Diploma thesis

LINKING CONNECTIVITIES AND GENE EXPRESSION PATTERNS IN THE MOUSE BRAIN

September 30, 2022

Tilman Hinnerichs
Matrikelnummer: 4643427
Technische Universität Dresden

Tutor: Dr. Nico Scherf
MPI for CBS

Summer semester 2022

proper title?

Abstract

Contents

To be sorted somewhere

- Variability and different interpretations of different graph convolutional neural filters [Kipf and Welling, 2016, Li et al., 2020, Feng et al., 2022] etc.
- DeepGOPlus for feature generation [Kulmanov and Hoehndorf, 2019]
- discussion of different PPI network databases [Szklarczyk et al., 2014]
- discussion of potential databases associating gene expression data with their spatial distribution [Hawrylycz et al., 2011]
- discussion of best neural learning/graph convolutional methods [Paszke et al., 2019, Fey and Lenssen, 2019]
- how to handle highly imbalanced data, metrics, preprocessing, sampling, modification of loss function [Jeni et al., 2013] and optimization over them (with Adam[Kingma and Ba, 2015])
- maybe introduction of PhenomeNET for MP/GO for more sophisticated protein representation [Hoehndorf et al., 2011, Ashburner et al., 2000, Carbon et al., 2020, Smith and Eppig, 2009] and derive features from DL2vec [Chen et al., 2020, Mikolov et al., 2013]
- evaluation of „Using ontology embeddings for structural inductive bias in gene expression data analysis“[Trebacz et al., 2020]
- take some ideas from Zitnik and Leskovec [2017] with title „Predicting multicellular function through multi-layer tissue networks“. (OhmNet)
- potentially group results based on InterPro[Blum et al., 2020] families eventually
- choice of model organism?!

1 Introduction

General thread for introduction and motivation:

- Gene expression patterns are difficult to analyze in humans → take mouse as model organisms
- why do we study sane mice and not a
- The brain is a multi-level system in which the high-level functions are generated by low-level genetic mechanisms. Thus, elucidating the relationship among multiple brain levels via correlative and predictive analytics is an important area in brain research. Currently, studies in multiple species have indicated that the spatiotemporal gene expression patterns are predictive of brain wiring. Specifically, results on the worm Caenorhabditis elegans have shown that the prediction of neuronal connectivity using gene expression signatures yielded statistically significant results.
- no in-depth analysis of mouse brain genetic patterns and their relation to different connectivity patterns has been made yet
- Why are we concerned with gene expression prediction and what could it tell us?
  - [Twine et al., 2011] show the importance of gene expression patterns, by linking gene expression abberation with increase in Alzheimer's disease
  - studies have shown circadian patterns of gene expression in human brain and the disruption of those in depressive disorder [Li et al., 2013]
  - first understand sane brain and its circuits, before tackling pathological data
- why is finding (low-dimensional) patterns important here?
- What is structural and functional connectivity and what are associated hypotheses?
  - [Fornito et al., 2015] elaborate on the connectomics of brain disorders and its complexity in connectivity. Understanding how brain networks respond to pathological perturbations is crucial for understanding brain disorders and behavior
- Why is finding a link or to connectivity from gene expression desirable?
- Why do we think that GCNs could help finding such patterns?
  - Guilt by association over gene networks [Oliver, 2000, Gillis and Pavlidis, 2012] in genetic networks
  - protein function prediction from PPI networks [Vazquez et al., 2003]
  - GCNs have been applied successfully to variety of tasks over different types of graphs.
- what are our contributions?
  - Showed that graph convolution over PPI graphs helps finding patterns in gene expression data
  - Contributed an implementation of our method
  - contributed an implementation of KerGNN to PyTorch Geometric
  - built an open framework for parametric UMAP in torch to integrate non-UMAP graphs
- what is the outline of this script?

General Introduction of the Research Study

Research problem or Questions with Sub-Questions

Reasons or Needs for the Research Study/Motivation for my research

Definition and explanation of Key Terminology

Context of Research Study within th Greater Discipline

- Introduction to mouse brains as model organisms for insights into human brain
- Works on mouse brain in general and potential tasks
- works on gene expression in mouse brains
  - traditional approaches
  - importance of gene expression patterns in mouse brains
- neural networks for this purpose
  - how were
- gene expression for general tissue

2 Literature overview

2.1 Gene expression databases and prediction

Research in gene expression prediction and profiling has a long history in bioinformatics and systems biology, but was almost exclusively linked to cancer research. Moreover, with the rise of machine learning, and more specifically (deep) neural networks and its variants, this field became increasingly data reliant. The Human Genome Project [Watson, 1990], launched in 1990 and declared finished in 2003 while the first gapless assembly was finished in 2022, also sparked various works in relating these genetic representations to other tissue- and individual-specific properties and traits.

For comparison of gene expression profiling works there exist multiple prominent variables. Most significantly, the chosen organism is a crucial choice for both data availability and predictive complexity. Second, the chosen tissue is naturally important for the proposed hypotheses, especially with respect to tissue definitive cancer research, and its potential ability to generalize without transfer learning. While gene expression pattern analysis approaches frequently focus on tissues like *mamma*, (primarily female) breast, [Herschkowitz et al., 2007], liver [Flores-Morales et al., 2002], and skeletal muscle [Lecker et al., 2004] for exploration of diseases like cancer and atrophy, respectively, in humans.

However, the nervous system is often investigated separately as it bears different molecular processes and structure, anatomy and cell life cycles, while brain and spinal cord are even based in a separate nutritional circuit for mammals. Moreover, gene expression determination in the human brain may almost certainly remain an deadly intervention for most brain tissues, hence allowing only for careful extraction of specific tissues in living organisms. Also this disallows for *in-vivo* extraction of vital brain regions and structures, e.g. the brainstem. Furthermore, the human brain's gene expression patterns are varied and diversified [Ramasamy et al., 2014], aligning with its anatomical and embryogenesis complexity, and its compartments are exceptionally and deeply connective and collaborative [Fornito et al., 2015]. Both also hold for invertebrates, i.e. insects. Thus, full genetic profiles of expression are mandatory for a full understanding of the mammalian and invertebrate brain and primary nervous system, respectively. By the strong intervention of the tissue extraction, full genome atlases are fit together from various experiments on multiple individuals.

The human brain is among the most intricate and complicated networks we do know of, and is far from being fully understood. Additionally, full transcriptomic atlases of human brains are difficult to collect while raising decisive privacy concerns. Yet, there were multiple efforts and projects with rather small sample sizes. A detailed elaboration on dataset and organism choice, and their respective properties may be found in Section 3.2.

However, there have been works on numerous works for other tissues and other organisms. modENCODE Consortium et al. [2010] correlate activity patterns in the regulatory network within *Drosophile*, proposing their model for identification of functional elements "modENCODE". As this work was published back in 2010, the approach relies purely on statistical correlation and covariance. Chikina et al. [2009] follow a similar approach in *C. elegans* predicting tissue-specific gene expression in 2006 utilizing support-vector machines (SVM)[Noble, 2006].

More modern, data-oriented machine learning models such as (deep) neural networks (NN) were applied successfully to similar problems. Aromolaran et al. [2020] achieved to predict essential genes based on their respective sequence and functional features profiting off NNs, while transcriptomic interaction prediction was done based on functional gene data using deep learning in Yang et al. [2019] in *Drosophila* over different tissues.

Within humans, as mentioned previously, gene expression was primarily used for cancer and disease research. Schulte-Sasse et al. [2021] and Wang et al. [2021] were the first to apply graph convolutional neural networks to the task of gene expression prediction within humans. While Schulte-Sasse et al. [2021] was applied on data from The Cancer Genome Atlas (TCGA)[Tomczak et al., 2015] across multiple tissues, Wang et al. [2021]'s MOGONET is proposed as a general framework for gene expression prediction with example computations on ROSMAP dataset and TCGA. Both approaches implement the original formulation of GCNs[Kipf and Welling, 2016], which we will discuss in more detail in Section 3.3.2, over protein-protein interaction networks and accomplish outstanding performances and both measure biomarker importance for prediction in order to leverage explainability. The authors thereby exploit the ”guilt by association” principle [Oliver, 2000, Gillis and Pavlidis, 2012] over gene networks, adding background knowledge such as biological interaction and pathways.

Crucial for almost all classification tasks in machine learning is the choice of entity representation. In the mentioned works molecular [Schulte-Sasse et al., 2021, modENCODE Consortium et al., 2010, Noble, 2006] and phenotypical [Wang et al., 2021, Chikina et al., 2009] features were used for expression prediction, but never both combined. The combination of phenotypical and molecular features over GCNs was proven to raise predictive performance in drug-target interaction prediction [Hinnerichs and Hoehndorf, 2021] but remains an open challenge for this very task.

2.2 Finding spatial patterns in gene expression in mice brains

In this subsection we will constrain the issue of gene expression analysis to both ”spatial patterns”, mammals and the tissues of the brain, which we study in this work. The term *spatial patterns* is rather vague and allows for various interpretations, both discrete and continuous, which will form the classes for the following literature review.

In Pavlidis and Noble [2001] is the first first review paper on regional variation in genetic expression in mouse brain, up to our knowledge. Zapala et al. [2005] is also among the earliest works, showing that local structures beared ”transcriptional imprint” that coincide with the embryological origin of the examined regions. However, they only were able to identify up to 24 neural tissues. They further conclude that this may be important for functional collaboration within the adult mouse brain. The authors measure pairwise correlation show the existence of clusters over a heatmap.

The Allen Institute Brain Atlas (AIBA), is a collection various atlases such as Allen Mouse Brain Atlas (AMBA)[Lein et al., 2006, Daigle et al., 2018], the Allen Mouse Brain Connectivity Atlas (AMBCA) [Oh et al., 2014, Harris et al., 2019] and the Allen Mouse Brain Common Coordinate Framework (CCFv3) [Wang et al., 2020] to name only the ones related to adult mice’s brains. As it was the first coherent collection of spatially resolved expression values, mapping 2D expression images consistently to 3D coordinates, the AMBA has sparked a range of publications. Within Lein et al. [2006], the Allen Institute also published the ”Allen Reference Atlas”(ARA) proposing a number morphological and histologically induced sub-regions of the brain and hence a precisely defined parcelation. Moreover, they propose the ARA *ontology*, a semnatic hierarchy, providing a hierarchical cluster of all sub-structures and map them back to their coordinates with the CCF.

Name a few ”sparked” research works on this

Bohland et al. [2010] advance clustering of such expressions under usage of singular value decomposition (SVD) within mice, combined with an extensive analysis of similarities to neuroanatomy. Likewise, Takata et al. [2021] propose a flexible annotation atlas of the mouse brain, introducing a flexible ontology construction framework which may be used on the transcriptomic data such as the AMBA, leveraging anatomic structure and axonal projection data. Here, FAA focuses on consistent and reproducible regions-of-interest (ROIs) definition for other downstream tasks such as resting-state functional connectivity annotation. Further, this ontology may be seen as a pattern within mouse brain, while it may only detect connected structures.

The authors of Valk et al. [2020] analyze structural covariance of cortical thickness within primate brains, namely macaques, and its correlation to each cortical layers transcriptome. Further, transcriptomic variation was related to a continuum of functions by mapping them the brain anatomy, inducing a *continuous*, functional parcelation of the primates brain. Further, this study suggests a relation of functional and transcriptomic links. Similarly, Zeng et al. [2015] propose a range of deep learning methods for capturing spatiality of gene expression within the mouse brain. A notable addition is the work of Kelly and Black [2020], presenting an R package for simulating gene expression from graph structures over general biological pathways, that may be and was applied to (mammalian) brains prospectively.

While also focused on mouse brains, Partel et al. [2020] submits a novel database based on their own *in-situ* sequencing data, and a consecutive spatial gene expression analysis pipeline, and relates the results to tissue morphology and hence indirectly to the AMBA. Similarly to our proposed approach, brain parcelation are present as n-dimensional, continuous embeddings, representing closeness in gene expression space. Due to the similarity in the pipeline especially in their visualization utilizing UMAP, we will use the generated images of this work for a brief comparison in Section 4.

Up until now, GCNs were not applied to this issue.

2.3 Structural and functional connectivity prediction

In this section we will examine related work on brain connectivity prediction from transcriptomic data. We hereby separate axonal and functional connectivity due to their differing associated hypotheses.

The relation of gene expression patterns and *structural* connectivity was studied numerous times over various model organisms, especially*C. elegans*, *Mus musculus*, but also humans. We will categorize existing literature with respect to the underlying organism.

Kaufman et al. [2006] and Varadan et al. [2006] were among the earlier works on this research field, followed by Arnatkevičiūtė et al. [2018], showing the relation of axonal connectivity and gene expression within *C. elegans*. Further, that relation was shown by Rubinov et al. [2015], Fakhry et al. [2015] and Fulcher and Fornito [2016] for the mouse brain, while Parkes et al. [2017] and Goel et al. [2014] proved a correlation within human brains.

Fakhry and Ji [2015] is among the earlier works focusing on the predictive power across different mouse brain regions. They applied non-machine learning, computational models for axonal connectivity prediction in adult mouse brains. Similarly, Roberti et al. [2019] uses transcriptomic information to anatomical connectivity patterns and gene expression of neurons using (shallow) neural networks. Yield a 85% accuracy in prediction of unconnected and connected regions. Both shall serve as a baseline performance in chapter 4. Only recently, Wang et al. [2022] proposed a novel-network based method integrating molecular-based gene association networks such as protein-protein interaction networks with brain connectome data. They further link these gene expression patterns to four brain diseases, including Alzheimer's disease, Parkinson's disease, major depressive disorder and autism.

The correlation of *functional* connectivities and transcriptomic data is much more complex in nature than the previous task. We will again classify approaches by their respective model organism.

Whitfield et al. [2003] were one of the first to link transcriptomic data with behavior and hence functional patterns in individual honey bees back in 2003. The authors show that changes in the messenger RNA were connected to behavior and how changes to RNA directly influenced the other. Rankin [2002] first developed the idea of combining behavioral analyses of *C. elegans* with their genetics. Further, Sun and Hobert [2021] only recently described the distinct functional states and the corresponding distinct molecular states within the transcriptome. While honey bees and nematodes are rather simple model organisms, enabling both full transcriptomic analyses of the organisms, and their bearing and actions. However, "behavior" may be ambiguous and vague for such taxonomically distant animals, from the viewpoint of humans, and may only be linked to very basic meta-tasks such as basic routing, orientation and basic social interaction.

Research on humans further indicates correlation of transcriptomic patterns and "neural dynamics", concluded from e.g. fMRI data [Richiardi et al., 2015, Diez and Sepulcre, 2018, Vértes et al., 2016] or electrocorticography [Betzel et al., 2019]. We refer to Fulcher et al. [2021] for an extensive overview on the link to axonal and functional connectivity. Further, Zerbi et al. [2021] propose a computational model calibrated over 16 autistic mouse models, that reveals a range of functional connectivity subclasses and -types.

Brief Overview of LIterature Reviewed, Discussed and applied

Study Model and Process Aligning with literature reviewed

Hypotheses and justifications tied to prior sections and statements

The Scope of the study with theoretical assumptions and limitations

3 Materials and methods

In this study, we utilized and incorporated various approaches from other works and applied them to diverse datasets. The following section will give a brief overview over all modules of the proposed models, while the entire computational methods will be presented and described in the results section (Section 4). We further introduce the goals and scopes of our respective research questions and on our evaluation metrics for this purpose.

Introduction and general description, study method and study design

3.1 Problem description

Here we give a brief introduction to each of the three tackled issues and further summarize data properties, challenges and goals of each problem in this section and Section 3.2.

3.1.1 Spatial gene expression prediction

Firstly, the issue of spatial gene expression prediction is concerned with the following problem: Within a given structure or at a specific coordinate, and for a given gene, we want to determine whether the latter is expressed or not. While there are also ways to quantify the expression within a region, we only care about the *quality*, i.e. whether the is expressed or not. We treat all structure-gene pairs without a known expression as negatives, thus handling the dataset in a closed-world manner, and accordingly formulate the problem as a binary classification task. Naturally, as shown in Section 2.2, mammalian brains show a high correlation of gene expression and neuroanatomic substructures, suggesting importance of nearby and adjacent structures to the considered one. While spatial prediction and conditionals are hard to infuse into models, we will therefore start by predicting gene expression within single structures, constructing train and validation set over genes, or train from related or proximate regions.

3.1.2 Preserving dimensionality reduction in brains

The second studied issue is the task of preserving dimensionality reduction. Here, all regions or 3D-voxels are associated with a vector embedding of fixed dimensionality n, representing each section with the features of choice. As we are concerned with the influence and patterns of gene expression we will only use and consider embeddings of transcriptome for this task. The eventual task is to find a mapping $f_{emb} : \mathbb{R}^n \rightarrow \mathbb{R}^k$ with $k < n$ reducing each structure's representation dimensionality to k, such that the relative, pairwise distances are preserved. Hence, if two sections share similar expressions, they should be described similarly in $\mathbb{R}^k$, invariant to spatial distance. As we seek to visualize such embeddings for quality assessment of the embeddings and characterization of marked regions, a mapping into color-space, i.e. choosing $k \in \{1, 2, 3\}$, appears natural.

3.1.3 Structural and functional connectivity prediction

Eventually, we want to forecast the brain's connectome in our third formulation: Given two structures we want to predict whether there is a connection either of axonal or functional or both types. Therefore, similarly to the question of dimensionality, regions and voxels, respectively, are represented by their expression characteristics in vector space. Likewise to other works, introduced and explained in Section 4, and to the issue described in 3.1.1, our analysis is invariant to the eventual "strength" of the connection, but focuses on the quality of connectivity. Hence, we define a cut-off threshold converting the issue to a binary classification task, whereas non-positive and unknown links are treated as negatives.

Assumptions of study method and study design with implied

3.2 Datasets and preprocessing

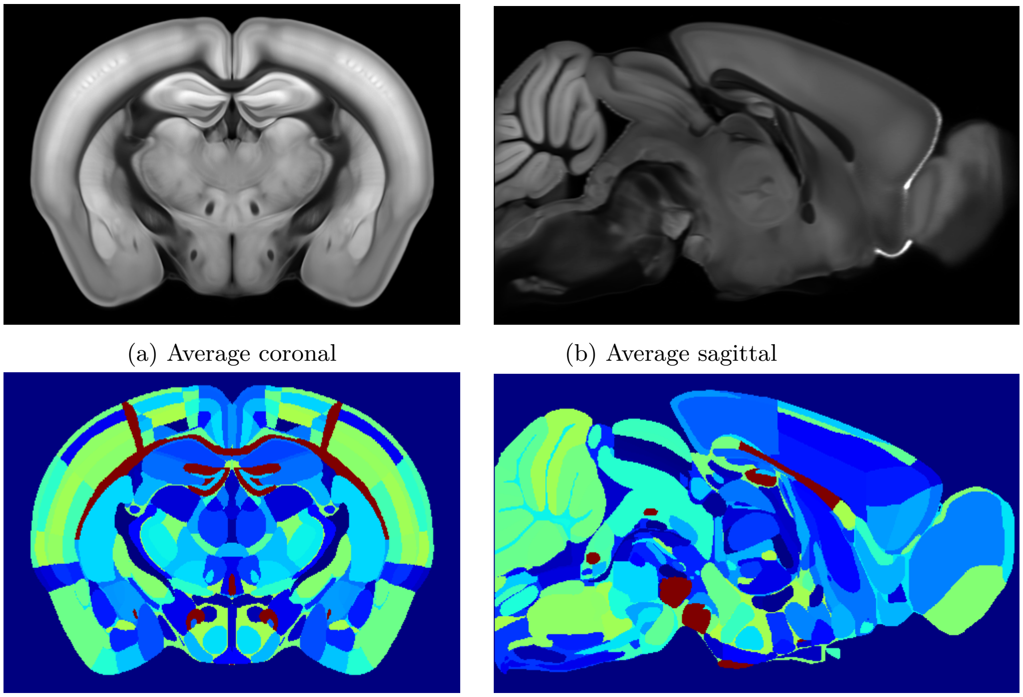
As human brains are among the most complex in structure and connectivity within nature, a full transcriptomic atlas may be very valuable for the research community and our experiments in this work. However, full transcriptomic atlases of homo sapiens are ethically difficult to gather. Additionally, as a valuable, public genetic atlas of deceased relatives may provide highly critical information about the remaining, living ones, such as genetic diseases, genetic markers for correlating with addiction and other social behavior, or ancestry in general, this raises tremendous privacy concerns. As we aim to investigate transcriptomic patterns in the brain and their relation to structural and functional connectivity as a generalized, organism-invariant methodology, we also want our experiments to be as understandable and replicable as possible. However, there have been multiple initiatives towards collaborative and open human brain data, such as the Allen Human Brain Atlas (AHBA) [Hawrylycz et al., 2011] also published by the Allen Institute and the Human Brain Atlas (HBA) [Roland et al., 1994]. While both are almost complete, e.g. AHBA considers over 20000 genes, but these were collected from just 6 human individuals. In combination with the Human Connectome Project (HCP) [Van Essen et al., 2013] this atlas provides a valuable, matched data resource. As rodents, and more specifically mice, are more simplistic and well studied in behavior and due to their taxonomic proximity to humans serve as model organisms for diverse genetic, social and medical experiments, we opted for mice as the study organism. The ultimate goal still shall be the further understanding of brains of our species.

Furthermore, immense effort was put into enormous projects and databases for model invertebrates, namely *Drosophila* (specifically *Drosophila melanogaster*, also called *fruit fly*) and *Caenorhabditis elegans* (short: *"C. elegans"*, colloquially also called *roundworm*) with the two projects "Virtual Fly Brain" [Milyaev et al., 2012] (https://virtualflybrain.org/) and "Wormbase" [Lee and Sternberg, 2003, Davis et al., 2022] (https://wormbase.org), respectively. Yet, we wanted to stay within the same taxonomic phylum leading our choice towards mouse brains.

3.2.1 Spatial gene expression values in mice brain

The AMBA[Lein et al., 2006] is a genome-wise and comprehensive, digital 3D-map of spatial gene expression of the adult mouse central nervous system (CNS). AMBA utilizes *in situ* hybridization (ISH) for expression measurement and is publicly and freely available. In literature there are several approaches, compromising sequencing depth, accuracy, throughput and spatial resolution. Generally, one can identify two classes of gene expression measurement while preserving spatial information. The first approach is to store spacial coordinates first, followed by a the sequencing of single-cell RNA where Achim et al. [2015] and Chen et al. [2017] propose the mapping and the Geo-seq protocol for this method. The second method includes the usage of "barcodes", decoded in the tissue sample, while running a parallel analysis of numerous mRNAs [Ke et al., 2013, Moffitt et al., 2016]. Here, AMBA and Partel et al. [2020]'s dataset are constructed using the first and the latter method, respectively. Sample extraction is very invasive and hence samples are collected from multiple individuals. Please see Ng et al. [2007] for more information and an overview on applied 3D reconstruction and registration of ISH images.

In AMBA over 17000 genes were measured in voxels in resolutions of $10\mu m$, $25\mu m$, $50\mu m$ or $100\mu m$. An "average brain", i.e. averaging expression intensities of each voxel among all genes, is shown in Figure 1a and 1b as coronal and sagittal cross-sections. The most important supportive dataset for AMBA is the Common Coordinate Framework Version 3 (CCFv3)[Wang et al., 2020] structuring and mapping all brain voxels into identifiable sub-structures, based on their morphology. Unfortunately, functional connectivity (FC) measurement is often measured with respect to regions-of-interest (ROIs) that may coincide with these structures. Therefore, full brain FC datasets on voxel level may not



(a) Average coronal (b) Average sagittal

(c) Annotation coronal (d) Annotation sagittal

Figure 1: (a)-(b): Coronal and sagittal view on the "average brain", an averaged map of all gene expression values gathered from AMBA; (c)-(d): Coronal and sagittal view on the annotation atlas, partitioning the reference brain into ≈ 900 sub-structures.

be publicly available in the same resolution. Hence, for a full integration of all datasets, we will use the CCFv3 grouped voxel for structures, thus redefining all three problem descriptions from voxel to structure level, i.e. a set of voxels. A visualization of these structures based on the CCFv3 annotation volume is depicted in Figure 1c and 1d.

Expression values in AMBA are provided per structure as expression densities, intensities and energies. Let g be the considered gene, d the considered division or substructure, $P_{g,d}$ be the set of pixels within an ISH image and $P_{g,d,e}$ the set of expressing pixels, displaying expression of g in d. Both $P_{g,d}$ and $P_{g,d,e}$ are given as measured expression intensities for the respective pixels. Thus, the three measurements are defined as follows:

Metric name	Formula	Description
Expression density	$E_d := \frac{\lvert P_{g,d,e}\rvert}{\lvert P_{g,d}\rvert}$	sum of expressing pixels / sum of all pixels in division
Expression intensity	$E_i := \frac{\sum_{p_e \in P_{g,d,e}} p_e}{\lvert P_{g,d,e}\rvert}$	sum of expressing pixel intensity / sum of expressing pixels
Expression energy	$E_i := \frac{\sum_{p_e \in P_{g,d,e}} p_e}{\lvert P_{g,d}\rvert}$	sum of expressing pixel intensity / sum of all pixels in division

The respective description is provided by the official Allen Mouse Brain Documentation, which we note as a reference for further information on these metrics. As expression energy provides an in-division normalization within itself, is invariant to structure size and is dependent on the actually measured intensities, it appears to be the natural choice. It was further used by all previously mentioned studies over AMBA as ground-truth; we will hence do likewise, and thus use expression value and expression energy interchangeably.

We perform a gene-set enrichment analysis (GSEA) on the expression values. GSEA is a method facilitating recognition of classes that are expressed frequently, and thus may be over-represented in the expression data. We hereby follow the works of Subramanian et al. [2007] and Kuleshov et al. [2016] for this purpose, and integrate them into our preprocessing pipeline.

With S the set of structures and G the set of genes, we are now able to construct a matrix $M_{GE} \in \mathbb{R}^{S \times G}$. For simplicity, we will only consider structures with at least one non-zero gene expression value and genes that have a non-zero expression energy in at least one structure. This leaves us with $|S| = 843$ structures and $|G| = 16679$ genes. Note further, that the measured pixel-wise intensities are not normalized and hence the matrix entries, i.e. expression energies, may not be within the interval $[0, 1]$. Hence, we will first normalize the matrix. However, the briefly put process of normalization remains non-trivial, not within its computational complexity, but within its neuro-biological implications and constraints. More specifically, there may be ”effective” and ”ineffective” genes. ”Effective genes” may have a high impact on other downstream processes with even a few low expression intensities on one side, while ”ineffective” genes have low impact albeit high expression intensities and high spread among the considered substructure. Summing up, the expression entries in M_{GE} may not be proportional for its actual importance.

An example may be genes describing and piloting cell proliferation. As cell division is naturally lowered in brains, in comparison to e.g. bone marrow, due to the density of neural cells, those genes may not be expressed as much. A significant increase of such may hence indicate functional deviant purpose or even pathological, i.e. cancerous, tissues, but works on low expression levels thus representing an ”effective” gene. On the other hand, genes managing nutritional supply, may be highly expressed in all cells at all times, over-ruling the above ”effective gene” in its expression intensities, thus constituting an ”ineffective” gene.

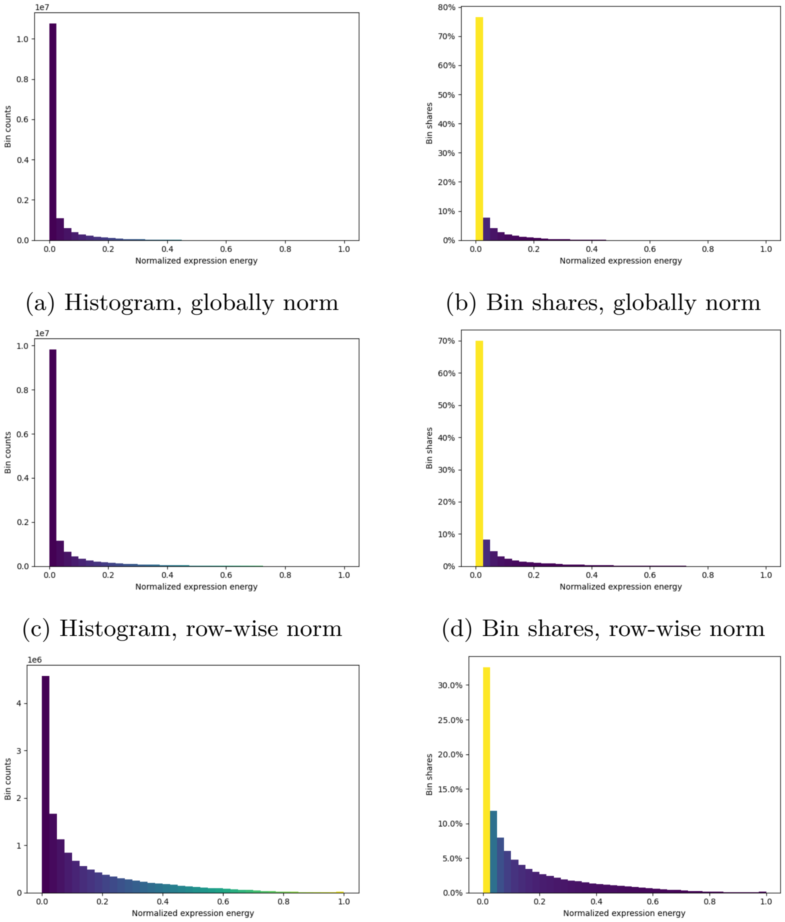
Thus, there exist three different normalization schemes for this matrix:

1. A global normalization by dividing all matrix values by its global maximum,
2. a row-wise or per-structure normalization scheme, and
3. a column-wise or per-gene normalization.

While the global normalization remains the most common, it potentially leads to even lower values for ”effective” and preserves over-represented and exaggerated expression energies for ”ineffective” genes. Further per-structure, i.e. row-wise, normalization allows for highlighting of expression values that are significantly lower or higher among a single structure. Unfortunately, a row-wise normalization scheme will not break the above bias, but migrates the issue from the global to a structure-level scale. The third normalization scheme helps us to identify values associated with a specific gene, which are highly expressed, significantly lower or higher among the samples, i.e. *relative* expression intensities. This scheme helps us solving the issue of invariance to effectiveness of genes and allows for a fair representation of the transcriptome. The overall distribution of all all normalization schemes are depicted in Figure 2a to 2f.

Edit figure caption

Especially for our binary classification task of gene expression prediction (see Section 3.1.1), we need mapping of those entries to $\{0, 1\}$. Moreover, neural networks prefer inputs in the interval $[0, 1]$ as we want to use M_{GE} as input for the problems described in Sections 3.1.2 and 3.1.3. Thus, we threshold the resulting normalized expression energies, by applying a fixed cut-off threshold $t \in [0, 1]$. We experiment with various expression thresholds as described in results Section 4.



(a) Histogram, globally norm

(b) Bin shares, globally norm

(c) Histogram, row-wise norm

(d) Bin shares, row-wise norm

(e) Histogram, column-wise norm

(f) Bin shares, column-wise norm

Figure 2: $|S| = 843$, $|G| = 16679$

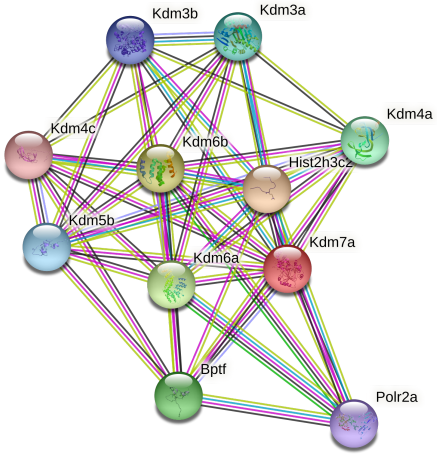


Figure 3: Protein-protein interaction graph for *Kdm7a*, an histone demethylase required for brain development, that is closely related to e.g. *Hist2h3c2*, a protein playing central role in transcription regulation and thus proliferation activity indicating e.g. malign tumors

3.2.2 Protein-protein interaction graph

As shortly sketched in the introduction (see Section 1), we want to apply graph convolutional neural networks over protein-protein interaction (PPI) graphs and identify patterns within. However, there are numerous databases for PPI graphs publicly and freely available. Common sources are CPDB[Lo et al., 2009], iRefIndex[Razick et al., 2008], MultiNet[Sengupta et al., 2023] and STRING[Szklarczyk et al., 2014]. Due to our experience with the database and its success in other related tasks [Schulte-Sasse et al., 2021, Wang et al., 2021, Hinnerichs and Hoehndorf, 2021], we will use STRING (Version 10) as ground truth data. As STRING provide probabilities and hence confidence scores for each interaction, we re-use the recommended threshold of 0.7 in order to retrieve only high-confidence interactions. STRING database contains 300000 interactions between over 20000 genes gathered from other databases and literature for *Mus musculus*. An example interaction graph for *Kdm7a*, an histone demethylase required for brain development, is shown in Figure 3.

3.2.3 Structural and functional connectivity databases

in-depth description of the study design/datasets used and motivation why they were used for these experiments

Explanation of Sample used in the study

- how many positives, samples, individuals
  - over all intensities →
  - per structure →
  - per gene →
- experimental setup from Allen Institute for axonal projection data
- paraphrase description of „Technical tour: Explore the Allen Mouse Brain Connectivity Atlas“
- why were these datasets used and not others?ss

- How did we achieve the matching?
- what are premises of the dataset?
- transfer learning working for other structure/regions
- dataset: Allen Mouse brain atlas vs.
    - phenoview impc data
    - mousephenotype
    - HPO/MP project expression data
- Mouse brain CCFv3
- Where is data coming from? [Pallast et al., 2019] (AIDAmri)
- How to calculate functional connectivity matrix → AIDAconnect (no paper yet? cite dataset?
- How to combine functional connectivity for multiple samples?

Four graphs were used in this study:

- Protein-protein interaction graph from STRING
- structure hierarchy/ontology from [Lein et al., 2006]
- structural connectivity data from (Mouse Projection data)
- functional connectivity data from [Pallast et al., 2019]

3.3 Model

Explanation of Measurement, Definitions, Indexes, Reliabililty and Validity of study method and study design

Description of Analytical Tehcniques to be Applied and justification for them

Reliability and validity of internal/external design and related subtypes

3.3.1 Feature generation

Data preparation for regression task

- unbalanced data for prediction task

3.3.2 Graph convolutional neural layers

We include these molecular and ontology-based sub-models within a graph neural network (GNN) [Kipf and Welling, 2016]. The graph underlying the GNN is based on the protein–protein interaction (PPI) graph. The PPI dataset is represented by a graph $G = (V, E)$, where each protein is represented by a vertex $v \in V$, and each edge $e \in E \subseteq V \times V$ represents an interaction between two proteins. Additionally, we introduce a mapping $x : V \to \mathbb{R}^d$ projecting each vertex v to its node feature $x_v := x(v)$, where d denotes the dimensionality of the node features.

A graph convolutional layer [Kipf and Welling, 2016] consists of a learnable weight matrix followed by an aggregation step, formalized by

$$\mathbf{X}' = \hat{\mathbf{D}}^{-1/2}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-1/2}\mathbf{X}\mathbf{\Theta} \tag{1}$$

where for a given graph $G = (V, E)$, $\hat{A} = A + I$ denotes the adjacency matrix with added self-loops for each vertex, D is described by $\hat{D}_{ii} = \sum_{j=0} \hat{A}_{ij}$, a diagonal matrix displaying the degree of each node, and Θ denotes the learnable weight matrix. Added self-loops enforce that each node representation is directly dependent on its own preceding one. The number of graph convolutional layers stacked equals the radius of relevant nodes for each vertex within the graph.

The update rule for each node is given by a message passing scheme formalized by

$$\mathbf{x}'_i = \mathbf{\Theta} \sum_{j}^{N} \frac{1}{\sqrt{\hat{d}_j \hat{d}_i}} \mathbf{x}_j \tag{2}$$

where both $\hat{d}_i, \hat{d}_j$ are dependent on the edge weights e_{ij} of the graph. With simple, single-valued edge weights such as $e_{ij} = 1 \ \forall (i, j) \in E$, all $\hat{d}_i$ reduce to d_i, i.e., the degree of each vertex i. We denote this type of graph convolutional neural layers with GCNConv.

While in this initial formulation of a GCNConv the node-wise update step is defined by the sum over all neighboring node representations, we can alter this formulation to other message passing schemes. We can rearrange the order of activation function σ, aggregation AGG, and linear neural layer MLP with this formulation as proposed by [Li et al., 2020]:

$$\mathbf{x}'_i = \mathrm{MLP}\left(\mathbf{x}_i + \mathrm{AGG}\left(\left\{\sigma\left(\mathbf{x}_j + \mathbf{e_{ji}}\right) + \epsilon : j \in \mathcal{N}(i)\right\}\right)\right) \tag{3}$$

where we only consider $\sigma \in \{\mathrm{ReLU}, \mathrm{LeakyReLU}\}$. We denote this generalized layer type as GENConv following the notation of PyTorch Geometric [Fey and Lenssen, 2019]. While the reordering is mainly important for numerical stability, this alteration also addresses the vanishing gradient problem for deeper convolutional networks [Li et al., 2020]. Additionally, we can also generalize the aggregation function to allow different weighting functions such as learnable SoftMax or Power for the incoming signals for each vertex, substituting the averaging step in GCNConv. Hence, while GCNConv suffers from both vanishing gradients and signal fading for large scale and highly connected graphs, each propagation step in GENConv emphasizes signals with values close to 0 and 1. The same convolutional filter and weight matrix are applied to and learned for all nodes simultaneously. We further employ another mechanism to avoid redundancy and fading signals in stacked graph convolutional networks, using residual connections and a normalization scheme [Li et al., 2019] [Li et al., 2020] as shown in Supplementary 3. The residual blocks are reusable and can be stacked multiple times.

- what is GATConv?
- what is KerGNN and what is its idea?
- add some sentences to the section above
- node vs. graph classification vs. link prediction

3.3.3 Dimensionality reduction techniques

3.3.3.1 Principal component analysis (PCA)

3.3.3.2 tSNE

3.3.3.3 UMAP

3.3.3.4 Parametric UMAP

3.3.4 Hyperparameter tuning

- RayTune[Liaw et al., 2018] for automated hyperparameter tuning

3.4 Evaluation and metrics

- AUC and AUPR for gene expression prediction
- self-built metric for evaluation of dim-red
- AUC and AUPR for conn pred

4 Results

4.1 Gene expression prediction

- We originally started from the per section prediction in order to paste its performance and results to other ”related” structures within in the mouse brain. We propose multiple ideas .... As mentioned we used three different feature types in this study. ...(molecular features, phenotypical features, pure taxonomic features (InterPro embedding))) ...Due to the poor performance of the predictor with all three used feature types, we abandoned these plane
- what is our baseline here? -> no study on prediction yet in adult mouse brains
  -
- structure specific features?
  - structural ontology / closeness
  - developmental hierarchy of tissue

Our model also allows us to test different ways of representing omics data. We tested different ways to normalize values assigned to genes as these normalizations convey different biological information; in the matrix of values assigned to genes from cancer samples, we can normalize values across the entire matrix, across each row (cancer sample), or across each column (gene). While a global normalization is more common, row-based normalization allows us to highlight values that are significantly higher or lower within one sample (e.g., which genes are expressed at high or low levels within a single sample), and column-based normalization allows us to highlight values assigned to a particular gene that are significantly higher or lower within one sample (e.g., whether a gene is expressed at higher or lower levels within one sample compared to all others). We find that column-based normalization performs better than row-based normalization, while the global normalization approach performs close to random. The best results are achieved when combining both row- and column-based normalization (Supplementary Table 2).

4.2 Dimensionality reduction and its combination with different graphs structures

- plot for showing validity of embeddings: K-means colour with respect to cluster
- plot colour parent structure all similar

4.3 On the linkage of connectivities and gene expression patterns

Brief Overview of Material

Findings (Results) of the Method of Study and Any Unplanned or Unexpected Situations that Occurred

Brief Descriptive Analysis Reliability and Validity of the Analysis

Explanation of the Hypothesis and Precise and Exact Data (Do Not Give Your Opinion)

5 Discussion

Brief Overview of Material

Full Discussion of Findings (Results) and Implications

Full Discussion of Research Analysis of Findings

Full Discussion of Hypothesis and of Findings

Post Analysis and Implications of Hypothesis and of Findings

Novelty:

- GCNs over gene expression was never applied here

6 Conclusion

Summary of Academic Study

Reference to Literature Review

Implications of Academic Study

Limitations of the Theory or Method of Research

Recommendations or Suggestions of Future Academic Study

- gene expression patterns within mouse brain and both possible hypothesis and tasks, and models over this
- gene knockout models and whether they can learn propagation of those?
- connection of FC and gene expression patterns and how to prove such interaction/correlation?
- possible gene knockout targets within mouse brain and possible structural influences

References

K. Achim, J.-B. Pettit, L. R. Saraiva, D. Gavriouchkina, T. Larsson, D. Arendt, and J. C. Marioni. High-throughput spatial mapping of single-cell rna-seq data to tissue of origin. *Nature biotechnology*, 33(5):503–509, 2015.

A. Arnatkevičiūtė, B. D. Fulcher, R. Pocock, and A. Fornito. Hub connectivity, neuronal diversity, and gene expression in the caenorhabditis elegans connectome. *PLoS computational biology*, 14(2): e1005989, 2018.

O. Aromolaran, T. Beder, M. Oswald, J. Oyelade, E. Adebiyi, and R. Koenig. Essential gene prediction in drosophila melanogaster using machine learning approaches based on sequence and functional features. *Computational and structural biotechnology journal*, 18:612–621, 2020.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000. doi: 10.1038/75556. URL `https://doi.org/10.1038/75556`.

R. F. Betzel, J. D. Medaglia, A. E. Kahn, J. Soffer, D. R. Schonhaut, and D. S. Bassett. Structural, geometric and genetic factors predict interregional brain connectivity patterns probed by electrocorticography. *Nature biomedical engineering*, 3(11):902–916, 2019.

M. Blum, H.-Y. Chang, S. Chuguransky, T. Grego, S. Kandasaamy, A. Mitchell, G. Nuka, T. Paysan-Lafosse, M. Qureshi, S. Raj, L. Richardson, G. A. Salazar, L. Williams, P. Bork, A. Bridge, J. Gough, D. H. Haft, I. Letunic, A. Marchler-Bauer, H. Mi, D. A. Natale, M. Necci, C. A. Orengo, A. P. Pandurangan, C. Rivoire, C. J. A. Sigrist, I. Sillitoe, N. Thanki, P. D. Thomas, S. C. E. Tosatto, C. H. Wu, A. Bateman, and R. D. Finn. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research*, 49(D1):D344–D354, Nov. 2020. doi: 10.1093/nar/gkaa977. URL `https://doi.org/10.1093/nar/gkaa977`.

J. W. Bohland, H. Bokil, S. D. Pathak, C.-K. Lee, L. Ng, C. Lau, C. Kuan, M. Hawrylycz, and P. P. Mitra. Clustering of spatial gene expression patterns in the mouse brain and comparison with classical neuroanatomy. *Methods*, 50(2):105–112, 2010.

S. Carbon, E. Douglass, B. M. Good, D. R. Unni, N. L. Harris, C. J. Mungall, S. Basu, R. L. Chisholm, R. J. Dodson, E. Hartline, P. Fey, P. D. Thomas, L.-P. Albou, D. Ebert, M. J. Kesling, H. Mi, A. Muruganujan, X. Huang, T. Mushayahama, S. A. LaBonte, D. A. Siegele, G. Antonazzo, H. Attrill, N. H. Brown, P. Garapati, S. J. Marygold, V. Trovisco, G. dos Santos, K. Falls, C. Tabone, P. Zhou, J. L. Goodman, V. B. Strelets, J. Thurmond, P. Garmiri, R. Ishtiaq, M. Rodríguez-López, M. L. Acencio, M. Kuiper, A. Lægreid, C. Logie, R. C. Lovering, B. Kramarz, S. C. C. Saverimuttu, S. M. Pinheiro, H. Gunn, R. Su, K. E. Thurlow, M. Chibucos, M. Giglio, S. Nadendla, J. Munro, R. Jackson, M. J. Duesbury, N. Del-Toro, B. H. M. Meldal, K. Paneerselvam, L. Perfetto, P. Porras, S. Orchard, A. Shrivastava, H.-Y. Chang, R. D. Finn, A. L. Mitchell, N. D. Rawlings, L. Richardson, A. Sangrador-Vegas, J. A. Blake, K. R. Christie, M. E. Dolan, H. J. Drabkin, D. P. Hill, L. Ni, D. M. Sitnikov, M. A. Harris, S. G. Oliver, K. Rutherford, V. Wood, J. Hayles, J. Bähler, E. R. Bolton, J. L. D. Pons, M. R. Dwinell, G. T. Hayman, M. L. Kaldunski, A. E. Kwitek, S. J. F. Laulederkind, C. Plasterer, M. A. Tutaj, M. Vedi, S.-J. Wang, P. D'Eustachio, L. Matthews, J. P. Balhoff, S. A. Aleksander, M. J. Alexander, J. M. Cherry, S. R. Engel, F. Gondwe, K. Karra, S. R. Miyasato, R. S. Nash, M. Simison, M. S. Skrzypek, S. Weng, E. D. Wong, M. Feuermann, P. Gaudet, A. Morgat, E. Bakker, T. Z. Berardini, L. Reiser, S. Subramaniam, E. Huala, C. N. Arighi, A. Auchincloss, K. Axelsen, G. Argoud-Puy, A. Bateman, M.-C. Blatter, E. Boutet, E. Bowler, L. Breuza, A. Bridge, R. Britto, H. Bye-A-Jee, C. C. Casas, E. Coudert, P. Denny, A. Estreicher, M. L. Famiglietti, G. Georghiou, A. Gos, N. Gruaz-Gumowski, E. Hatton-Ellis, C. Hulo, A. Ignatchenko, F. Jungo,

K. Laiho, P. L. Mercier, D. Lieberherr, A. Lock, Y. Lussi, A. MacDougall, M. Magrane, M. J. Martin, P. Masson, D. A. Natale, N. Hyka-Nouspikel, S. Orchard, I. Pedruzzi, L. Pourcel, S. Poux, S. Pundir, C. Rivoire, E. Speretta, S. Sundaram, N. Tyagi, K. Warner, R. Zaru, C. H. Wu, A. D. Diehl, J. N. Chan, C. Grove, R. Y. N. Lee, H.-M. Muller, D. Raciti, K. V. Auken, P. W. Sternberg, M. Berriman, M. Paulini, K. Howe, S. Gao, A. Wright, L. Stein, D. G. Howe, S. Toro, M. Westerfield, P. Jaiswal, L. Cooper, and J. Elser. The gene ontology resource: enriching a GOld mine. *Nucleic Acids Research*, 49(D1):D325–D334, Dec. 2020. doi: 10.1093/nar/gkaa1113. URL `https://doi.org/10.1093/nar/gkaa1113`.

J. Chen, S. Suo, P. P. Tam, J.-D. J. Han, G. Peng, and N. Jing. Spatial transcriptomic analysis of cryosectioned tissue samples with geo-seq. *Nature protocols*, 12(3):566–580, 2017.

J. Chen, A. Althagafi, and R. Hoehndorf. Predicting candidate genes from phenotypes, functions and anatomical site of expression. *Bioinformatics*, Oct. 2020. doi: 10.1093/bioinformatics/btaa879. URL `https://doi.org/10.1093/bioinformatics/btaa879`. advance access.

M. D. Chikina, C. Huttenhower, C. T. Murphy, and O. G. Troyanskaya. Global prediction of tissue-specific gene expression and context-dependent gene networks in caenorhabditis elegans. *PLoS computational biology*, 5(6):e1000417, 2009.

T. L. Daigle, L. Madisen, T. A. Hage, M. T. Valley, U. Knoblich, R. S. Larsen, M. M. Takeno, L. Huang, H. Gu, R. Larsen, et al. A suite of transgenic driver and reporter mouse lines with enhanced brain-cell-type targeting and functionality. *Cell*, 174(2):465–480, 2018.

P. Davis, M. Zarowiecki, V. Arnaboldi, A. Becerra, S. Cain, J. Chan, W. J. Chen, J. Cho, E. da Veiga Beltrame, S. Diamantakis, et al. Wormbase in 2022—data, processes, and tools for analyzing caenorhabditis elegans. *Genetics*, 220(4):iyac003, 2022.

I. Diez and J. Sepulcre. Neurogenetic profiles delineate large-scale connectivity dynamics of the human brain. *Nature communications*, 9(1):1–10, 2018.

A. Fakhry and S. Ji. High-resolution prediction of mouse brain connectivity using gene expression patterns. *Methods*, 73:71–78, 2015.

A. Fakhry, T. Zeng, H. Peng, and S. Ji. Global analysis of gene expression and projection target correlations in the mouse brain. *Brain Informatics*, 2(2):107–117, 2015.

A. Feng, C. You, S. Wang, and L. Tassiulas. Kergnns: Interpretable graph neural networks with graph kernels. *ArXiv Preprint: https://arxiv. org/abs/2201.00491*, 2022.

M. Fey and J. E. Lenssen. Fast graph representation learning with pytorch geometric. *CoRR*, abs/1903.02428, 2019. URL `http://arxiv.org/abs/1903.02428`.

A. Flores-Morales, H. Gullberg, L. Fernandez, N. Ståhlberg, N. H. Lee, B. Vennström, and G. Norstedt. Patterns of liver gene expression governed by trβ. *Molecular endocrinology*, 16(6):1257–1268, 2002.

A. Fornito, A. Zalesky, and M. Breakspear. The connectomics of brain disorders. *Nature Reviews Neuroscience*, 16(3):159–172, 2015.

B. D. Fulcher and A. Fornito. A transcriptional signature of hub connectivity in the mouse connectome. *Proceedings of the National Academy of Sciences*, 113(5):1435–1440, 2016.

B. D. Fulcher, A. Arnatkeviciute, and A. Fornito. Overcoming false-positive gene-category enrichment in the analysis of spatially resolved transcriptomic brain atlas data. *Nature communications*, 12(1): 1–13, 2021.

J. Gillis and P. Pavlidis. “guilt by association” is the exception rather than the rule in gene networks. *PLoS Computational Biology*, 8(3):e1002444, Mar. 2012. doi: 10.1371/journal.pcbi.1002444. URL `https://doi.org/10.1371/journal.pcbi.1002444`.

P. Goel, A. Kuceyeski, E. LoCastro, and A. Raj. Spatial patterns of genome-wide expression profiles reflect anatomic and fiber connectivity architecture of healthy human brain. *Human brain mapping*, 35(8):4204–4218, 2014.

J. A. Harris, S. Mihalas, K. E. Hirokawa, J. D. Whitesell, H. Choi, A. Bernard, P. Bohn, S. Caldejon, L. Casal, A. Cho, et al. Hierarchical organization of cortical and thalamic connectivity. *Nature*, 575 (7781):195–202, 2019.

M. Hawrylycz, R. A. Baldock, A. Burger, T. Hashikawa, G. A. Johnson, M. Martone, L. Ng, C. Lau, S. D. Larsen, J. Nissanov, L. Puelles, S. Ruffins, F. Verbeek, I. Zaslavsky, and J. Boline. Digital Atlasing and Standardization in the Mouse Brain. *PLOS Computational Biology*, 7(2):e1001065, Feb. 2011. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1001065. URL `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1001065`. Publisher: Public Library of Science.

J. I. Herschkowitz, K. Simin, V. J. Weigman, I. Mikaelian, J. Usary, Z. Hu, K. E. Rasmussen, L. P. Jones, S. Assefnia, S. Chandrasekharan, et al. Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome biology*, 8(5):1–17, 2007.

T. Hinnerichs and R. Hoehndorf. Dti-voodoo: machine learning over interaction networks and ontology-based background knowledge predicts drug–target interactions. *Bioinformatics*, 37(24):4835–4843, 2021.

R. Hoehndorf, P. N. Schofield, and G. V. Gkoutos. PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Research*, 39(18):e119–e119, July 2011. doi: 10.1093/nar/gkr538. URL `https://doi.org/10.1093/nar/gkr538`.

L. A. Jeni, J. F. Cohn, and F. De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 245–251, 2013. doi: 10.1109/ACII.2013.47.

A. Kaufman, G. Dror, I. Meilijson, and E. Ruppin. Gene expression of caenorhabditis elegans neurons carries information on their synaptic connectivity. *PLoS computational biology*, 2(12):e167, 2006.

R. Ke, M. Mignardi, A. Pacureanu, J. Svedlund, J. Botling, C. Wählby, and M. Nilsson. In situ sequencing for rna analysis in preserved tissue and cells. *Nature methods*, 10(9):857–860, 2013.

S. T. Kelly and M. A. Black. graphsim: An R package for simulating gene expression data from graph structures of biological pathways. *Journal of Open Source Software*, 5(51):2161, July 2020. ISSN 2475-9066. doi: 10.21105/joss.02161. URL `https://joss.theoj.org/papers/10.21105/joss.02161`.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907, 2016. URL `http://arxiv.org/abs/1609.02907`.

M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1):W90–W97, 2016.

M. Kulmanov and R. Hoehndorf. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics*, 36(2):422–429, 07 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz595. URL `https://doi.org/10.1093/bioinformatics/btz595`.

S. H. Lecker, R. T. Jagoe, A. Gilbert, M. Gomes, V. Baracos, J. Bailey, S. R. Price, W. E. Mitch, and A. L. Goldberg. Multiple types of skeletal muscle atrophy involve a common program of changes in gene expression. *The FASEB Journal*, 18(1):39–51, 2004.

R. Y. Lee and P. W. Sternberg. Building a cell and anatomy ontology of caenorhabditis elegans. *Comparative and Functional Genomics*, 4(1):121–126, 2003.

E. S. Lein, M. J. Hawrylycz, N. Ao, M. Ayres, A. Bensinger, A. Bernard, A. F. Boe, M. S. Boguski, K. S. Brockway, E. J. Byrnes, L. Chen, L. Chen, T.-M. Chen, M. C. Chin, J. Chong, B. E. Crook, A. Czaplinska, C. N. Dang, S. Datta, N. R. Dee, A. L. Desaki, T. Desta, E. Diep, T. A. Dolbeare, M. J. Donelan, H.-W. Dong, J. G. Dougherty, B. J. Duncan, A. J. Ebbert, G. Eichele, L. K. Estin, C. Faber, B. A. Facer, R. Fields, S. R. Fischer, T. P. Fliss, C. Frensley, S. N. Gates, K. J. Glattfelder, K. R. Halverson, M. R. Hart, J. G. Hohmann, M. P. Howell, D. P. Jeung, R. A. Johnson, P. T. Karr, R. Kawal, J. M. Kidney, R. H. Knapik, C. L. Kuan, J. H. Lake, A. R. Laramee, K. D. Larsen, C. Lau, T. A. Lemon, A. J. Liang, Y. Liu, L. T. Luong, J. Michaels, J. J. Morgan, R. J. Morgan, M. T. Mortrud, N. F. Mosqueda, L. L. Ng, R. Ng, G. J. Orta, C. C. Overly, T. H. Pak, S. E. Parry, S. D. Pathak, O. C. Pearson, R. B. Puchalski, Z. L. Riley, H. R. Rockett, S. A. Rowland, J. J. Royall, M. J. Ruiz, N. R. Sarno, K. Schaffnit, N. V. Shapovalova, T. Sivisay, C. R. Slaughterbeck, S. C. Smith, K. A. Smith, B. I. Smith, A. J. Sodt, N. N. Stewart, K.-R. Stumpf, S. M. Sunkin, M. Sutram, A. Tam, C. D. Teemer, C. Thaller, C. L. Thompson, L. R. Varnam, A. Visel, R. M. Whitlock, P. E. Wohnoutka, C. K. Wolkey, V. Y. Wong, M. Wood, M. B. Yaylaoglu, R. C. Young, B. L. Youngstrom, X. F. Yuan, B. Zhang, T. A. Zwingman, and A. R. Jones. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445(7124):168–176, Dec. 2006. doi: 10.1038/nature05453. URL `https://doi.org/10.1038/nature05453`.

G. Li, M. Müller, A. Thabet, and B. Ghanem. Deepgcns: Can gcns go as deep as cnns? In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

G. Li, C. Xiong, A. Thabet, and B. Ghanem. Deepergcn: All you need to train deeper gcns. *CoRR*, abs/2006.07739, 2020.

J. Z. Li, B. G. Bunney, F. Meng, M. H. Hagenauer, D. M. Walsh, M. P. Vawter, S. J. Evans, P. V. Choudary, P. Cartagena, J. D. Barchas, et al. Circadian patterns of gene expression in the human brain and disruption in major depressive disorder. *Proceedings of the National Academy of Sciences*, 110(24):9950–9955, 2013.

R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica. Tune: A research platform for distributed model selection and training, 2018.

W.-C. Lo, C.-C. Lee, C.-Y. Lee, and P.-C. Lyu. Cpdb: a database of circular permutation in proteins. *Nucleic acids research*, 37(suppl_1):D328–D332, 2009.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013. URL `http://arxiv.org/abs/1310.4546`.

N. Milyaev, D. Osumi-Sutherland, S. Reeve, N. Burton, R. A. Baldock, and J. D. Armstrong. The virtual fly brain browser and query interface. *Bioinformatics*, 28(3):411–415, 2012.

modENCODE Consortium, S. Roy, J. Ernst, P. V. Kharchenko, P. Kheradpour, N. Negre, M. L. Eaton, J. M. Landolin, C. A. Bristow, L. Ma, et al. Identification of functional elements and regulatory circuits by drosophila modencode. *Science*, 330(6012):1787–1797, 2010.

J. R. Moffitt, J. Hao, G. Wang, K. H. Chen, H. P. Babcock, and X. Zhuang. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences*, 113(39):11046–11051, 2016.

L. Ng, S. Pathak, C. Kuan, C. Lau, H.-w. Dong, A. Sodt, C. Dang, B. Avants, P. Yushkevich, J. Gee, et al. Neuroinformatics for genome-wide 3-d gene expression mapping in the mouse brain. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(3):382–393, 2007.

W. S. Noble. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567, 2006.

S. W. Oh, J. A. Harris, L. Ng, B. Winslow, N. Cain, S. Mihalas, Q. Wang, C. Lau, L. Kuan, A. M. Henry, et al. A mesoscale connectome of the mouse brain. *Nature*, 508(7495):207–214, 2014.

S. Oliver. Guilt-by-association goes global. *Nature*, 403(6770):601–602, Feb. 2000. doi: 10.1038/35001165. URL `https://doi.org/10.1038/35001165`.

N. Pallast, M. Diedenhofen, S. Blaschke, F. Wieters, D. Wiedermann, M. Hoehn, G. R. Fink, and M. Aswendt. Processing pipeline for atlas-based imaging data analysis of structural and functional mouse brain MRI (AIDAmri). *Frontiers in Neuroinformatics*, 13, June 2019. doi: 10.3389/fninf.2019.00042. URL `https://doi.org/10.3389/fninf.2019.00042`.

L. Parkes, B. Fulcher, M. Yücel, and A. Fornito. Transcriptional signatures of connectomic subregions of the human striatum. *Genes, Brain and Behavior*, 16(7):647–663, 2017.

G. Partel, M. M. Hilscher, G. Milli, L. Solorzano, A. H. Klemm, M. Nilsson, and C. Wählby. Automated identification of the mouse brain's spatial compartments from in situ sequencing data. *BMC Biology*, 18(1), Oct. 2020. doi: 10.1186/s12915-020-00874-5. URL `https://doi.org/10.1186/s12915-020-00874-5`.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

P. Pavlidis and W. S. Noble. Analysis of strain and regional variation in gene expression in mouse brain. *Genome Biology*, 2(10):research0042.1, Sept. 2001. ISSN 1474-760X. doi: 10.1186/gb-2001-2-10-research0042. URL `https://doi.org/10.1186/gb-2001-2-10-research0042`.

A. Ramasamy, D. Trabzuni, S. Guelfi, V. Varghese, C. Smith, R. Walker, T. De, L. Coin, R. De Silva, M. R. Cookson, et al. Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature neuroscience*, 17(10):1418–1428, 2014.

C. H. Rankin. From gene to identified neuron to behaviour in caenorhabditis elegans. *Nature Reviews Genetics*, 3(8):622–630, 2002.

S. Razick, G. Magklaras, and I. M. Donaldson. irefindex: a consolidated protein interaction database with provenance. *BMC bioinformatics*, 9(1):1–19, 2008.

J. Richiardi, A. Altmann, A.-C. Milazzo, C. Chang, M. M. Chakravarty, T. Banaschewski, G. J. Barker, A. L. Bokde, U. Bromberg, C. Büchel, et al. Correlated gene expression supports synchronous activity in brain networks. *Science*, 348(6240):1241–1244, 2015.

I. Roberti, M. Lovino, S. Di Cataldo, E. Ficarra, and G. Urgese. Exploiting gene expression profiles for the automated prediction of connectivity between brain regions. *International journal of molecular sciences*, 20(8):2035, 2019.

P. Roland, C. Graufelds, J. Wăhlin, L. Ingelman, M. Andersson, A. Ledberg, J. Pedersen, S. Åkerman, A. Dabringhaus, and K. Zilles. Human brain atlas: for high-resolution functional and anatomical mapping. *Human Brain Mapping*, 1(3):173–184, 1994.

M. Rubinov, R. J. Ypma, C. Watson, and E. T. Bullmore. Wiring cost and topological participation of the mouse brain connectome. *Proceedings of the National Academy of Sciences*, 112(32):10032–10037, 2015.

R. Schulte-Sasse, S. Budach, D. Hnisz, and A. Marsico. Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. *Nature Machine Intelligence*, 3(6):513–526, 2021.

K. Sengupta, A. Gambin, S. Basu, and D. Plewczynski. Multinet: A diffusion-based approach to assign directionality in protein interactions using a consensus of eight protein interaction datasets. In *Proceedings of International Conference on Frontiers in Computing and Systems*, pages 13–20. Springer, 2023.

C. L. Smith and J. T. Eppig. The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 1(3):390–399, Nov. 2009. doi: 10.1002/wsbm.44. URL `https://doi.org/10.1002/wsbm.44`.

A. Subramanian, H. Kuehn, J. Gould, P. Tamayo, and J. P. Mesirov. Gsea-p: a desktop application for gene set enrichment analysis. *Bioinformatics*, 23(23):3251–3253, 2007.

H. Sun and O. Hobert. Temporal transitions in the post-mitotic nervous system of caenorhabditis elegans. *Nature*, 600(7887):93–99, 2021.

D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, Peer, L. J. Jensen, and C. von Mering. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1): D447–D452, Oct. 2014. doi: 10.1093/nar/gku1003. URL `https://doi.org/10.1093/nar/gku1003`.

N. Takata, N. Sato, Y. Komaki, H. Okano, and K. F. Tanaka. Flexible annotation atlas of the mouse brain: combining and dividing brain structures of the Allen Brain Atlas while maintaining anatomical hierarchy. *Scientific Reports*, 11(1):6234, Mar. 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-85807-0. URL `https://www.nature.com/articles/s41598-021-85807-0`. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Brain;Functional magnetic resonance imaging;Neuroscience Subject_term_id: brain;functional-magnetic-resonance-imaging;neuroscience.

K. Tomczak, P. Czerwińska, and M. Wiznerowicz. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015.

M. Trebacz, Z. Shams, M. Jamnik, P. Scherer, N. Simidjievski, H. A. Terre, and P. Liò. Using ontology embeddings for structural inductive bias in gene expression data analysis. *CoRR*, abs/2011.10998, 2020.

N. A. Twine, K. Janitz, M. R. Wilkins, and M. Janitz. Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by alzheimer's disease. *PloS one*, 6(1): e16266, 2011.

S. L. Valk, T. Xu, D. S. Margulies, S. K. Masouleh, C. Paquola, A. Goulas, P. Kochunov, J. Smallwood, B. T. T. Yeo, B. C. Bernhardt, and S. B. Eickhoff. Shaping brain structure: Genetic and phylogenetic axes of macroscale organization of cortical thickness. *Science Advances*, 6(39):eabb3417, 2020. doi: 10.1126/sciadv.abb3417. URL `https://www.science.org/doi/abs/10.1126/sciadv.abb3417`.

D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

V. Varadan, D. M. Miller III, and D. Anastassiou. Computational inference of the molecular logic for synaptic connectivity in c. elegans. *Bioinformatics*, 22(14):e497–e506, 2006.

A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani. Global protein function prediction from protein-protein interaction networks. *Nature Biotechnology*, 21(6):697–700, May 2003. doi: 10.1038/nbt825. URL `https://doi.org/10.1038/nbt825`.

P. E. Vértes, T. Rittman, K. J. Whitaker, R. Romero-Garcia, F. Váša, M. G. Kitzbichler, K. Wagstyl, P. Fonagy, R. J. Dolan, P. B. Jones, et al. Gene transcription profiles associated with intermodular hubs and connection distance in human functional magnetic resonance imaging networks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1705):20150362, 2016.

Q. Wang, S.-L. Ding, Y. Li, J. Royall, D. Feng, P. Lesnar, N. Graddis, M. Naeemi, B. Facer, A. Ho, et al. The allen mouse brain common coordinate framework: a 3d reference atlas. *Cell*, 181(4): 936–953, 2020.

T. Wang, W. Shao, Z. Huang, H. Tang, J. Zhang, Z. Ding, and K. Huang. Mogonet integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nature Communications*, 12(1):1–13, 2021.

W. Wang, R. Han, M. Zhang, Y. Wang, T. Wang, Y. Wang, X. Shang, and J. Peng. A network-based method for brain disease gene prediction by integrating brain connectome and molecular network. *Briefings in Bioinformatics*, 23(1):bbab459, 2022.

J. D. Watson. The human genome project: past, present, and future. *Science*, 248(4951):44–49, 1990.

C. W. Whitfield, A.-M. Cziko, and G. E. Robinson. Gene expression profiles in the brain predict behavior in individual honey bees. *Science*, 302(5643):296–299, 2003.

Y. Yang, Q. Fang, and H.-B. Shen. Predicting gene regulatory interactions based on spatial gene expression data and deep learning. *PLoS computational biology*, 15(9):e1007324, 2019.

M. A. Zapala, I. Hovatta, J. A. Ellison, L. Wodicka, J. A. Del Rio, R. Tennant, W. Tynan, R. S. Broide, R. Helton, B. S. Stoveken, et al. Adult mouse brain gene expression patterns bear an embryologic imprint. *Proceedings of the National Academy of Sciences*, 102(29):10357–10362, 2005.

T. Zeng, R. Li, R. Mukkamala, J. Ye, and S. Ji. Deep convolutional neural networks for annotating gene expression patterns in the mouse brain. *BMC Bioinformatics*, 16(1):147, May 2015. ISSN 1471-2105. doi: 10.1186/s12859-015-0553-9. URL `https://doi.org/10.1186/s12859-015-0553-9`.

V. Zerbi, M. Pagani, M. Markicevic, M. Matteoli, D. Pozzi, M. Fagiolini, Y. Bozzi, A. Galbusera, M. L. Scattoni, G. Provenzano, A. Banerjee, F. Helmchen, M. A. Basson, J. Ellegood, J. P. Lerch, M. Rudin, A. Gozzi, and N. Wenderoth. Brain mapping across 16 autism mouse models reveals a spectrum of functional connectivity subtypes. *Molecular Psychiatry*, Aug. 2021. doi: 10.1038/s41380-021-01245-4. URL `https://doi.org/10.1038/s41380-021-01245-4`.

M. Zitnik and J. Leskovec. Predicting multicellular function through multi-layer tissue networks. *CoRR*, abs/1707.04638, 2017. URL `http://arxiv.org/abs/1707.04638`.