Diploma thesis

TOWARDS GENE EXPRESSION PREDICTION IN MOUSE BRAINS

September 30, 2021

Tilman Hinnerichs
Matrikelnummer: 4643427
Technische Universität Dresden

Tutor: Dr. Nico Scherf
MPI for CBS

Summer semester 2021

proper title?

Abstract

Contents

To include in some chapter

Predict gene expression per section/structure:
Take region as input and predict gene expression
Challenges:

- how to normalize expression intensity (see discussion in DeepMOCCA paper, Sara Alghamdi), as there are regions with much more activity than others (e.g. bone narrow vs. bone boarder); thresholds for intensity varies across genes
  - over all intensities →
  - per structure →
  - per gene →

    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).
- transfer learning working for other structure/regions
- dataset: Allen Mouse brain atlas vs.
  - phenoview impc data
  - mousephenotype
  - HPO/MP project expression data
- structure specific features?
  - structural ontology / closeness
  - developmental hierarchy of tissue
- how descriptive are the following for gene expression? How to encode structural properties
  - molecular protein function (DeepGOPlus)
  - phenotypical features
  - fix-point stuff
- predict knock-out of genes, does this relate over
  - functional graph [Valk et al., 2020]
  - structural ontology?
  - developmental ontology

Compare the following embeddings:

1. tSNE
2. UMAP
3. Parametric UMAP
4. Parametric UMAP with custom graph
5. Parametric Umap with custom graph and GNN in network

To be investigated

- predict against axonal/structural connectivity
- predict against functional connectivity

Possible hypotheses

- predict gene expression for a given single structure
- (predict structure from gene expression pattern)
- (predict structure form gene expression and image)
- (predict cancer type from morphology/pathologic image of cancer)
- simulate loss of function/expression by removing one node of graph/add remove gene expression
  - evaluate against what data? see other section
- see influence on **structural** connectivity
  - evaluate against what data?
- see influence on **functional** connectivity
  - evaluate against what data? [Zerbi et al., 2021]
  - transgenic data from AMBA?

Ideas for page-filling plots

- show distribution (histo, mean, median, boxplot?) of expression densities see 'get_ge_structure_mat'
- show predictability of gene expression across over multiple structure vs. distance over onto?

1 Introduction

General thread for introduction and motivation:

- Gene expression patterns are difficult to analyze in humans → take mouse as model organisms
- no in-depth analysis of mouse brain genetic patterns and their relation to different connectivity patterns has been made yet
- we analyze

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

Brief Overview of Theoreticla Foundations Utilized in the study

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

2.1 To be searched

- read across citations of DeepMOCCA/Takata et al. [2021]
- find other papers on
  - 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

2.2 To be sorted somewhere

- Variability and different interpretations of different graph convolutional neural filters [Kipf and Welling, 2016, Li et al., 2020, Hamilton et al., 2017] etc.
- Guilt by association over gene networks [Oliver, 2000, Gillis and Pavlidis, 2012]
- protein function prediction from PPI networks [Vazquez et al., 2003]
- DeepGOPlus for feature generation [Kulmanov and Hoehndorf, 2019]
- discussion of DeepMocca by Sara [Althubaiti et al., 2021]
- 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
- RayTune[Liaw et al., 2018] for automated hyperparameter tuning

2.3 Spatial patterns of gene expression

Data discussion, hypotheses and traditional approaches:

- [noa]
- Possible effects of rabies virus on gene expression[Prosniak et al., 2001] for potential knockout targets
- Review paper on regional variation in gene expression in mouse brain [Pavlidis and Noble, 2001]

Modern approaches on learning from gene expression patterns in mouse brain:

- Deep learning methods for capturing spatiality w.r.t. gene expression withing the brain [Zeng et al., 2015]
- R package for simulating gene expression from graph structures over general biological pathways [Kelly and Black, 2020]

Read this

2.4 Gene knockout and pathology hypotheses

- Gene expression for different kinds of stress within mouse brain [Flati et al., 2020]

2.5 Structural connectivity

- experimental setup from Allen Institute for axonal projection data
- paraphrase description of „Technical tour: Explore the Allen Mouse Brain Connectivity Atlas“

2.6 Functional connectivity

- Where is data coming from? [Pallast et al., 2019]
- How to calculate functional connectivity matrix → AIDAconnect (no paper yet? cite dataset?
- How to combine functional connectivity for multiple samples?

3 Methods

Introduction and general description, study method and study design

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

- why were these datasets used and not others?
- How did we achieve the matching?
- what are premises of the dataset?

Explanation of Sample used in the study

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 oand validity of internal/external design and related subtypes

Assumptions of study method and study design with implied

3.1 Problem description

3.2 Datasets

- Allen mouse brain atlas [Lein et al., 2006]
- STRING for PPI network and how we chose suitable interactions [Szklarczyk et al., 2014]

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 [Pallast et al., 2019]
- functional connectivity data from [Pallast et al., 2019]

3.3 Model

3.3.1 Feature generation

Data preparation for regression task

- unbalanced data for regression → k-means and weighting w.r.t. bin

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]:

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

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.

References

Clustering of spatial gene expression patterns in the mouse brain and comparison with classical neuroanatomy - ScienceDirect. URL `https://www.sciencedirect.com/science/article/abs/pii/S1046202309002035?casa_token=1ZHnepKbbpgAAAAA:1S2m5q8_x_uYFTeJivzAG59F4WNz6HgkJEDxuBGq2X3FguxdTIshHxlP7Nxz56GazM05bVkEpSk`.

S. Althubaiti, M. Kulmanov, Y. Liu, G. V. Gkoutos, P. Schofield, and R. Hoehndorf. Deepmocca: A pan-cancer prognostic model identifies personalized prognostic markers through graph attention and multi-omics data integration. *bioRxiv*, 2021. doi: 10.1101/2021.03.02.433454. URL `https://www.biorxiv.org/content/early/2021/03/02/2021.03.02.433454`.

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`.

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`.

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

T. Flati, S. Gioiosa, G. Chillemi, A. Mele, A. Oliverio, C. Mannironi, A. Rinaldi, and T. Castrignanò. A gene expression atlas for different kinds of stress in the mouse brain. *Scientific Data*, 7(1), Dec. 2020. doi: 10.1038/s41597-020-00772-z. URL https://doi.org/10.1038/s41597-020-00772-z.

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.

W. L. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. In *NIPS*, 2017.

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.

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.

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. 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.

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.

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.

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`.

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`.

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`.

M. Prosniak, D. C. Hooper, B. Dietzschold, and H. Koprowski. Effect of rabies virus infection on gene expression in mouse brain. *Proceedings of the National Academy of Sciences*, 98(5):2758–2763, Feb. 2001. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.051630298. URL `https://www.pnas.org/content/98/5/2758`. Publisher: National Academy of Sciences Section: Biological Sciences.

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`.

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.

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.

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`.

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`.

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`.