Sparse Autoencoder and Deep belief network (Chen et al., 2017) | Finding sources of shared variation across all modalities of interest | Not tested in neuroanatomical data yet | Sharing a common dimension |
| Deep Coupling Autoencoder (Ma et al., 2018) | Finding sources of shared variation across all modalities of interest | Not tested in neuroanatomical data yet | Non |

modalities. In the following paragraphs we highlight a number of additional studies that follow that example.

The ConnectiTIP approach was developed in Klingler et al. (2018) and combines MAPseq with scRNA-seq data. The utility of such method is to identify transcriptional modules whose co-expression patterns correlate with multiple single-neuron projections to other brain regions. For example, in Klingler et al. (2018) they identified modules related to intra-cortical connections to the primary motor cortex (MOp), secondary somatosensory cortex (SS) and subcortical areas (Sub).

In Nandi et al. (2020), they integrated electrophysiology, morphology and transcriptomics in the same framework to build biologically inspired neuron models. They generated $\sim 10^8$ single-cell computational models with active and passive conductances, by varying the passive conductance properties across the models. The models were formulated based on experimental data that comprised hundreds of patch clamp recordings and the corresponding morphological reconstructions (Gouwens et al., 2019). Furthermore, they used scRNA-seq to validate their models by correlating differences between inhibitory and excitatory cells across modalities. Specifically, they found that differences in the conductances between the model-based inhibitory and excitatory cells were correlated with the transcriptional and electrophysiological differences between the experimentally characterized cells. Hence, this framework was able to reconcile the different single-cell-related modalities, thereby strengthening the cell-type definitions and testing causal relationships (Nandi et al., 2020).

An intriguing follow-up question to multi-modal cell-type classification is how behavioral states are encoded by fine grained cell-types that have been defined in this multi-modal fashion. In a recent study (Xu et al., 2020), they evaluated the specificity of transcriptomically-defined cell-types in encoding behavioral states, by combining calcium imaging during multiple behavioral tasks with ex vivo multiplexed RNA fluorescent in situ hybridization. They found a highly accurate decoding of behavioral states from different combinations of cell-type-specific activation patterns in the hypothalamic paraventricular nucleus (PVH), an area crucial for behavior. This finding is a prime example of advance towards an ambitious aim called Rosetta Brains (Marblestone et al., 2014): bridging the gap between molecular and systems Neuroscience by integrating data specific to cellular gene expression, development, connectivity and activity in relation to behavior, and understanding this relationship in health and brain disease.

While we have demonstrated a number of studies that have combined gene expression, morphology and electrophysiology within a single experiment, integrating behavioral encoding and cellular development are more ambitious goals and will require further development of novel techniques and more studies. In Section 6, we will discuss the feasibility of this aim and the extent to which it relates to MCIF.

Clustering

Clustering refers to a family of unsupervised machine learning methods (Bishop, 2006), which partitions the datapoints of a dataset into categories or clusters. The resulting clusters describe the internal organization of the data based on a proximity measure such as the euclidean distance or the Pearson correlation coefficient (Tan et al., 2005). A typical use-case of clustering is to provide a summary of the data through visualization of the clusters. Since visualization can only be achieved in the 2D or 3D space, clustering is often accompanied by dimensionality reduction methods, such as the commonly used t-SNE (van der Maaten and Hinton, 2008). Besides visualization, clustering results can be compared with additional data to assess whether they reflect categories of significance. An example in transcriptomics is the application of clustering to scRNA-seq data, with the purpose of identifying the number of underlying cell-types (Poulin et al., 2016). A typical clustering objective function to be minimized is the intra-cluster variance (Tan et al., 2005):

$$J = \sum_{j=1}^{K} \sum_{i=1}^{n} K(x_i, c_j) \delta_j(x_i)$$

(1)

here $c_j$ is the centroid or representative of cluster $j$, $x_i$ is the $i$th datapoint, $K(x_i, c_j)$ is a function quantifying the proximity between the datapoints and the cluster centroids or representatives and $\delta_j$ is an indicator function for the $j$th cluster where $\delta_j(x_i) = 1$ if $x_i \in C_j$ otherwise it is 0.

Other clustering techniques, such as Spectral Clustering (Shi and Malik, 2000), has been developed to overcome limitations of standard clustering algorithms. This approach is particularly useful for datasets with complex structures and overlapping clusters.

![Fig. 5. Cell-type classification by integrating different modalities related to cell structure and function, such as transcriptomics, electrophysiology and morphology, provides a more precise definition of cell-types compared to unimodal approaches (Klingler et al., 2018; Gouwens et al., 2019; Nandi et al., 2020). Panels a and b correspond to excitatory and inhibitory cell-types, respectively. The cell-types have been defined using both dendritic morphologies and electrophysiological responses as data features, which is illustrated by exemplar cells in each column. The numbers below each column correspond to a subclass defined using transcriptomic data, which was used to further group the cell-types. Panels a and b are reprinted by permission from Springer Nature: Nature Neuroscience (Gouwens et al., 2019), copyright 2019.](image-url)
perform eigenvalue decomposition on the proximity matrix and cluster the data based on the eigenvectors with the k largest eigenvalues. More recent clustering techniques, such as DBSCAN or dip-means are density-based and do not require the clusters to be spherical (Ester et al., 1996; Kalogeratos and Likas, 2012). Moreover, not all clustering techniques require the number of clusters to be specified in advance. Hierarchical clustering is a family of clustering techniques that create a dendrogram of clusters, with the root node corresponding to a unified cluster and the leaf nodes corresponding to the individual datapoints, while nodes are connected based on a proximity measure such as the Sum of Squared Error (Tan et al., 2005):

\[
\text{SSE} = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - c_i||^2
\]

where \(C_i\) is the \(i\)th cluster or node in a given hierarchy of the dendrogram, \(x\) is the set of datapoints that are members of \(C_i\), and \(c_i\) is the centroid of the cluster.

An interesting clustering variation are the bi-clustering techniques. While traditional clustering focuses on partitioning the samples of a dataset, bi-clustering partitions the data at the level of samples and features simultaneously (Prelic et al., 2006; Busygin et al., 2008; Mishne et al., 2019). This can be particularly useful in transcriptomics, where different sub-groups of genes are assumed to co-express in specific groups of cells, leading to diagonal blocks in the ordered cell-gene matrix (Poulin et al., 2016). Moreover, a combined feature matrix of multiple modalities can be co-clustered to provide fine-grained descriptions of cell-types, as shown in Gouwens et al. (2019) with identified morphology-electrophysiology-based cell-types.

4.2. Spatial cell density inference

Spatial cell density is an important constraint for the connectome. In order to make the leap from fine-grained cell-type classification to cell-type-specific projection patterns, the densities or counts of the source and target cell-types should be spatially registered (Chevete et al., 2018; Busygin et al., 2008; Mishne et al., 2019). This can be particularly useful in transcriptomics, where different sub-groups of genes are assumed to co-express in specific groups of cells, leading to diagonal blocks in the ordered cell-gene matrix (Poulin et al., 2016). Moreover, a combined feature matrix of multiple modalities can be co-clustered to provide fine-grained descriptions of cell-types, as shown in Gouwens et al. (2019) with identified morphology-electrophysiology-based cell-types.
anatomical space is not known to sufficient detail and spatial registration is required. These methods aim to infer the counts, densities and positions of cells in brain space.

High throughput volumetric methods that yield RNA expression patterns can be combined with scRNA-seq to estimate the densities or locations of cells in space. In Grange et al. (2014) they used a penalized least-squares method to integrate the ISH and scRNA-seq data, resulting in densities of multiple cell-types over 200 μm voxels that were validated by experimental data (see Fig. 6 b). ( Li et al., 2017 ) produced similar results with the ISH data using the DLSC technique ( Table 2 ), a dimensionality reduction approach based on sparse representations of basis vectors ( Mairal et al., 2010 ). Both approaches demonstrate that the spatial expression of gene modules can be used to approximate densities of various cell-types.

Another approach integrating scRNA-seq and ISH data is the SEURAT technique (see box “SEURAT approach”), which was used to infer the spatial locations of cells from zebrafish embryos, a number of which were rare cell-types ( Satija et al., 2015 ) (see Fig. 6 c). In Ortiz et al. (2020) , they combined SEURAT with supervised classification for mapping cells in the adult mouse brain, due to its higher complexity compared to zebrafish embryos. Specifically, they used SEURAT to map a subset of cells, which were then used as labels to train Support Vector Machine (SVM) and Artificial Neural Network models for mapping the remaining cells ( Cortes and Vapnik, 1995; Bishop, 2006 ). This resulted in a whole-brain molecular atlas consisting of hundreds of gene clusters with distinct spatial profiles.
In Erö et al. (2018), a different approach was followed (see Fig. 6a). Specifically, they used a Monte-Carlo algorithm to generate the positions of $10^6$ cells in the mouse brain given spatial constraints provided by the high resolution Nissl volumes of the mouse brain (Fig. 6). Moreover, they used the ISH data of AMBA to determine the cell-type of each cell based on the gene marker with the highest expression in the same spatial location. This approach can be useful for inferring cell location and counts given cell density data for the purpose of model building or statistical comparison with transgenic animal models for disease. The approach is not limited to Nissl volumes, but it could also be applied to densities estimated using other techniques such as EDC-CLARITY (Sylwestrak et al., 2016) and qBrain (Kim et al., 2017). Thus, it constitutes a complementary strategy to the cell-density inference approaches mentioned above (Grange et al., 2014; Li et al., 2017).

The density of specific pre-synaptic and post-synaptic puncta has been shown to be a significant marker of meso-scale projection patterns (Zhu et al., 2018). Thus, a potential data fusion scheme could augment cell-type-specific connectivity representations by overlapping synaptic puncta positions or densities together with the corresponding cell-type counts or densities, morphologies and projection densities in brain space. In the following section we will discuss ways to augment the already existing mesoconnectome models using morphology and projection density data. In Section 4.4, we will highlight and discuss a number of fusion approaches.

### Spatial cell density inference approaches

**SEURAT** is a model that integrates ISH and scRNA-seq data, with the aim of mapping cells in brain space using their genetic profile and imputing missing gene values in ISH data (Sattija et al., 2015). When applied to the zebrafish embryo, it yielded an ROC score of 0.96 for the mapping quality. The steps of SEURAT can be summarized as follows.

1. cell densities are estimated from Nissl volumes.
2. a random voxel V is selected which has a corresponding intensity $dV$
3. a random probability $p$ is generated
4. a cell is generated in a random position inside V if $p < dV$
5. step 2 is repeated if a total number of cells $N$ is not reached, otherwise the loop halts.
6. gene markers are used to label the cells as inhibitory or excitatory neurons or glial cells.

Note that for the total number of cells $N$, three different numbers were used for the isocortex, cerebellum and the rest of the brain based on literature (Herculano-Houzel et al., 2011).

### 4.3. Connectivity augmentation

The term connectivity augmentation is used to characterize strategies whose goal is to directly improve the representation of the connectome given connectivity data as input. An appropriate goal would be that an augmented connectome is cell-type-specific and volumetric with cellular-level resolution of at least 10 μm. Given the available data, an augmentation strategy should be able to integrate different cre-lines and axonal morphologies to allow for laminar and cell-type specificity. Additionally, it should incorporate DTI data in order to account for subject-specific connectivity estimates.

In computational terms, the output of such methods would be a weighted directed graph or a 2D matrix whose nodes or rows and columns correspond to connected source and target brain areas at a level of delineation as close to cell-type-specific as possible. Appropriate methods discussed below are non-parametric kernel regression, community detection and Bayesian relational learning methods.

Non-parametric kernel regression models have been used for integrating the injection and projection densities of tract-tracing experiments into a unified connectivity matrix (Oh et al., 2014; Ypma and Bullmore, 2016; Gamanut et al., 2018; Knox et al., 2018). A number of these models have been incorporated to the MCM toolbox provided by the Allen Institute of Brain Science (https://alleninstitute.org/). In Knox et al. (2018), they provided the first non-parametric regression model based on the 100 μm volumetric version of the wild-type tract-tracing data from the AMBWA (see Fig. 7a). The model was based on the Nadaraya-Watson estimator (Nadaraya, 1964) and provided a voxel-by-voxel connectivity array. This model can be used to corroborate structural connectivity by fusing multiple cre-line tract-tracing experiments into a unified, volumetric, layer-specific and cell-class-specific connectivity array. An example dataset for application is the cre-line tract-tracing experiments from the AMBWA (Harris et al., 2019) (see Section 3.1).

### Non-parametric kernel regression

The Nadaraya-Watson model (Nadaraya, 1964) is a non-parametric kernel regression model that was used in Knox et al. (2018) to estimate a voxel-based connectivity matrix from a tract-tracing dataset (see Section 4.3). It was adapted to integrate projections from multiple

---

**Fig. 8.** A multi-modal data fusion example using a shared factorisation strategy to reduce the high dimensionality of multiple spatially overlaid datasets and to generate hypotheses for animal models of disease (see Section 5.2). The Linked ICA method is used in order to find linked independent components explaining variance that is shared across two volumetric modalities (Grovess et al., 2011). The modalities were obtained by ISH (Lein et al., 2007) and tract-tracing experiments in wild type mice (Oh et al., 2014), respectively. This example displays a spatial map which highlights brain areas with high variance in a given independent component. The blue-to-lightblue color represents voxels with large negative values (below 1st-percentile) and the red-to-yellow color represents voxels with large positive values (above 99th-percentile), see colorbar. A number of highlighted subcortical areas give the impression of being located outside of brain space. This is explained by the low density of the Nissl volume in those areas, which serves as the anatomical template. This template has been plotted overlaid with the spatial map, based on CCF v3.0 (Wang et al., 2020). The figure is reprinted from Timonidis et al. (2021).
discrete injections into a smooth projection pattern and can incorporate additional data when it becomes available. The procedure is as follows:

\[
W_i = \sum_{c_i \in c} K(||y_i - c_i||)\tilde{Y}_i
\]

(3)

here \(S_i\) is the set of tract-tracing experiments, \(\tilde{Y}\) contains the projection densities, \(c\) represents the center of masses for the different injections, \(v\) represents voxels part of or close to the injection sites and indices \(i,j\) and \(e,f\) correspond to target voxel, source voxel and injection respectively (Knox et al., 2018). The Gaussian radial basis function kernel was used for estimating the distance between the injection mass centres and the target voxels \(v\) is an inverse scaling parameter:

\[
K_v(d) = \exp \left( -\frac{d^2}{2\sigma^2} \right)
\]

(4)

The following step is to integrate morphology data. An exemplary approach can be found in Reimann et al. (2019), where a constrained version of the Lookavin community detection algorithm (Blondel et al., 2008) was applied to tract-tracing data from the AMBCA. Specifically, they applied the algorithm to convert a meso-scale connectivity matrix into an undirected proximity matrix. Afterwards, they used a modification to construct a directed tree graph and account for directed connections. Then, they estimated the innervation probabilities of pairs of brain areas by single-neuron axonal projections. Finally, they updated the connectivity matrix using the estimated innervation probabilities (see Fig. 7b). The single-neuron data were part of the axonal reconstructions of the MouseLight project (Economou et al., 2016; Wimubst et al., 2019). As a result, this work constitutes the first important effort of combining morphology data with tract-tracing data such that the innervation patterns of single-neurons influence the long-range projection densities.

However, the number of available single-neuron morphologies is still not enough to cover the whole mouse brain. For that reason, it will be necessary to impute missing projections. A highly successful family of imputation methods applied in connectomics are the non-parametric Bayesian relational learning models (Ambrosen et al., 2013; Hinne et al., 2014, 2017; Betzel and Bassett, 2017) (see Fig. 7c). Variations of the Infinite relational model (IRM) have been applied to network problems, such as the IFRM and the IHRM (Miller et al., 2009; Xu et al., 2012) (see Table 2 for the abbreviations). Examples of accurate connectivity predictions include the C. elegans connectome and the mouse retina microcircuit (Jonas and Kording, 2014). While these techniques have been used in multiple connectome-related studies in the past, the MCMC sampling approaches (Table 2) are slow in convergence for large volumetric datasets (Neal, 2011), such as the ones mentioned in Section 3. To work on such large-scale datasets, it is more appropriate to develop a relational learning approach that is based on a variational inference (VI) strategy, such as AEVB (Attias, 2000; Kingma and Welling, 2013) (see box “Variational Inference”). This is because VI approaches use stochastic and distributed optimization techniques that scale better compared to MCMC for large datasets (Blei et al., 2017).

**Non-parametric relational learning models**

These methods assume that the data is being generated by a latent space and that the connection probability of two nodes is related to the proximity of these nodes in the latent space:

\[
\tilde{z} \sim \text{CRP}(\lambda)
\]

(5)

\[
p(\tilde{z}) = \frac{\Gamma(\lambda) \lambda^k}{\Gamma(\lambda+k)} \prod_{i=1}^k \Gamma(n_i)
\]

(6)

\[
\phi_{ij} \sim \text{Beta}(a,b)
\]

(7)

\[
\lambda_{ij} \sim \text{Bernoulli}(\phi_{ij} \tilde{z})
\]

(8)

where Eq. (5) corresponds to the partitioning prior distribution estimated using the Chinese Restaurant Process (CRP) (Aldous, 1985), \(\lambda\) denotes the distribution scale and \(\tilde{z}\) denotes the realization of a partition drawn from the CRP. CRP assumes a latent space of infinite size and partitions the nodes into a finite, non-empty subset of the latent space. This partitioning can be estimated as a limit of the CRP distribution as the number of latent variables goes to infinity, as shown in Eq. (6). There, \(n\) denotes the number of nodes belonging to the \(k\)th cluster, i.e., with \(\tilde{z}_k = k\). Moreover, Eq. (7) corresponds to the linkage probability \(\phi_{ij}\) between a pair of latent variables \(k\) and \(l\), which is realized by the Beta distribution. Finally, Eq. (8) corresponds to the distribution of a connection between two network nodes, given the affinity between the latent variables that they have been partitioned as Eq. (7), which is realized by the Bernoulli distribution.

The various IRM-based models differ in their prior distribution assumptions, for which in the aforementioned case the CRP, Gamma, Beta and Bernoulli distributions have been chosen (see Eqs. (5)–(8)). For instance, the IFRM includes an additional Bernoulli distribution for assigning multiple features to the latent variables (Miller et al., 2009), and the IHRM includes a Dirichlet distribution to assign multiple links between the latent variables (Xu et al., 2012). In all cases, the posterior distribution of the latent space is derived using MCMC sampling approaches (Neal, 2011).

Tract-tracing and DTI data provide different types of estimates for the mesoconnectome, such as directionality in the case of tract-tracing, as well as subject-specificity and spatial smoothness in the case of DTI. Moreover, DTI data have reached a 43 µm resolution in rodents (Calabrese et al., 2015), which makes it comparable to the voxel-based connectivity data in AMBCA. Hence, both modalities could complement each other in an integrated approach. An additional advantage is that DTI can act as an intermediate registration target between AMBCA and other imaging modalities such as resting state fMRI, which can provide estimates of functional connectivity to complement a structural analysis (Grandjean et al., 2017).

In Chen et al. (2015a) a framework was provided for spatially anchoring DTI data to the AMBCA. This was achieved by annotating and parcellating the DTI data using the CCF v1.0 under various scales of resolution. Subsequently, AMBCA was used as a ground truth to optimize the parcellation derived using two main DTI parameters, fractional anisotropy (FA) threshold and angular threshold. This was achieved by estimating the maximum Area Under the ROC Curve (AUC) value for the different parcellations. The highest performance was reached at a parcellation of 96 regions, with an AUC of 72% for ipsilateral and 68% for contralateral connections. This study serves as an example on how high resolution DTI data can be optimized for integration with a connectome model based on AMBCA (see Fig. 7d).

**Variational Inference**

The idea behind variational inference is that Bayesian inference can be approximated by estimating a tractable lower bound. A number of complicated models can be formulated and solved using this approach, such as Linked ICA (Groves et al., 2011). Given a set of observations \(X\) and a set of latent variables \(Z\), the posterior distribution of \(Z\) given \(X\) would be estimated as:

\[
p(Z|X) = \frac{p(X|Z)p(Z)}{p(X)}
\]

(9)

from this equation, \(p(X)\) can be estimated using the integral:

\[
p(X) = \int dp(Z|X)\]

(10)

this integral can be approximated by taking the variational lower bound referred to as ELBO (Attias, 2000; Kingma and Welling, 2013). ELBO is the lower bound of the log probability of \(X\):

\[
\log p(X) \geq E_{\hat{q}(Z)}[\log p(X|Z)] + H[\hat{q}]
\]

(11)
where \( q \) is a distribution that approximates the posterior distribution \( p(Z/X) \) and can be estimated using a parametric function. The proper choice of \( q \) is central for minimizing the ELBO, which depends on minimizing the Shannon entropy of \( q, H(Z): \)

\[
H(Z) = -E_q[\log(q(Z))]
\]

The advantage of this approximation is that it is tractable under stochastic-gradient based methods such as Stochastic Gradient Descent (Attias, 2000; Kingma and Welling, 2013).

4.4. Shared factorisation

Previous strategies focused on improving the representations of specific data modalities. These modalities can be directly integrated via spatial registration and overlaid in a brain atlas. However, the massive amounts of available data will require a strategy to reduce the computational resources required for their integration. Therefore, shared factorisation can be used to express different data modalities in terms of common factors, often representing spatial profiles. These techniques thereby can identify biologically relevant sources of shared variation and minimize redundant information.

A number of factorisation approaches is based on minimizing the least-squares objective function under L1 or L2 norm sparsity constraints on the coefficients and factors, such as Lasso Regression (Grange et al., 2014; Ji et al., 2014; Fakhry et al., 2015), Sparse Reweighted-Loss Rank-Sparse Matrix (Kovak et al., 2019), 2D CCA (Qadar et al., 2019) and Sparse CCA (Boute and Liu, 2010) (Table 2). These approaches work by estimating a sparse set of coefficients that link two modalities, under a formulation as shown in Eq. 13, where \( Y \) representing modality 1 and \( H \) representing modality 2, or in Eq. 14 with \( k = 2 \) modalities. A classical use case is linking gene expression with structural connectivity (Ji et al., 2014; Fakhry et al., 2015; Timonidis et al., 2019; Huang et al., 2020) or with electrophysiology (Kovak et al., 2019). A disadvantage of this strategy is that it only works with two modalities, in contrast to multi-modal fusion approaches such as the ones that will be discussed below.

Multi-modal factorisation approaches have also been proposed for finding sources of shared variation for more than two modalities. Examples techniques are Group-PCA (Smith et al., 2014), Concat-ICA (Calhoun et al., 2006), Tensor-ICA (Beckmann and Smith, 2005) and Matrix Tri-Factorisation (Zitnik and Zupan, 2015) (Table 2). However, these approaches assume that all modalities have the same noise levels across components. As a consequence, components will be filtered out in cases where some modalities have a higher ratio of explained variance than others.

Linked ICA assumes for each modality, a relative contribution to the cost function (see Eq. 15), where concomitantly individual degrees-of-freedom are obtained using a Variational Bayes approach (Groves et al., 2011). Linked ICA has been previously applied to functional, structural and diffusion-MRI data, whose subject dimension was the multi-modal shared element (Kincses et al., 2013; Itahashi et al., 2015; Wolters et al., 2017; Llera et al., 2019; Maglanoc et al., 2020). However, the approach can be modified to use another dimension as the shared one, such as the voxel-dimension (see Fig. 8).

An additional advantage of Linked ICA is that it can be \( \approx 10 \) times faster than a conventional variational inference method, because it has an elimination strategy for components that do not contribute to the total data variance (Groves et al., 2011). Linked ICA also creates a generative model that can reconstruct the original data or create new test data, since it models the distribution of the independent components that link the data modalities (Groves et al., 2011). Given that all parts of the generative model are being explicitly modelled, in the next section we discuss for a number of use cases where this model can be helpful for generating and testing hypotheses in neuroscience studies.

Shared Factorisation Formulation

Classical matrix factorisation aims to represent a data matrix \( Y \) as a linear combination of a feature coefficient matrix \( X \), a latent variable matrix \( H \), and a residual matrix \( E \):

\[
Y = XH + E.
\]

where \( H \) has a number of rows equal or less than the original matrix and maximizes the explained variance of \( Y \), under a set of constraints imposed by the model priors. Similarly, shared factorisation aims to maximize the explained variance of multiple data modalities using the same formula:

\[
Y^k = X^kW^kH + E^k.
\]

where \( k \) corresponds to a data modality, \( X^k \) contains the feature coefficients for modality \( k \), \( H \) contains the shared latent variables and \( E^k \) contains the residuals of modality \( k \). In case where the contribution of each modality can be set independently instead of equal across all components, then the factorisation can be reformulated as:

\[
Y^k = X^kW^kH + E^k.
\]

5. Use cases

In this section we highlight a number of use cases in which experimental and theoretical neuroscience could benefit from the suggested fusion strategies.

5.1. Activity simulation

Big initiatives such as the Virtual Brain project and the Blue Brain project have developed large-scale computational models of the rodent or human brains in which the connectome plays a central role (Markram, 2006; Markram et al., 2011; Sanz Leon et al., 2013). The connectome’s graph-theoretic representation serves as a bridge between experimental data and computational models. A prime example is the emergence of functional connectivity dynamics (FCD) from the structural connectivity, under various models used for the computational nodes (Hansen et al., 2015; Choi and Mihalas, 2019).

A framework generating function from structure in the mouse brain is the Virtual Mouse Brain (TVMB) (Sanz-Leon et al., 2013; Ritter et al., 2013; Woodman et al., 2014). TVMB is an open source software that simulates large scale brain network dynamics, related to healthy and pathological conditions, by incorporating computational models and neuroimaging data from mice (Melozzi et al., 2017).

An interesting use case of TVMB can be found in Melozzi et al. (2019). In this study, they generated resting-state FCD in mice by giving as input the structural connectivity provided by AMBCA and by diffusion-MRI data that was obtained from 19 mice (see Fig. 9). The
mean field activity of brain regions was derived from the connectivity strength of the predicted connectomes and from additional assumptions induced by the use of the reduced Wong-Wang model for the nodes representing brain regions (Wong and Wang, 2006).

By incorporating the augmented connectome suggested by the fusion strategies and using cell-densities for constraining the mean field approximations, one can use TVMB to test cell-type-specific structural connectivity patterns in-silico by generating the corresponding FCD and

---

**Fig. 9.** The Virtual Mouse Brain (TVMB) is a computational framework for simulating resting state activity from structural connectivity data derived from the mouse brain (Melozzi et al., 2017). Specifically, individual mice are scanned using fMRI and diffusion-MRI in order to obtain structural connectivity (SC) and functional connectivity (FC) data (Melozzi et al., 2019). Afterwards, the TVMB simulates BOLD activity from the SC data and estimates an FC matrix by taking a voxel-wise correlation of the activity across all voxels. Likewise, an experimental FC matrix is estimated from the experimental BOLD activity of the FC data. Finally, the voxel-wise similarity of both FC matrices is assessed. This procedure is repeated using the wild-type tract-tracing data from AMBCA (Oh et al., 2014). Both similarity matrices are compared to determine which experimental data better predicts functional connectivity dynamics. The figure is reprinted from Melozzi et al. (2019).
evaluate its variance across subjects, different cell-type-specific patterns and different manipulations of structural connectivity.

5.2. Generation of hypotheses for animal models of disease

Identifying multi-modal components of shared variance across various parts of the mouse brain could lead to the generation of new hypotheses on mechanisms of disease. These hypotheses can be validated using other data or tested experimentally. Any shared factorisation approach, such as Linked ICA, will link the related data modalities in brain space. Hence, each component will be comprised of spatial patterns, coefficients of the different modalities and contribution of the modalities to the component (Groves et al., 2011). Therefore, a user can identify a component of interest by visualizing its spatial map and identifying high variance in an area or sub-area of interest. This subsequently will lead to finding a subset of interesting coefficients, such as genes in the case of gene expression data or tracts in DTI data. Afterwards, a post-hoc analysis through correlation with other sources of information might lead to mechanistic insights.

An example would be to identify a gene subset associated with a particular projection affected in the disease and to apply ontology enrichment analysis for assessing its functional relevance (Rice, 2007; Rivals et al., 2007). If a post-hoc analysis results in a significant correlation, which for example might be a significant enrichment in GABergic pathways, the appropriate cell populations could be targeted by optogenetics to mitigate the disease symptoms, or induce them in control subjects.

5.3. Next steps

As a follow-up step, we propose the use of Neuroinformatics-based platforms such as EBRAINS (“https://ebrains.eu/”) for multiple pre-processing, analysis and modeling steps related to the data produced by the fusion strategies or the connectivity simulations. The advantage of using EBRAINS includes data curation and sending feedback to the providers of the original data sources. Examples of feedback include a description of required pre-processing steps, the compatibility for usage in MCIF and 3D visualization of the integration output. Moreover, enabling public access to the data and the workflow of MCIF could allow users to generate and test their hypotheses in silico (see Fig. 1).

6. Discussion

During the past decade, spurred by large-scale funding programmes such as the Brain Initiative (Alivisatos et al., 2015), rapid progress has been made in new neuroanatomical methods as well as their application as part of high-throughput approaches, which were combined with efficient, well-curated and open-source processing pipelines. In this study we reviewed these novel neuroanatomical techniques and pipelines, with a focus on applying data fusion techniques to support a multi-scale integrative framework necessary for obtaining a cell-type-specific resolution of the mouse mesoconnectome. Current experimental techniques provide comprehensive data for characterizing the cell types and projections, with a focus on applying data fusion techniques to support a multi-scale integrative framework necessary for obtaining a cell-type-specific resolution of the mouse mesoconnectome. Current experimental techniques provide comprehensive data for characterizing mouse brain cell types based on their transcriptomic, morphological, electrophysiology and projection properties not only with sufficient quality to support large-scale models (Alivisatos et al., 2013; Billeh et al., 2020) but also to serve as input for the data fusion techniques as part of the MCIF framework. These data fusion techniques perform classification of cell types, spatial inference of cell density patterns, augment the representation of structural connectivity and simplify the representation using multi-modal factorisation. Taken together, these advances can drive a large number of use-cases, of which we highlighted two that are exemplary for being of great value for either experimental or theoretical neuroscientists, specifically simulations of functional connectivity dynamics and generators of hypotheses for animal models of disease.

The approaches not only benefit from large-based infrastructures such as EBRAINS to store, process and analyse data related to the prior approaches, but these may also be required given the high demands on computational resources necessary for the data integration pipelines and storage. In the backdrop of these infrastructures, we believe that a decisive step towards a cell-type-resolved mesoconnectome can be made in the coming five years provided there is sufficient funding and coordination between scientific partners, similar to recent large projects or initiatives, such as the Blue Brain Project (Markram, 2006), the Human Brain Project (Markram et al., 2011), the Virtual Brain Project (Sanz Leon et al., 2013) or the BRAIN initiative (Jorgenson et al., 2015).

An investment in such a project would nevertheless imply targeted development of only a subset of techniques among the ones reviewed here. Based on not only the intrinsic strength of a technique, which also considers cost in time as well as money, but also the ability to combine with other techniques and integrate in a pipeline, we argue for the following selection. First, a robust barcoding technique along the lines of MAPseq, BARseq and BRISeq (see Section 3.5) (Kebischull et al., 2016; Chen et al., 2019; Huang et al., 2020), because it yields the combinatorial multiplexed labeling of cell-type-specific axonal projection patterns and gene expression that cannot be achieved by classical anatomical tracing, scRNA-seq or ISH approaches. Second, a retrograde-anterograde viral labeling combination that filters out unwanted neuronal labeling by using the pseudotyped rabies virus (Wickersham et al., 2007; Zingg et al., 2017) would be useful for labeling cell-type-specific projections across the ones tagged by the barcode approach. Third, a synaptic-tagging technique such as SYNMAP (Zhu et al., 2018) would quantify the pre- and post-synaptic counts and densities across the mapped projections. Fourth, high-resolution microscopy, such as LSFM, combined with tissue clearing techniques, followed by an image processing pipeline (see Section 3.4) (Dodt et al., 2007; Tonom et al., 2014; Pan et al., 2016; Tainaka et al., 2018; Gao et al., 2019; Zhu et al., 2020; Kist et al., 2020), because it ensures fast scanning of brain volumes containing the tagged axonal projections and synapses at micron resolution. Lastly, the final steps of MCIF, namely connectivity augmentation (Section 4.3) and multi-modal factorisation (Section 4.4), would be necessary for achieving data integration and providing spatially smooth connectivity patterns for visualization purposes.

The selection strategy in the preceding paragraphs is in line with the program referred to as Rosetta Brains, which sets an even more ambitious goal of integrating cell-type-specific transcriptomics, projections, activity in different behavioral states and circuit development. To achieve this, (Marblestone et al., 2014) proposed a combination of fluorescent in situ sequencing, single-nucleus sequencing for lineage tracking (Zhizhov et al., 2016; Shapiro et al., 2013), cellular recordings and barcode tagging of pre-synaptic and post-synaptic pairs (see Sections 3.2–3.5 for related techniques), in the context of hypothesis-driven experiments in a behavioral setting. The proposal was, and still is, theoretical, as the extent to which it can be realised depends on the further development of techniques reviewed in the preceding sections (Section 3), similar to the limitations of MCIF. For example, further development of in situ sequencing would provide direct spatial localization of single-cell data and make spatial inference techniques in the data fusion pipeline redundant. Likewise, improvement of barcode-sequencing would make it feasible to obtain projection, activity and gene expression data in a single experiment and minimize the need to integrate data from different studies to define multi-modal cell-types. Finally, producing whole-brain volumes using high-resolution microscopy would reduce the reliance on data with spatial discrepancies that are induced by aligning tissue sections across the same or different subjects and registering them to a common atlas (Fürth et al., 2018).

This review focused on the mouse given the massive amount of available data. An open question is to what extent the insights of mouse studies can be translated to the human brain. Primate brains are known to differ in the elaboration of cortical circuits (Kang et al., 2011; Lui
et al., 2011), as well as basic features such as structure of the neuronal membranes (Bozek et al., 2015). Reviewing translational research was beyond the scope of this review, but a number of techniques can be directly applied (post-mortem) to human brains, such as polarized light microscopy. Moreover, the rapid developments in brain organoids (DiLullo and Kriegstein, 2017) offer some hope that circuit structure can be predicted with significant accuracy in human brains, such as polarized light microscopy.

References

Ahrlund-Richter, S., Xuan, Y., van Lunteren, J.A., Kim, H., Ortiz, C., Pollak Dorocic, I., Chon, U., Vanselow, D.J., Cheng, K.C., Kim, Y., 2019. Enhanced and unified anatomical and functional data fusion. IEEE J. Biomed. Health Inform. 23 (1), 47–62. https://doi.org/10.1109/JBHI.2018.2867192.

Alvati, A.P., Andrews, A.M., Boyden, E.S., Chun, M., Church, G.M., Deisseroth, K., Donoghue, J.P., Fraser, S.E., Lippincott-Schwartz, J., Looger, L.L., Masmanidis, S., McEuen, P.L., Nurmikko, A.V., Park, H., Peterka, D.S., Reid, C., Roukes, M.L., Scherer, A., Schütz, M., Sejnowski, T.J., Shepard, K.L., Tsao, D., Turjáni, G., Weiss, P.S., Xu, C., Yuste, R., Zhen, X., 2013. Nanotools for neuroscience and brain activity mapping. ACS Nano 7 (3), 1850–1866. https://doi.org/10.1021/nn3041284.

Alvati, A.P., Chun, M., Church, G.M., GreenSPAN, R.J., Roukes, M.L., Yuste, R., 2015. A national network of neuroscience technology centers for the brain initiative. Neuron 88 (3), 435–448. https://doi.org/10.1016/j.neuron.2015.05.015.

Ahn, J.H., Lee, M., 2019. A whole-brain atlas of monosynaptic input targeting neurons. J. Neurosci. Methods 337, 298–311. https://doi.org/10.1016/j.jneumeth.2018.07.011.

Alessi, D., Koibuchi, N., Shimizu, T., Yonekura, S., Motomura, H., Jonassen, D., 2015. The power of machine learning to extract and interpret brain data. Bioinf. Comput. Biol. Solutions 2, 1–27. https://doi.org/10.1186/s40360-015-0025-3.

Alzamora, M., Vazquez-Medina, R., Rueda, R.M., 2019. Modeling cell states using single-cell RNA sequencing data. Bioinformatics 35 (14), 2135–2140. https://doi.org/10.1093/bioinformatics/btz054.

Amann, R., Fuchs, B.M., 2008. Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques. Nat. Rev. Microbiol. 6, 329–348. https://doi.org/10.1038/nrmicro1808.

Anavy, A., Nestler, E.J., 2015. Induced pluripotent stem cells in drug discovery. Nat. Rev. Drug Discov. 14 (5), 356–372. https://doi.org/10.1038/nrd4618.

Aronov, K.S., Kerem, A., 2016. Towards a unifying theory of neural circuit development. Bioessays 38 (9), 746–757. https://doi.org/10.1002/bies.20166.

Aronov, A., Landau, D., Regev, A., 2017. Convolutional neural networks for single-cell data. Cell Syst. 4 (2), 223–236. https://doi.org/10.1016/j.cels.2017.03.014.

Aronov, A., Regev, A., 2018. Deep learning for single-cell data analysis. Cell 173 (1), 125–140. https://doi.org/10.1016/j.cell.2018.01.026.

Aronov, A., Regev, A., 2019. Single-cell resolution mapping of brain connectivity at single-cell resolution. Neuron 100 (2), 261–279. https://doi.org/10.1016/j.neuron.2019.01.018.

Aronov, A., Regev, A., 2021. Data assimilation for single-cell analysis. Nat. Rev. Genet. 22 (3), 155–168. https://doi.org/10.1038/s41576-020-0049-5.

Arras, T., Dueker, L., 2019. Machine learning in neuroscience: a review. Front. Neurosci. 13, 379. https://doi.org/10.3389/fnins.2019.00379.

Aronov, A., Regev, A., 2019. Single-cell resolution mapping of brain connectivity at single-cell resolution. Neuron 100 (2), 261–279. https://doi.org/10.1016/j.neuron.2019.01.018.

Aronov, A., Regev, A., 2021. Data assimilation for single-cell analysis. Nat. Rev. Genet. 22 (3), 155–168. https://doi.org/10.1038/s41576-020-0049-5.

Aronov, A., Regev, A., 2019. Single-cell resolution mapping of brain connectivity at single-cell resolution. Neuron 100 (2), 261–279. https://doi.org/10.1016/j.neuron.2019.01.018.

Aronov, A., Regev, A., 2021. Data assimilation for single-cell analysis. Nat. Rev. Genet. 22 (3), 155–168. https://doi.org/10.1038/s41576-020-0049-5.

Aronov, A., Regev, A., 2019. Single-cell resolution mapping of brain connectivity at single-cell resolution. Neuron 100 (2), 261–279. https://doi.org/10.1016/j.neuron.2019.01.018.
