Diss. ETH no. 27623

Connectome++: Microtubule Reconstruction and Neurotransmitter Prediction from Large Scale Electron Microscopy Volumes

A thesis submitted to attain the degree of
Doctor of Sciences of ETH Zurich
(Dr. sc. ETH Zurich)

presented by
Nils Eckstein
MSc. ETH Zurich
born on 25.03.1992
citizen of Germany

accepted on the recommendation of

Prof. Dr. Richard Hahnloser
Dr. Matthew Cook
Dr. Jan Funke
Prof. Dr. Albert Cardona

2021

Contents

List of Figures

List of Tables

Abstract

This work builds upon the recent advances in the automated data analysis of terabyte scale electron microscopy (EM) volumes of brains for Connectome generation and extends the amount of information that can be extracted from EM datasets. To this end we present three novel methods for EM data analysis that enable the automatic reconstruction of microtubules, allow for semi-automated discovery of so far unknown phenotypic features, and that are able to extract neurotransmitter type of synapses directly from EM data.

The first chapter of this thesis presents a method for microtubule tracking in electron microscopy volumes. Our method first identifies a sparse set of voxels that likely belong to microtubules. Similar to prior work, we then enumerate potential edges between these voxels, which we represent in a candidate graph. Tracks of microtubules are found by selecting nodes and edges in the candidate graph by solving a constrained optimization problem incorporating biological priors on microtubule structure. For this, we present a novel integer linear programming formulation, which results in speed-ups of three orders of magnitude and an increase of 53% in accuracy compared to prior art (evaluated on three 1.2 x 4 x 4μm volumes of *Drosophila melanogaster* neural tissue). We also propose a scheme to solve the optimization problem in a block-wise fashion, which allows distributed tracking and is necessary to process very large electron microscopy volumes.

In the second part of the thesis, we propose a method for neural network interpretability by combining feature attribution with counterfactual explanations to generate attribution maps that highlight the most discriminative features between pairs of classes. We show that this method can be used to quantitatively evaluate the performance of feature attribution methods in an objective manner, thus preventing potential observer bias. We evaluate the proposed method on three diverse datasets, including a challenging artificial dataset and real-world biological data. We show quantitatively and qualitatively that the highlighted features are substantially more discriminative than those extracted using conventional attribution methods and argue that this type of explanation is better suited for understanding fine grained class differences as learned by a deep neural network.

The last chapter shows that in *Drosophila melanogaster* artificial convolutional neural networks can confidently predict the type of neurotransmitter released at a synaptic site from EM images alone. The network successfully discriminates between six types of neurotransmitters (GABA, glutamate, acetylcholine, serotonin, dopamine, and octopamine) with an average accuracy of 87% for individual synapses and 94% for entire neurons, assuming each neuron expresses only one neurotransmitter. This result is surprising

as there are often no obvious cues in the EM images that human observers can use to predict neurotransmitter identity. We show that the classifier generalizes across brain regions, neurons and datasets. We predict all automatically detected synapses in a whole-brain EM dataset and analyse global neurotransmitter distribution in the *Drosophila melanogaster* brain. Furthermore, we use the method presented in chapter three to identify a set of hypothetical discriminators that could be used by humans to distinguish the different neurotransmitter phenotypes by eye and potentially reveal the relationship between structure and function of synapses. Finally, we apply the proposed method to predict the neurotransmitter identity of all neurons of 89 hemilineages. We show that in contrast to the Ventral Nervous System (VNS) our predictions are inconsistent with the hypothesis that all neurons within a hemilineage express the same fast-acting neurotransmitter in the brain of *Drosophila melanogaster*.

Zusammenfassung

Diese Arbeit baut auf den Fortschritten in der automatischen Datenanalyse von Terabyte grossen Elektronen Mikroskopie (EM) Datensätzen für Connectomics auf und erweitert die Menge an Informationen, die aus EM Datensätzen extrahiert werden können. Wir präsentieren drei neue Methoden für EM Datenanalyse, die es erlauben automatisch Mikrotubuli zu rekonstruieren, semi-automatische Entdeckungen von bis dahin unbekannten Phänotypen zu machen und Neurotransmitter Identität zu bestimmen.

Das erste Kapitel dieser Arbeit präsentiert eine Methode für Mikrotubuli Rekonstruktion in EM Volumen. Unsere Methode identifiziert erst eine Menge von Voxeln, die wahrscheinlich zu Mikrotubuli gehören. Ähnlich zu früheren Arbeiten identifizieren wir mögliche Kanten zwischen dieser Menge von Voxeln und repräsentieren sie in einem Kandidaten-Graph. Dann finden wir Mikrotubuli in dem Kandidaten-Graph, indem wir ein Optimisierungsproblem mit Nebenbedingungen unter Berücksichtigung von biologischem Vorwissen lösen. Dafür präsentieren wir eine neue ganzzahlige lineare Optimisierungs (GLO) Formulation, die den Algorithmus drei Grössenordnungen schneller macht und zu einer Verbesserung von 53% Genauigkeit führt (ausgewertet auf 1.2 x 4 x 4 μm Volumen von *Drosophila melanogaster* neuronalem Gewebe). Weiterhin schlagen wir ein Schema vor um die GLO in Blöcken zu lösen, was verteilte Mikrotubuli Rekonstruktion erlaubt, und notwendig ist um sehr grosse EM Volumen zu prozessieren.

In dem zweiten Teil der Arbeit schlagen wir eine neue Methode vor um zu interpretieren was tiefe künstliche neuronale Netzwerke gelernt haben, indem wir Merkmal-Zuweisungserklärungen mit kontrafaktischen Erklärungen verbinden um Attributionskarten zu generieren, die die stärksten diskriminierenden Eigenschaften eines Klassenpaares zeigen. Wir demonstrieren, dass diese Methode benuzt werden kann, um die Genauigkeit von Merkmal-Zuweisungserklärungen objektiv zu vergleichen und so hilft Beobachter-Bias zu vermeiden. Wir evaluieren die vorgeschlagene Methode auf drei diversen Datensätzen, einschliesslich ein herausfordernder künstlicher Datensatz und ein biologischer Datensatz. Wir zeigen quantitativ und qualitativ, dass die so extrahierten Merkmale substanziell mehr diskriminierend sind als herkömmliche Merkmal-Zuweisungen und argumentieren, dass diese Art von Erklärungen besser geeignet ist um zu verstehen, wie ein künstliches neuronales Netzwerk feine Klassenunterschiede wahrnimmt.

Das letzte Kapitel zeigt, dass in *Drosophila melanogaster* künstliche, faltende, neuronale Netzwerke den Neurotransmitter Typ einer Synapse von EM Bildern dieser Synapse bestimmen können. Das Netzwerk ist in der Lage zwischen sechs verschieden Neurotransmittern (GABA, glutamate, acetylcholine, sero-

tonin, dopamine und octopamine) mit einer durchschnittlichen Genauigkeit von 87% für individuelle Synapsen und 94% für ganze Neuronen zu unterscheiden. Dies gilt unter der Annahme, dass jedes Neuron nur einen Neurotransmitter freisetzt. Dieses Resultat ist überraschend weil es oft keine auffälligen Merkmale in den EM Bildern gibt, die menschliche Beobachter benutzen können um die Neurotransmitter Identität zu bestimmen. Wir zeigen, dass der Klassifikator in der Lage ist die Neurotransmitter Identität in verschieden Gehirnregionen, Neuronen und Datensätzen zu ermitteln. Weiterhin bestimmen wir alle automatisch identifizierten Synapsen in dem *FAFB* Datensatz und analysieren die globale Neurotransmitter Verteilung in dem *Drosophila melanogaster* Gehirn. Ausserdem benutzen wir die in Kapitel drei vorgeschlagene Methode, um hypothetische, diskriminerende Faktoren zwischen den Neurotransmittern zu bestimmen. Wir wenden die Methode auch an um alle Neurotransmitter in 89 *Hemi-Abstammungen* zu bestimmen und zeigen, dass im Unterschied zum Strickleiternervensystem unsere Vorhersagen inkonsistent sind mit der Hypothese, dass alle Neuronen innerhalb einer *Hemi-Abstammungslinie* den selben, klassischen Neurotransmitter im Gehirn von *Drosophila melanogaster* freisetzen.

Acknowledgments

This thesis would not exist without the help and guidance of many people. First I would like to express my deepest gratitude towards my advisor and friend Jan Funke, who had the biggest impact on this work and me during this entire PhD. I am very thankful for the many opportunities Jan gave me and the countless scientific advices I received. I could not have wished for anything better. Thank you Jan! I also would like to thank Matthew Cook who taught me a lot about how to live in science without loosing your mind. Matt, I find you very inspiring and I really hope one day I also find the secret sauce that enables the playful and free approach to science that I feel you have. Furthermore a big thank you to the rest of my committee, Richard Hahnloser who enabled that I could work with Jan and Matt and Albert Cardona who inspired the work in the first chapter of this thesis.

Thanks also to all my fellow PhD students who made this experience so special: Julia Buhmann is actually not a Buhmann and really quite nice, Moritz Milde is also pretty cool and meeting both of you might just be the best thing about this entire thesis. A capital thank you also goes out to the entirety of the lab: Arlo who made Janelia oh so sweet, Caroline and Will who are the coding wizards everyone needs in their life and Steffen whose expertise and joy are a blessing to have around. Furthermore I would like to thank all the other people in Janelia who welcomed me so warmly there. Just to name a few who made my life particularly enjoyable: Chris, Claire, Jonny, Igor, John, Larissa, Stephan, Matt and Amelia, thank you!

Next on this list are my parents, without whom I would definitely not be writing this thesis and instead still be stuck in the ether, awaiting my turn. Not only did they conceive me, but with their love and unconditional support, they also gave me all the tools I needed to do what I want in life. Thank you both very much! Thank you Emi & Opa for building this incredible family and showing my younger self the world and all its colors, a kiss to all the Eckstein kings & queens! Thank you Oma Trauti for your love and taking care of me when I was still too young to do so myself.

A massive thanks to all you lovely Zürich people: Nico as my primary co-pilot during the past 22 years or so, Julia for teaching me how to be cool, Moritz for keeping the rain out and the people together, Amelie & Luise for the hottest flat in Züri-Wescht, Alex for your genuine interest in all of it, Philipp & Fabian for companionship among the quarks and stars, Alex & Jenny for the coolest couple in town, Pascal for never being too excited and Moritz (the other one) for being who you are. I probably forgot a bunch of people, but know I love you all. Another thundering thank you goes out to the serious people from M-town: Yannik, Philip, Michael and Dominic.

Last and very much first, thank you Dr. Bobcat. You have been with me from the beginning of this PhD to the very end and I could not have asked for a better, more loving partner in crime than you have been. We sailed this ship over an entire ocean and shared too many amazing stories to write them all down here. I am beyond thankful for all of your support and for who you are, every day.

Chapter 1

Introduction

The goal of neuroscience is to understand brains and more generally nervous systems. The brain is an important part of every animal and understanding it would allow us to understand animal (including human) behavior, learning and memory. This would enable us to design better drugs for treating cognitive illnesses and potentially reverse engineer intelligence for artificial agents. Since its birth, neuroscience developed from identifying the brain as the center of cognition (Breitenfeld et al., 2014), to studying the effect of large brain areas via lesions (Pearce, 2009), to finally discovering and studying single nerve cells: Neurons (Guillery, 2005). Subsequently the neuron was deemed the *fundamental computing unit* of the brain and an understanding emerged, that neurons communicate with each other via electro-chemical signals using synapses as release sites (Guillery, 2005). Since then the understanding of molecular mechanisms, development, physiology, neuroanatomy and many more subdisciplines of neuroscience has become increasingly precise. The view that the brain is an information processing unit, perhaps influenced by the advent of computers, is arguably the dominant interpretation of what brains are doing today. With this perspective comes an understanding that the functionally important parts of brains are circuits, comprised of neurons and connected by synapses. While the molecular mechanisms of individual neurons are still not fully understood there is hope, fueled by evidence from model organisms such as *Drosophila melanogaster*, that we can understand brain function on the circuit level of abstraction (e.g. Takemura et al. (2013, 2015, 2017); Eschbach et al. (2020); Li et al. (2020); Ohyama et al. (2015); Helmstaedter et al. (2013); Eichler et al. (2017)).

A necessary condition for understanding circuits is mapping them and generating a so-called connectome. With the advent of high resolution, high throughput, electron microscopes (EM) it became possible to image entire

brains of small model organisms such as *C. elegans* and *Drosophila melanogaster* at the required resolution to see individual neurons and synapses (White et al., 1986; Zheng et al., 2018; Xu et al., 2020; Witvliet et al., 2020). However, even for the comparatively small brain of *Drosophila melanogaster* the amount of data generated from an entire brain is too large to manually analyze (Zheng et al., 2018) and as a result a breadth of techniques for the automatic detection of synapses and segmentation of neurons have been developed in recent years (Funke et al., 2018; Januszewski et al., 2018; Lee et al., 2019; Sheridan et al., 2021; Kreshuk et al., 2015; Staffler et al., 2017; Buhmann et al., 2019). With them we now have access to the first partial connectomes of the adult *Drosophila melanogaster* central brain (Xu et al., 2020). However, generating a connectome is still not a streamlined process. In addition to imaging times of many months for *Drosophila melanogaster* the accuracy of computational approaches for detecting synapses and segmenting neurons is not yet sufficient for immediate use. Instead, every neuron and its synaptic connections has to be manually proofread by human experts, which is expensive, slow and the current bottleneck in connectome acquisition. For example, the recently completed *Hemibrain* dataset required 50 person-years of proofreading over a time span of two years (Xu et al., 2020). Additionally, the connectome alone is not sufficient to decode circuit function because crucial information is missing from the connectivity matrix. For example in *Drosophila melanogaster* it is not immediately obvious from the EM data whether a given synapse is excitatory (has a positive sign) or inhibitory (has a negative sign) or how strong it is (Xu et al., 2020; Barnes et al., 2020).

To overcome these issues, this thesis builds on the work that went into developing the current techniques to generate connectomes from EM data and looks beyond synapse detection and neuron segmentation. We present methods that exploit the vast amount of information that is stored in electron microscopy datasets but has received comparably less attention so far. In an effort to augment the connectomes from today, here we present a method to cheaply add neurotransmitter data and thus the sign of any synapse to connectomes, develop a method to understand how synaptic phenotype relates to neurotransmitter identity and propose a method for tracing microtubules in cells with the potential to speed up connectome proofreading and provide insight into cell biological processes: Connectome++.

1.1 Connectomics

Connectomics is the construction and study of the cellular wiring diagram of nervous systems. For that, the brain of a model organism is extracted, fixated, imaged slice by slice and aligned to create a 3D image volume of the

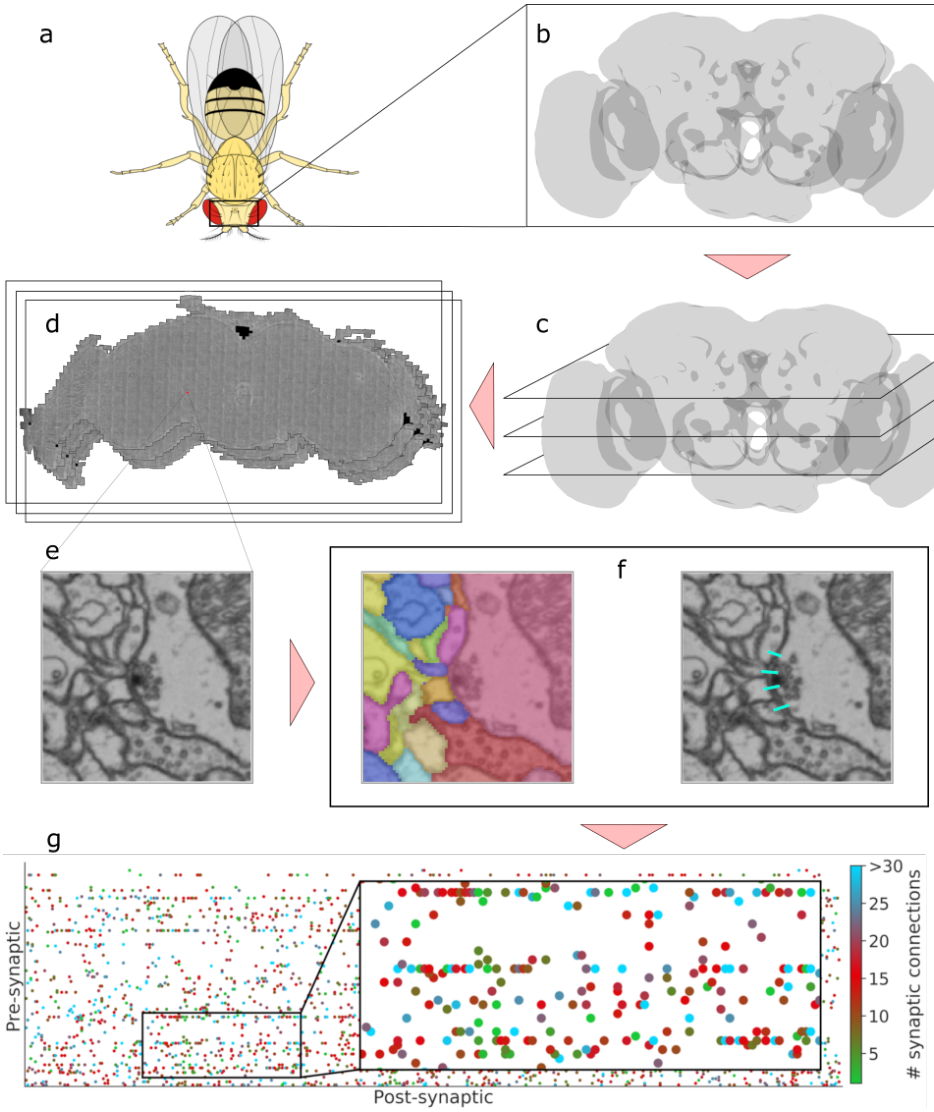


Figure 1.1: Overview of a connectomics pipeline, from the model organism to the connectivity matrix. **a**) Illustration of *Drosophila melanogaster*. For generating a *Drosophila melanogaster* connectome, its brain is **b**) dissected, fixed, stained with heavy metals, and **c**) imaged one nanometer-thick slice at a time with an electron microscope to produce a **d**) 3D electron microscopy volume of the *Drosophila melanogaster* brain. **e**) From the raw EM data each neuron is segmented and **f**) synaptic partners are detected, to generate **g**) a connectivity matrix, a connectome. The connectivity matrix shows the pre-synaptic neuron on the y-axis and post-synaptic partners on the x-axis. Number of synaptic connections between two neurons are indicated by color. Segmentation results shown are taken from *FlyWire* (Dorkenwald et al., 2020). Connectivity matrix adapted from Buhmann et al. (2019).

brain (see Xu et al. (2020) for a recent protocol). Subsequently all neurons and synapses are identified to create a map of the nervous system which can then be analyzed to elucidate structure and function of neural circuits (see Fig. 1.1 for an overview). The first complete Connectome ever generated was of the worm *C. elegans* (White et al., 1986) and it is used to this day to inform research of the worms nervous system (Larson et al., 2018). During the past few years the model organism of choice in large scale, cellular resolution connectomics, and the primary focus of this thesis, has been *Drosophila melanogaster*. This is largely because of its relatively small size, fast reproduction cycles, availability of genetic tools, comparatively easy culturing and its surprisingly complex behaviors e.g. during courtship and olfactory, associative learning tasks (Jennings, 2011; McKellar and Wyttenbach, 2017). However, it is an insect and possesses a different brain anatomy than humans and mammals. It lacks a cerebral cortex, arguably the most interesting part of the brain in mammals as it seems to enable higher order cognition and instead has a structure called the mushroom body, which enables complex associative learning (Heisenberg, 2003; Menzel, 2012; Waddell, 2013; Owald and Waddell, 2015). In addition *Drosophila melanogaster* has some peculiar cellular features, for example it has so-called polyadic synapses (also known as ribbon synapses and present in the retina of mammals) that connect one pre-synaptic site to multiple receiving neurons (Meinertzhagen and O'neil, 1991), synapse sizes are very homogeneous, and it lacks myelinated axons (Xu et al., 2020). Due to the small size of synaptic clefts, which are around 200 nm in thickness (Palay, 1956), creating a connectome requires the usage of high throughput, high resolution microscopy techniques. Currently the only microscopy technique that is able to consistently reach this resolution is electron microscopy. However, other approaches such as expansion microscopy show promising first results and may be able to augment or replace electron microscopy pipelines in the future (Shen et al., 2020).

1.1.1 Electron Microscopy

For imaging entire brains, high resolution alone is not sufficient. Because of the large size of these volumes, the microscopes need to be fast. Even with throughputs of $\sim$50 MPix/s (Zheng et al., 2018), the acquisition process can take many months. For example, the first full adult fly brain EM dataset (*FAFB*) has a total volume of $8 \times 10^7 \mu\text{m}^3$ and took 16 months to acquire (Zheng et al., 2018), and fault tolerance over these timelines needs to be guaranteed. This is achieved by modifying commercial EM systems to operate on these time and spatial scales by automating commonly manual tasks such as section collection and sample placement (Zheng et al., 2018; Xu et al., 2020; Hayworth et al., 2006). Electron microscopy technologies currently used in connectomics can be broadly distinguished in two categories: 1. (Serial section) transmission electron microscopy (ssTEM) (Williams and

Carter, 1996) and 2. (Focused ion beam) scanning electron microscopy (FIB-SEM) (Von Ardenne, 1938). ssTEMs operate by cutting a large, three dimensional, stained tissue sample into thin slices ($\geq$ 40nm), which are then imaged one by one by shooting high energy electrons through the sample and detecting where and how many electrons pass through the sample (Williams and Carter, 1996; Knott et al., 2008). Locations where many electrons arrive are unstained, such as the cytosol, while locations where few electrons arrive are stained, thus block electrons and appear dark in the final image, e.g. cell membranes. Each imaged slice is subsequently registered, aligned (Saalfeld et al., 2012), and normalized to generate the final volumetric electron microscopy dataset (Hanslovsky et al., 2015, 2017). FIB-SEM technology utilizes a different electron microscopy and sectioning approach. Instead of cutting individual sections and detecting electrons that pass through one section, (FIB-)SEM shoots electrons on the surface of a sample and detects scattered electrons instead (Von Ardenne, 1938). This has the advantage that no slices need to be cut and instead subsequent tissue surfaces are revealed and imaged by removing thin sample layers via a focused ion beam. This leads to significantly increased resolution ($\geq$ 2nm) in the axis perpendicular to the cutting plane (z-axis), which enables isotropic resolutions of up to 8 nm for large volumes such as the *Drosophila melanogaster* brain (Xu et al., 2020). The first and still only full *Drosophila melanogaster* brain electron microscopy dataset available (*FAFB*), has been acquired via ssTEM. However, since the acquisition of the *FAFB* dataset with ssTEM technology, FIB-SEM technology has seen increasing popularity for current efforts after overcoming its limited field of view via blockwise processing (Hayworth et al., 2015; Kornfeld and Denk, 2018; Xu et al., 2020). This is partially motivated by the insight that automatic neuron segmentation performs better on isotropic datasets (Xu et al., 2020). In contrast to light microscopy, where genetic tools and optical sensors allow for multi channel recordings that can be used to distinguish multiple objects, for example neurons (Hampel et al., 2011), electron microscopy does not currently offer similar capabilities at high label densities. As a result, to generate a connectome each neuron has to be segmented from the raw electron microscopy images.

1.1.2 Neuron Reconstruction

A neuron is defined as the connected region separated from other neurons by its cellular membrane. In EM the neuron membrane appears as dark outlines around the brighter cytosol (see Fig.1.1 e for an example) and the task in neuron reconstruction is to identify each individual cell in the volume. Neurons are large, complex objects that can extend over hundreds of microns, innervate multiple brain areas and contain many branch points. Combined with sometimes noisy data, misalignment of consecutive sections or entirely missing sections this makes neuron reconstruction a challeng-

ing task even for human experts. However, spurred by advancements in computer vision and in particular deep learning, recent years have seen major advances in the accuracy of automatic neuron segmentation pipelines (see Sheridan et al. (2021) for a recent comparative study). For large brain areas the current approach is thus to first generate an automatic neuron segmentation which is subsequently proofread by human experts to fix so-called split and merge errors. A split error occurs when a single neuron is wrongly assigned to two or more distinct segments, and a merge error occurs if two distinct neurons are assigned to the same segment. Both of these errors are potentially catastrophic for circuit analysis. Reducing these errors during automatic segmentation and simplifying the correction of these errors by human proofreaders has thus been a primary research focus during the past years. Early algorithms for automatic neuron segmentation worked by segmenting the cellular membrane, performing boundary detection of each cell and subsequent segmentation via a variant of the watershed algorithm (Ciresan et al., 2012; Liu et al., 2012). However, with this approach small prediction errors in the boundary can cause large morphological changes in the final neuron segmentation. This is undesirable and later algorithms try to remedy this by using structured loss functions to penalize morphological errors (Turaga et al., 2009; Funke et al., 2018; Beier et al., 2017) or by directly predicting some volumetric representation of the neuron such as flood filling networks (Januszewski et al., 2018) or local shape descriptors (Sheridan et al., 2021). In addition, software tools for correcting split and merge errors became increasingly advanced and together the time for generating a neuron segmentation for large brain areas has improved by 10-100x over the last years, enabling the generation of accurate large scale neuron segmentations (Zhao et al., 2018; Berning et al., 2015).

1.1.3 Synapse Detection

In addition to neuron segmentation, generating a connectome requires the identification of synaptic partners. A synapse is an electro-chemical contact between neurons used for signal communication among neurons. In electron microscopy volumes, synapses have a characteristic appearance, the most prominent being the so-called synaptic cleft, which appears as a dark, connected region between two synapses. In addition most synapses feature prominent pre-synaptic densities, called T-Bars as they look like the letter T (Meinertzhagen and O'neil, 1991). The pre-synaptic site, which is the region in the pre-synaptic neuron close to the cleft, also often features a number of different neurotransmitter containing vesicles for release into the synaptic cleft (Palay, 1956; Xu et al., 2020). Similar to neuron segmentation, manual detection of all synapses in brain sized EM volumes is not feasible and automatic approaches are used to identify synaptic partners and detect synaptic clefts. For that, current approaches rely mostly on deep neural net-

works and have been used to identify all synaptic partners in the *FAFB* and the *Hemibrain* dataset (Kreshuk et al., 2015; Dorkenwald et al., 2017; Staffler et al., 2017; Buhmann et al., 2019).

1.1.4 Proofreading Bottleneck

The limiting factor for generating connectomes is proof-reading of automatic neuron segmentations and synapse detection. The running time of these algorithms is on the order of days on appropriate hardware, while identifying and correcting errors takes on the order of months to years (Xu et al., 2020). As such, improving proof reading times is essential for the future of connectomics, in particular when moving to larger organisms such as the mouse. There are multiple options for reducing proof-reading times, such as increasing the accuracy of automatic neuron segmentation and synapse detection, secondary systems that automatically detect and correct errors in the initial predictions (Rolnick et al., 2017; Schubert et al., 2019), and better proof-reading tools and protocols (Zhao et al., 2018; Berning et al., 2015; Plaza, 2014; Cardona et al., 2012). In addition, it might be feasible to use additional information available in the EM data that constraints neuron morphology and circuitry. One such secondary information channel may come from the neuronal ultrastructure visible in EM, such as microtubules (Schneider-Mizell et al., 2016).

Microtubules

Neurons contain a multitude of different structures that enable cellular function such as mitochondria, vesicles, endoplasmic reticulum, microtubules and many more. Of particular interest for connectomics are microtubules because they provide structural stability to the cell and as a result closely follow the morphology of the neuron. As such they are well suited as an additional structural prior for neuron shape, beyond neuron membranes (Schneider-Mizell et al., 2016). However, with an outer diameter of around 24 nm, microtubules are close to the resolution limit of EM and thus difficult to reconstruct and detect. In this thesis we propose an algorithm for automatic detection and reconstruction of microtubules to eventually aid the accuracy of neuron segmentation and more generally understand cellular ultrastructure.

1.1.5 Circuits

Having generated a connectome, the next step is to use the wiring diagram to study the circuitry and its function for behavior. Reviewing all relevant literature is outside the scope of this work, but we refer the reader to the recently published work originating from the first dense, partial connectome in *Drosophila melanogaster* (Xu et al., 2020): Li et al. (2020) describe structure

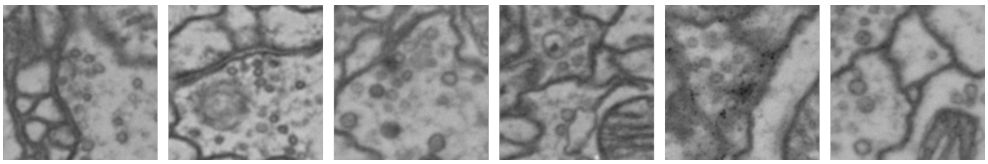


Figure 1.2: Appearance of synaptic sites from neurons with different neurotransmitters. From left to right GABA, acetylcholine, glutamate, serotonin, octopamine and dopamine.

and function of the mushroom body, the center for associative learning in the fly and Scheffer et al. (2020) provides a detailed analysis of the circuits in the entire central brain. However, as exemplified in the aforementioned studies, availability of the connectome is merely necessary but not sufficient for understanding circuits and often additional functional experiments are needed. For example, one property that is missing from the raw *Drosophila melanogaster* connectome is neurotransmitter identity to establish the sign of the connection between two neurons.

1.1.6 Neurotransmitters

The sign of a synaptic connection, i.e., whether it has an inhibitory or excitatory effect on its synaptic partners is determined by the neurotransmitter the synapse releases. There are a large variety of neurotransmitters in the *Drosophila melanogaster* brain but the three most common transmitters are the so-called classical or fast-acting transmitters GABA, glutamate and acetylcholine (Meissner et al., 2019). Glutamate and acetylcholine are often excitatory transmitters while GABA is inhibitory. In addition to fast acting transmitters, there exist so-called monoamines such as dopamine, norepinephrine, octopamine and serotonin, as well as various neuropeptides such as oxytocin, galanin or neurokinin. All transmitters are contained in vesicles and released into the synaptic cleft when a synapse is firing, transmitting the signal to its post-synaptic partners (Goyal and Chaudhury, 2013). So far, adding neurotransmitter identity to connectomes requires light microscopy (LM) in order to image fluorescent tags that have been genetically attached to neurotransmitter related molecules (Henry et al., 2012; Konstantinides et al., 2015; Davie et al., 2018; Davis et al., 2020; Hyatt and Wise, 2001; Long et al., 2017; Meissner et al., 2019). In addition, the neuron as seen in LM need to be traced and matched to neuron tracings in EM (Bates et al., 2019b,a; Costa et al., 2016). This is an expensive process and practically impossible to scale to an entire brain. In this thesis we propose an alternative route and show how a simple convolutional neural network is able to con-

fidently predict the neurotransmitter identity from EM images of synaptic sites alone. Because human experts are generally not able to perform this task in *Drosophila melanogaster* (Xu et al., 2020) we also investigate which features the classifier relies on and develop a novel neural network interpretability method to do this. For an illustration of appearance of synaptic sites with different neurotransmitters see Fig. 1.2.

1.2 Neural Network Interpretability

With an increase of deep neural networks deployed in computational pipelines across all industries, there has been an increased need for making their decision process more transparent. In particular for applications where errors have potentially devastating consequences such as cancer detection and medical AI in general (Cruz-Roa et al., 2013; Lipton, 2018; Ahmad et al., 2018), it is crucial to understand the features a neural network is using to make sure it is functioning correctly. Besides debugging and increasing trust in these systems neural network interpretability is crucial for exposing data and algorithmic biases (Tan, 2018). Furthermore, being able to interpret and distill what a network has learned may enable us to generate new (scientific) knowledge if there are tasks that neural networks can perform but humans cannot (Roscher et al., 2020). However, because neural networks are complicated objects with millions of parameters and non-linear relationships between inputs and outputs, understanding how they work can be difficult even for a single input. To deal with this issue, many methods have been proposed to perform neural network interpretability that are able to e.g. highlight the regions in a given input that contributed most to a classifiers decision, so-called attribution methods (e.g. Selvaraju et al. (2017); Sundararajan et al. (2017); Ancona et al. (2018)). Other approaches

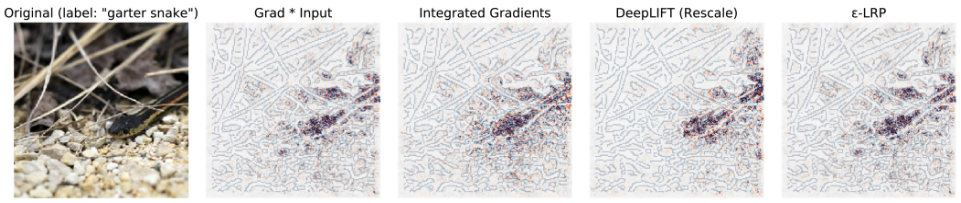


Figure 1.3: Illustration of the behaviour of popular, gradient based attribution methods applied to an Inception V3 network trained on natural image classification. Figure adapted from Ancona et al. (2018). All methods show considerable noise and it is unclear whether a human would understand that the snake is the relevant object if we did not have such strong innate visual priors about natural images.

try to approximate deep neural networks locally with an interpretable classifier (Ribeiro et al., 2016) or fitting of a rule based system (Cranmer et al., 2020). However, many of the proposed methods are based on black box optimization processes themselves and it is difficult to evaluate whether any given explanation is correct, in particular in domains such as EM where humans have little to no priors. Here we develop a novel interpretability method for EM data, which allows us to validate any given interpretation and use it to extract the previously unknown relation between neurotransmitter and synaptic phenotype. For an illustration of attribution methods on natural images see Fig. 1.3.

1.3 Thesis overview

Chapter 2 In this chapter we present our method for microtubule tracking in electron microscopy volumes, compare its performance to the prior state of the art and show its applicability to a diverse set of data sources.

Chapter 3 Here we propose a novel deep neural network interpretability method (*DAC*) for knowledge extraction from deep neural networks (DNNs). We conduct a range of experiments on three diverese datasets, including EM images of synaptic sites, to show that our approach improves on prior interpretability methods for the generation of visual explanations of DNN decision boundaries.

Chapter 4 In this chapter we show that in *Drosophila melanogaster* artificial convolutional neural networks can confidently predict the type of neurotransmitter released at a synaptic site from EM images alone. We show generalizability of the method across developmentally distinct neurons, brain regions, and datasets. We apply the method to the prediction of all automatically extracted synapses in the *FAFB* dataset, show that hemilineages likely express more than one neurotransmitter, and use the method from chapter three, to identify a set of hypothetical discriminators, that could be used by humans to distinguish the different neurotransmitter phenotypes by eye.

Chapter 2

Microtubule Tracking in Electron Microscopy Volumes[1]

2.1 Introduction

Microtubules are part of the cytoskeleton of a cell and crucial for a variety of cellular processes such as structural integrity and intracellular transport of cargo (Nogales, 2000). They are of particular interest for the connectomics community, as they directly follow the morphology of neurons. Tracking of microtubules therefore provides additional structural information that can potentially be leveraged for guided proof-reading of neuron segmentation and aid in the identification of neural subcompartments such as backbones and twigs (Schneider-Mizell et al., 2016).

Manual tracking of microtubules faces the same limitations as neuron segmentation and synapse annotations. The resolution needed to discern individual structures of interest like neural arbors, synapses, and microtubules can only be achieved with high resolution electron microscopy (EM), which results in large datasets (several hundred terabytes) even for small model organisms like *Drosophila melanogaster* (Zheng et al., 2018). With datasets of these sizes, a purely manual analysis becomes impractical. Consequently, the field of connectomics sparked a surge of automatic methods to segment neurons (for recent advances see Funke et al. (2018); Januszewski et al. (2018); Lee et al. (2019); Lee et al. (2017)), annotate synapses (Buhmann et al., 2018; Kreshuk et al., 2015; Staffler et al., 2017; Buhmann et al., 2019; Heinrich et al., 2018; Huang et al., 2018; Dorkenwald et al., 2017), and identify other structures of biological relevance such as microtubules (Buh-

[1]This chapter is based on the paper Eckstein et al. (2020b), except section 2.4 which describes our contribution to the COSEM project preprint: Heinrich et al. (2020). Julia Buhmann helped to generate the ground truth microtubule blocks used in this work. Tri Nguyen wrote the software we use for blockwise processing.

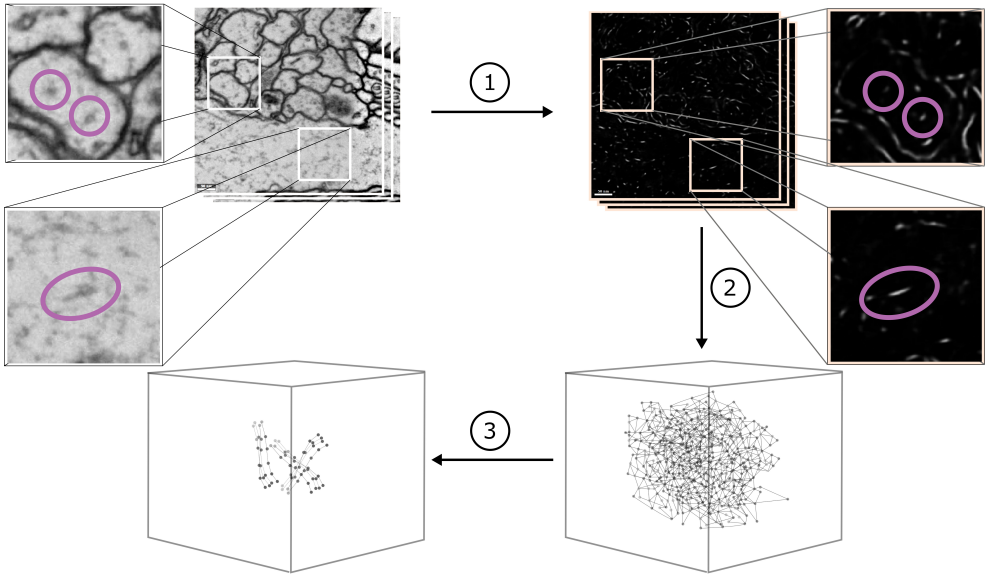


Figure 2.1: Overview of the proposed method. **1.** Microtubule scores are predicted via a 3D UNet (Ronneberger et al., 2015). Inlets show two microtubules that run perpendicular to the imaging plane (appearing as circles) and one that deviates from a 90 degree angle of incidence (appearing as a line segment). The corresponding (noisy) microtubule scores show the necessity of post processing. **2.** Candidate microtubule segments are extracted and represented as vertices in a 3D graph, where vertices are connected within a threshold distance. **3.** Final microtubule trajectories are found by solving a constrained optimization problem.

mann et al., 2016) or mitochondria (Xiao et al., 2018; Cheng and Varshney, 2017; Dorkenwald et al., 2017). Large scale automatic reconstruction of microtubules is a particularly challenging problem. With an outer diameter of 24 nm, microtubules are close to the resolution limit of serial section EM[2]. Especially in anisotropic EM volumes, the appearance of microtubules changes drastically depending on their angle of incidence to the imaging plane. Furthermore, they are often locally indistinguishable from other cell organelles (like endoplasmic reticulum) or noise.

Our method for microtubule tracking is based on the formulation proposed in Buhmann et al. (2016), with significant improvements in terms of efficiency and accuracy. Similar to Buhmann et al. (2016), we first predict a score for each voxel to be part of a microtubule. We then identify promis-

[2]Resolution is around $4 \times 4 \times 40$ nm for ssTEM, and $8 \times 8 \times 8$ nm for FIB-SEM (Takemura et al., 2015).

ing candidate points and possible links between them in a candidate graph as nodes and edges. Finally, we solve a constraint optimization problem incorporating biological priors to find a subset of edges that constitute microtubule tracks (for an overview see Fig. 2.1).
Our four main contributions are as follows: 1. We propose a new integer linear program (ILP) formulation, which decreases the time needed to solve the constraint optimization by several orders of magnitude. 2. We devise a scheme to solve the resulting optimization problem in a block-wise fashion in linear time, and thus are able to process real-world sized volumes. 3. Our formulation allows tracking of microtubules in arbitrary orientations in anisotropic volumes by introducing a non-maxima suppression (NMS) based candidate extraction method. 4. We improve the voxel-level classifier by training a 3D UNet (Ronneberger et al., 2015; Funke et al., 2018) on skeleton annotations, leading to more accurate microtubule scores.

We evaluate our method on a new benchmark comprising 153.6 μm^3 of densely traced microtubules, demonstrating a 53% increase in accuracy (0.517 $\rightarrow$ 0.789 F1 score) compared to the prior state of the art. Source code and datasets are publicly available at https://github.com/nilsec/micron.

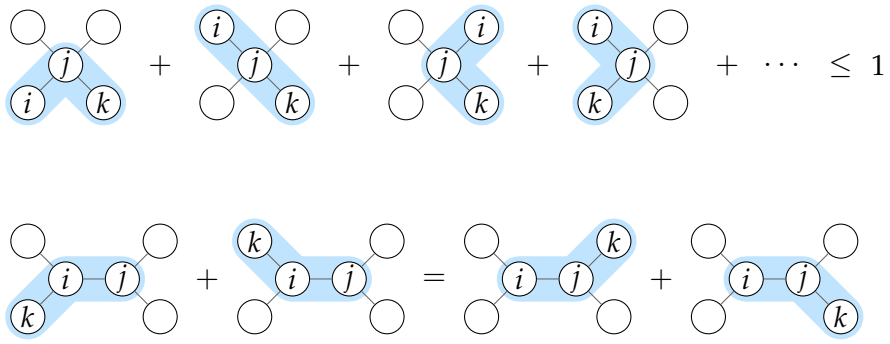
2.2 Method

2.2.1 Predictions

Starting from the raw EM input data, we train a 3D UNet (Ronneberger et al., 2015) to predict a microtubule score $m \in [0, 1]$ for each voxel. We generate microtubule scores for training from manually annotated skeletons by interpolating between skeleton markers on a voxel grid followed by Gaussian smoothing. In addition, we train the network to predict spatial gradients of the microtubule score up to second order. This is motivated by the idea that the spatial gradient encodes the local shape of a predicted object. Since microtubule segments have locally line-like shapes this auxiliary task potentially regularises microtubule score predictions.

2.2.2 Candidate Extraction

Given the predicted microtubule score we perform candidate extraction via two NMS passes, to guarantee that two successive candidates of a single microtubule track are not farther apart than the distance threshold θ_d we will use to connect two candidates with each other. In a first pass, we perform NMS and thresholding with a stride equal to the NMS window size, guaranteeing at least one candidate per NMS window if the maximum is above the threshold. This strategy is problematic if the local maximum lies on the border or corner of a NMS window as this produces multiple, in the worst case eight, candidates that are direct neighbors of each other. We



(a) Consistency constraints (top row) and no-branch constraints (bottom row).

$$
\begin{aligned}
&\min_{I_{i,j,k}} && \sum_{i \in V} c_i \cdot I_i + \sum_{(i,j) \in E} c_{i,j} \cdot I_{i,j} + \sum_{(i,j,k) \in T} c_{i,j,k} \cdot I_{i,j,k} \\
&\text{s.t.} \\
&\forall\, (i,j,k) \in T : && I_i, I_{i,j}, I_{i,j,k} \in \{0,1\} \\
&\forall\, i \in V : && 2I_i - \sum_{(i,j) \in E} I_{i,j} && = 0 \\
&\forall\, (i,j) \in E : && 2I_{i,j} - I_i - I_j && \leq 0 \\
&\forall\, (i,j,k) \in T : && 2I_{i,j,k} - I_{i,j} - I_{j,k} && \leq 0 \\
& && -I_{i,j,k} + I_{i,j} + I_{j,k} && \leq 1
\end{aligned}
$$

(b) ILP following Buhmann et al. (2016).

$$
\begin{aligned}
&\min_{I_{i,j,k}} && \sum_{(i,j,k) \in T} c_{i,j,k} \cdot I_{i,j,k} \\
&\text{s.t.} \\
&\forall\, (i,j,k) \in T : && I_{i,j,k} \in \{0,1\} \\
&\forall\, j \in V : && \sum_{(i,j,k) \in T} I_{i,j,k} && \leq 1 \\
&\forall\, (i,j) \in E : && \sum_{(k,i,j) \in T} I_{k,i,j} - \sum_{(i,j,k) \in T} I_{i,j,k} && = 0
\end{aligned}
$$

(c) Reformulated ILP on triplet indicators.

Figure 2.2: Constraint optimization on the candidate graph. We formulate an ILP on binary triplet indicators, which encode the joint selection of two incident candidate edges. The constraints shown in (a) ensure that found tracks are not crossing or splitting. Although mathematically equivalent to the formulation in (b), our formulation (c) is orders of magnitudes more efficient (see Fig. 2.4).

remove this redundancy by performing a second NMS pass on the already extracted maxima, providing us with the final set of microtubule segment candidate detections C.

2.2.3 Constrained Optimization

Following Buhmann et al. (2016), we represent each candidate microtubule segment $i \in C$ as a node in a graph with an associated position $p_i = (x_i, y_i, z_i)$. A priori we do not know which microtubule segments $i \in C$ belong together and form a microtubule. Thus, we connect all microtubule candidates with each other that are below a certain distance threshold θ_d. More formally, we introduce an undirected graph $G = (V, E)$, where $V = C \cup \{S\}$ is the set of microtubule candidate segments C augmented with a

special node S and $E \subset V \times V$ is the set of possible links between them. The special node S is used to mark the beginning or end of a microtubule track and is connected to all candidates in C. We further define a set $T = \{(i,j,k) \in V \times C \times V \mid (i,j),(j,k) \in E, i \neq k\}$ of all directly connected triplets on G.

As observed in Buhmann et al. (2016), we can make use of the fact that microtubules do not branch and have limited curvature (Gittes et al., 1993). We encode these priors as constraints and costs respectively, and solve the resulting optimization problem with an ILP. As outlined in Fig. 2.2, and in contrast to Buhmann et al. (2016), we formulate consistency and "no-branch" constraints on triplets of connected nodes $(i,j,k) \in T$ only, leading to an orders of magnitude improvement in ILP solve time (see Fig. 2.4). To this end, we introduce a binary indicator variable $I_{i,j,k} \in \{0,1\}$ for each $(i,j,k) \in T$ and define selection costs $c_{i,j,k}$ for each triplet by propagating costs c_i on nodes and $c_{i,j}$ on edges as follows:

$$c_i = \begin{cases} \theta_S & \text{if } i = S \\ \theta_P & \text{else} \end{cases} \qquad \begin{aligned} c_{i,j} &= \theta_D \operatorname{dist}(i,j) + \theta_E \operatorname{evid}(i,j) + c_i + c_j \\ c_{i,j,k} &= \theta_C \operatorname{curv}(i,j,k) + c_{i,j} + c_{j,k} \end{aligned}, \tag{2.1}$$

where θ_S is the cost for beginning/ending a track and $\theta_P < 0$ is the prior on node selection. $\operatorname{dist}(i,j) = ||p_i - p_j||$ measures the distance between candidates i and j, whereas $\operatorname{evid}(i,j) = \sum_{p \in P_{i,j}} m(p)$ accumulates the predicted evidence for microtubules on all voxels on a line $P_{i,j}$ connecting i and j. $\operatorname{curv}(i,j,k) = \pi - \angle(i,j,k)$ measures deviations of a 180 degree angle between two pairs of edges, and thus introduces a cost on curvature. The values $\theta_S, \theta_P, \theta_D, \theta_E, \theta_C \in \mathbb{R}$ are free parameters of the method and found via grid search on a validation dataset.

2.2.4 Blockwise Processing

In order to be able to apply the constraint optimization to arbitrary sized volumes, we decompose the candidate graph spatially into a set of blocks B. For each block $b \in B$, we define a constant-size context region $\overline{b}$, which encloses the block and is chosen to be large enough such that decisions outside the context region are unlikely to change the ILP solution inside the block. We next identify sets $S_i \subset B$ of blocks that are pairwise conflict free, where we define two blocks a and b to be in conflict if a overlaps with $\overline{b}$. All blocks of a subset S_i can then be distributed and processed in parallel. The corresponding ILP for each block $b \in S_i$ is solved within $\overline{b}$, however, assignments of the binary indicators are only stored for indicators corresponding to nodes in b. To obtain consistent solutions across block boundaries, existing indicator assignments from previous runs of conflicting blocks are acknowledged by adding additional constraints to the block ILP. See supplement for an illustration.

2.2.5 Evaluation

To evaluate reconstructed tracks against groundtruth, we resample both reconstruction and groundtruth tracks equidistantly and match nodes based on distance using Hungarian matching. Results are reported in terms of precision and recall on edges, which we consider correct if they connect two matched nodes that are matched to the same track.

2.3 Results

2.3.1 Dataset

We densely annotated microtubules in eight 1.2x4x4μm (30x1000x1000 voxel) volumes of EM data in all six CREMI [3] volumes A, B, C, A+, B+, C+ using Knossos (Boergens et al., 2017) and split the data in training (A+, B+, C+), validation (B+$_v$, B$_v$) and test (A, B, C) sets.

2.3.2 Comparison

*NMS_** models refer to the model described in the methods section, where NMS_SM uses a 3D UNet predicting microtubule score only, NMS_GRAD additionally predicts spatial gradients of the microtubule score up to second order and NMS_RFC uses a random forest classifier (RFC) instead of a 3D UNet. For each, we first select the best performing UNet architecture (for NMS_RFC we interactively train an RFC using Ilastik (Sommer et al., 2011)) and NMS candidate extraction threshold in terms of recovered candidates on the validation datasets, followed by a grid search over the distance threshold θ_d and ILP parameters for 150 different parameter combinations. For the NMS candidate extraction we use a window size of 1x10x10 voxels for the first NMS pass to offset the anisotropic resolution of 40x4x4 nm. For the second NMS pass we use a window size of 1x3x3 voxels, removing double detections.

Baseline refers to an adaptation[4] of the method in Buhmann et al. (2016), that uses an RFC for prediction, z section-wise connected component (CC) analysis on the thresholded microtubule scores for candidate extraction, and a fixed orientation estimate for each microtubule candidate pointing in the z direction[5]. For the baseline we interactively train two (for microtubules of different angles of incidence on the imaging plane) RFCs on training volumes A, B, C using Ilastik (Sommer et al., 2011). We find the threshold

[3] MICCAI Challenge on Circuit Reconstruction in EM Images, https://cremi.org.

[4] We use our ILP formulation, which was necessary to process larger volumes.

[5] Orientation estimate used in Buhmann et al. (2016) (direct communication with authors).

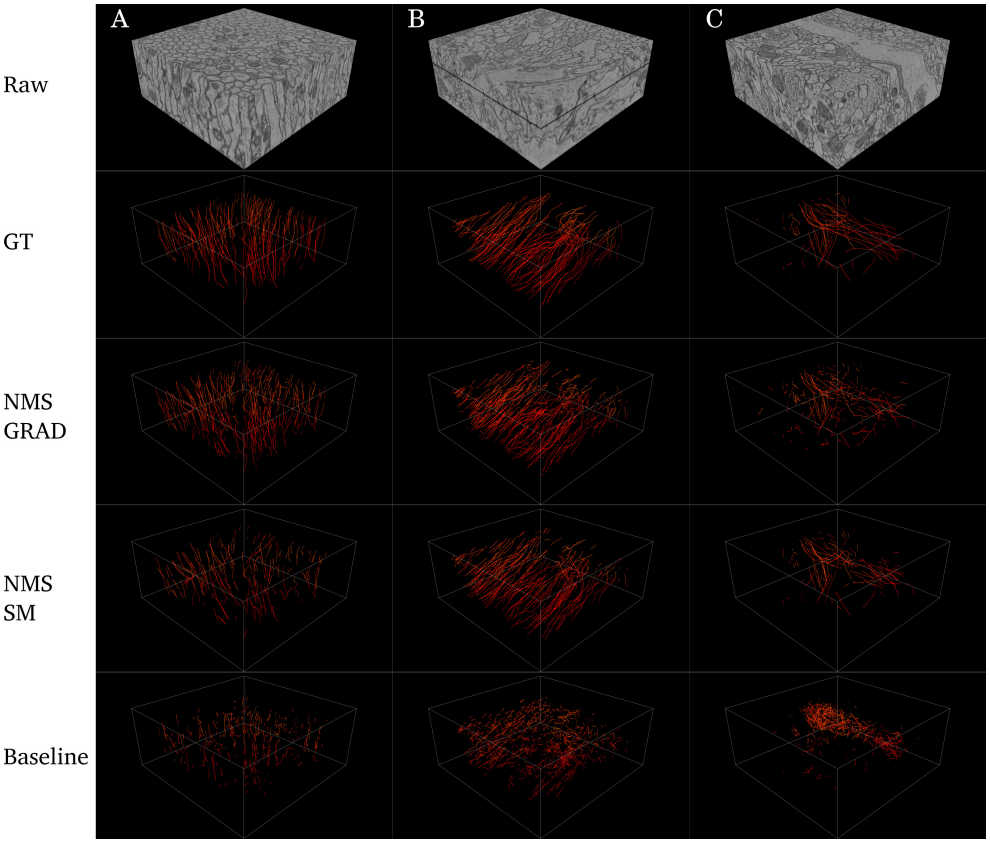


Figure 2.3: 3D rendering of raw EM volumes (Raw), manual tracing (GT) and automatically reconstructed microtubules in CREMI volumes A,B,C for our method (NMS_GRAD and NMS_SM) and the considered baseline (Buhmann et al., 2016) using validation best ILP parameters (best viewed on screen).

for CC candidate extraction, distance threshold θ_d and ILP parameters via grid search over 242 parameter configurations on the validation set. For an overview see Table 2.1.

2.3.3 Test Results

Fig. 2.4 shows that both variants of our proposed model outperform the prior state of the art (Buhmann et al., 2016) substantially. Averaged over test data sets A,B,C, we demonstrate a 53% increase in accuracy for NMS_GRAD. Table 2.1 further shows test best F1 scores for each individual dataset. In accordance with the qualitative results shown in Fig. 2.3, NMS_GRAD performs substantially better for test set A while NMS_SM is more accurate for volumes B and C. Ablation experiments show that CC candidate extraction leads to overall less accurate reconstructions. Exchanging the UNet with an RFC while retaining NMS candidate extraction seriously harms perfor-

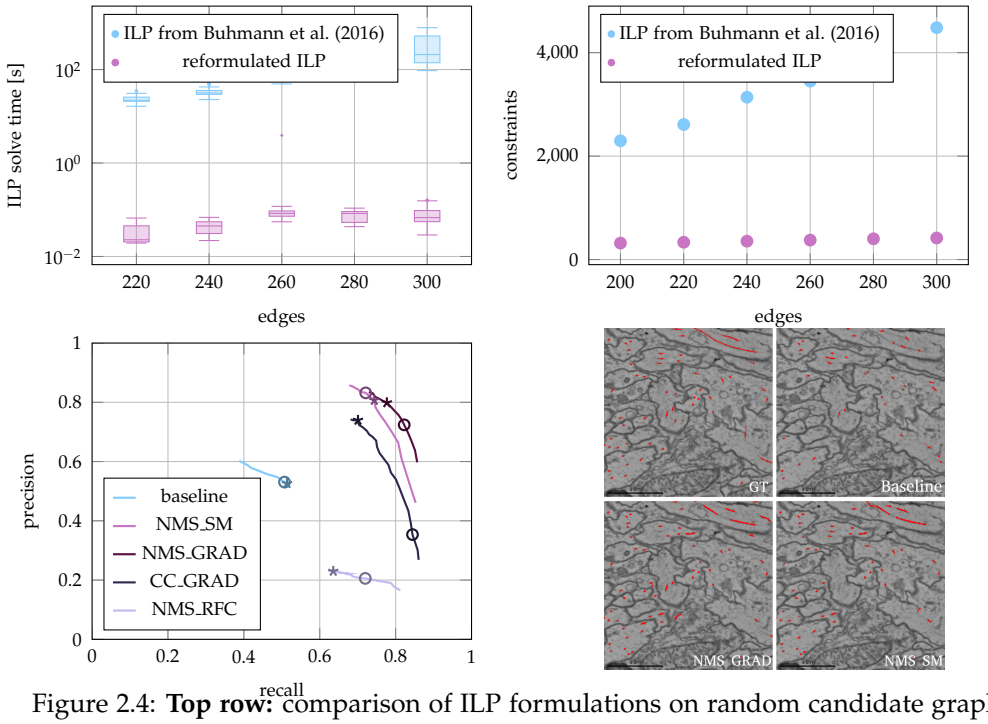


Figure 2.4: **Top row:** comparison of ILP formulations on random candidate graphs in terms of solve time (left) and number of constraints (right). Solve times have been obtained from 54 different ILP parameter configurations $\theta_{S,P,D,E,C}$ on an Intel Xeon(R), 2.40GHz x 16 CPU processor using the Gurobi optimizer. **Bottom row, left:** Comparison of our method (NMS_SM and NMS_GRAD) to the baseline (Buhmann et al., 2016) and two ablation experiments CC_GRAD (NMS replaced with connected component candidate extraction) and NMS_RFC (UNet replaced with RFC). Shown are precision and recall for varying values of the start/end edge prior θ_S averaged over the test datasets A,B,C. The validation and test best are highlighted with circles and stars, respectively. **Bottom row, right:** Qualitative results on sample B (best viewed on screen).

mance, resulting in large numbers of false positive detections. For extended qualitative results, including reconstruction of microtubules in the Calyx, a 76 x 52 x 64 μm region of the *Drosophila Melanogaster* brain, see supplement.

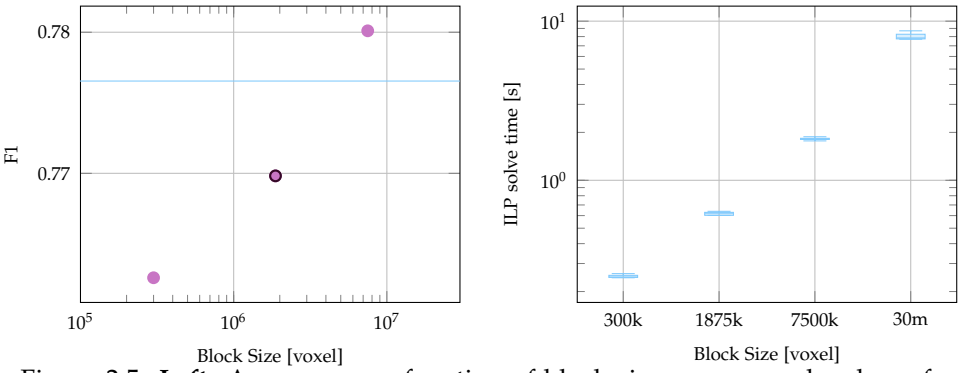


Figure 2.5: **Left**: Accuracy as a function of block size over several orders of magnitude. Shown are the F1 scores, averaged over test data sets A, B and C, using validation best parameters NMS_GRAD. Interestingly, for some sizes, solving the ILP block-wise results in higher F1 scores than solving the ILP to global optimality (blue line). However, it should be noted that the differences in F1 score are minor and likely not significant. The black circle indicates the block size we used for all reported results. **Right**: Box plot of ILP solve time per block as a function of block size. Shown is the wall-clock time needed to solve the ILP for one block, measured for ten runs, on test cube B using validation best parameters NMS_GRAD. Note that in contrast to accuracy, solve time is strongly affected by block size. This implies that we are able to process large volumes by solving the ILP in a blockwise manner, without a significant decrease in accuracy.

Table 2.1: Model overview and test best F1 score by data set.

Model	Prediction	Cand. Extr.	Edge Score	A	B	C	Avg
NMS_GRAD	UNet+GRAD	NMS	Evidence	**0.784**	0.827	0.757	**0.789**
NMS_SM	UNet	NMS	Evidence	0.711	**0.828**	**0.785**	0.775
Baseline	RFC	CC	Orientation	0.454	0.547	0.549	0.517
CC_GRAD	UNet+GRAD	CC	Evidence	0.660	0.723	0.537	0.640
NMS_RFC	RFC	NMS	Evidence	0.366	0.375	0.302	0.348

2.4 Whole Cell Microtubule Tracking

We apply the proposed method to the automatic segmentation of all microtubules in four FIB-SEM stacks of entire cells as part of the *COSEM* project with the aim to automatically segment all organelles in a given cell (Heinrich et al., 2020). The FIB-SEM volumes analysed here are two HeLa cells (jrc_hela-2 & jrc_hela-3) a jurkat cell (jrc_jurkat-1) and a macrophage cell (jrc_macrophage-2). In contrast to the full adult fly brain *FAFB* (Zheng et al., 2018), this data is isotropic with a resolution of 4x4x4nm. Additionally, we replace the classifier described here by the network from Heinrich et al. (2020) and apply the constrained optimzation as a refinement step on the network predictions.

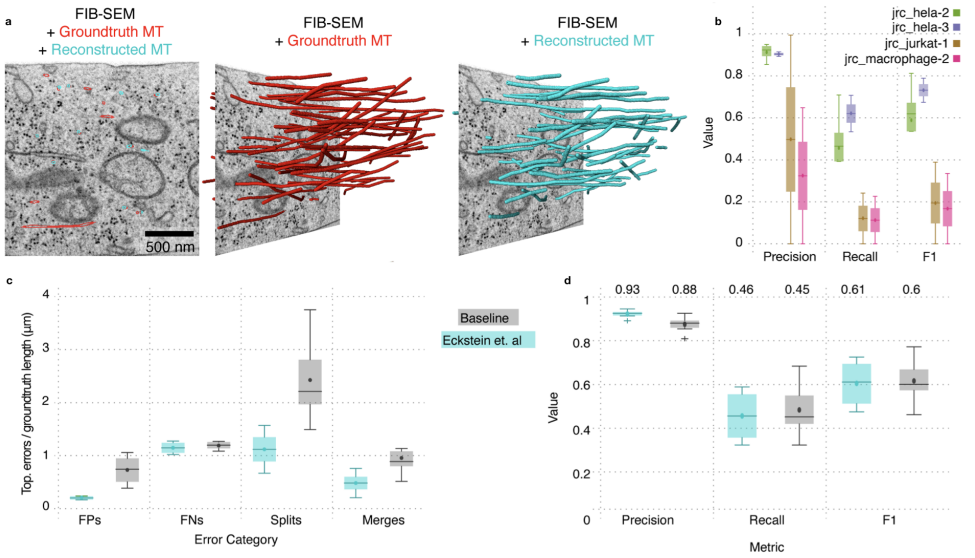


Figure 2.6: Overview of automatic microtubule tracking results in FIB-SEM stacks of entire cells. (a) Shown are raw 2D FIB-SEM slices together with microtubule ground truth tracks (red) and corresponding reconstructions (cyan) for a test block in jrc_hela-2. (b) Comparison of reconstruction accuracy over all considered cells. For each cell we show precision, recall and F1-score in two test ground truth cubes with 2μm edge length. (c) Comparison of the presented method to a simple baseline microtubule reconstruction method on isotropic 4x4x4nm resolution FIB-SEM data in terms of topological errors on entire tracks. (d) Comparison of the presented method to the considered baseline in terms of precision and recall as defined before. All comparison results have been generated via 6-fold cross validation over 4 ground truth blocks. The median value of the cross validation runs are shown above each column in (d). Figure adapted from Heinrich et al. (2020).

2.4.1 Evaluation

In order to validate that the presented method translates to isotropic data we compare the performance of our method to a baseline method (see 2.6 c, d) via 6-fold cross validation over 4 densely traced 2x2x2μm FIB-SEM blocks in jrc_hela-2. The baseline we consider consists of:

1. Thresholding of the microtubule predictions
2. Morphological closing of the thresholded predictions
3. Connected component analysis
4. Size filtering of connected components
5. Skeletonization of each connected component, where we restrict the skeleton to have no branches.

We used the following baseline hyperparameters after grid search over two out of the four ground truth blocks for each cross validation run:

1. Prediction threshold $t = 0.4$.
2. Morphological closing filter size $f = 4$
3. Connected component size filter $c = 500$

For our method we find hyperparameters in the same way (θ_E = 200, θ_C = 22, θ_S = 200 and θ_P = -200 and θ_d = 180nm) and solve the ILP with a block size of 400x400x400 nm. Using our method results in a significant reduction of morphological errors on full tracks, in particular a reduced number of false positives, splits and merges w.r.t. the baseline. Similarly we improve upon the baseline in all considered metrics in terms of individual edge accuracy. We also find that reconstruction accuracy is highly dependent on the cell. We achieve 0.6-0.8 edge accuracy for the two considered HeLa cells but perform worse for the Jurkat and Macrophage cell (Fig. 2.6 b). This is mostly caused by varying quality of the neural network predictions for microtubule scores m, which in turn is influenced by the imaging quality and amount of available training data. For a full analysis of microtubule morphology and relationship to other organelles see Heinrich et al. (2020).

2.5 Discussion

Although some of our improvements in accuracy can be attributed to the use of a deep learning classifier, the presented method relies mostly on an effective way of incorporating biological priors in the form of constraint optimization. In particular our ablation studies (CC_GRAD) show that the strided NMS-based candidate extraction method positively impacts accuracy: Since a single microtubule could potentially extend far in the x-y imaging plane, it is not sufficient to represent candidates in one plane by a single node, as done in Buhmann et al. (2016). The strided NMS detections homogenize the candidate graph and is likely decisive for the ability of our method to generalize to datasets of different resolutions, such as the presented FIB-SEM stacks. A potential downside is poor precision when combined with extremely noisy microtubule score predictions m (see NMS_RFC). In this case NMS on a grid extracts too many candidate segments, and besides structural priors, the only remaining cost we use to extract final microtubule tracks is directly derived from the (noisy) predicted microtubule score m (see equation (2.1)). Note that the baseline does not suffer as much from noisy microtubule scores, because it uses a fixed orientation prior and is thus limited to a subset of microtubules in any given volume. Finally, the reformulation of the ILP and the block-wise processing scheme result in a dramatic speed-up and the ability to perform distributed, consistent tracking, which is required to process petabyte-sized datasets.

Chapter 3

Discriminative Attribution from Counterfactuals[1]

3.1 Introduction

Machine Learning—and in particular Deep Learning—continues to see increased adoption in crucial aspects of society such as industry, science, and healthcare. As such, it impacts human lives in significant ways. Consequently, there is a need for understanding how these systems work and how they make predictions in order to increase user trust, limit the perpetuation of societal biases, ensure correct function, or even to gain scientific knowledge. However, due to the large numbers of parameters and non-linear interactions between input and output, deep neural networks (DNNs) are generally hard to interpret. In particular, it is not clear which input features influence the output of a DNN.

A popular approach for explaining DNN predictions is provided by so-called feature attribution methods. These methods output the importance of each input feature w.r.t. the output of the DNN. In the case of image classification—the primary focus of this work—the output is a heatmap over input pixels, highlighting and ranking areas of importance. A large number of approaches for feature attribution have been proposed in recent years (for a recent review see Samek et al. (2021) and related work below). Although those have been used successfully to interpret model behavior for some applications, there is still debate about the effectiveness, accuracy, and trustworthiness of these approaches (Kindermans et al., 2019; Adebayo et al., 2018; Ghorbani et al., 2019; Alvarez-Melis and Jaakkola, 2018). In addition, objectively evaluating feature attribution methods remains a difficult task (Samek et al., 2016; Hooker et al., 2018).

[1]Alexander S. Bates and Gregory S.X.E. Jefferis helped with the data collection of the Synapses dataset and provided valuable feedback.

A complementary approach for explaining DNN decisions are so called counterfactual explanations (Martens and Provost, 2014; Wachter et al., 2017). In contrast to feature importance estimation, counterfactual approaches attempt to explain a DNN output by presenting the user with another input that is close to the original input, but changes the classification decision of the DNN to another class. For humans, this representation is natural and it provides an intuitive means for elucidating DNN behaviour.

While counterfactual explainability methods have seen increased adoption in structured data domains, they are comparatively less popular for image data, where feature attribution methods arguably remain the dominant tool for practitioners. The popularity of feature importance methods is partly driven by their ease of use, availability in popular Deep Learning frameworks, and intuitive outputs in the form of pixel heatmaps. In contrast, generating counterfactual explanations typically involves an optimization procedure that needs to be carefully tuned in order to obtain a counterfactual with the desired properties. This process can be computationally expensive and does, in general, not allow for easy computation of attribution maps (Verma et al., 2020). To address these issues, we present a simple method that bridges the gap between counterfactual explainability and feature importance for image classification by building attribution maps from counterfactuals (DAC: Discriminative Attribution from Counterfactuals, see Fig. 3.1 for a visual summary). Crucially, our method can be used to quantitatively evaluate the attribution in an objective manner on a target task, a missing feature in current attribution methods. We use a cycle-GAN (Zhu et al., 2017) to translate real images $\boldsymbol{x}_{\mathrm{R}}$ of class i to counterfactual images $\boldsymbol{x}_{\mathrm{C}}$ of class $j \neq i$, where we validate that the translation has been successful by confirming that $f(\boldsymbol{x}_{\mathrm{R}}) = i$ and $f(\boldsymbol{x}_{\mathrm{C}}) = j$, where f is the classifier to interpret. We repurpose a set of common attribution methods by introducing their discriminative counterparts, which are then able to derive attribution maps from the paired real and counterfactual image. We show that this approach is able to generate sparse, high quality feature attribution maps that highlight the most discriminative features in the real and counterfactual image more precisely than standard attribution methods. Furthermore, subsequent thresholding of the attribution map allows us to extract binary masks of the features and quantify their discriminatory power by performing an intervention and replacing the highlighted pixels in the counterfactual with the corresponding pixels in the real image. The difference in output classification score of this hybrid image, compared to the real image classification, then quantifies the importance of the swapped features. We validate our method on a set of three diverse tasks, including a challenging artificial dataset, a real world biological dataset (where a DNN solves a task human experts can not), and MNIST. For all three datasets we show quantitatively and qualitatively that our method outper-

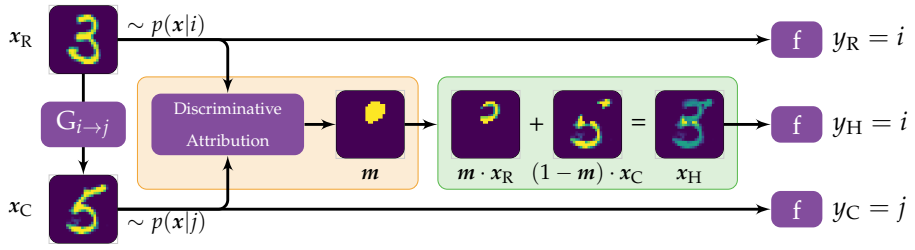


Figure 3.1: Overview of the proposed method: An input image x_R of class i is converted through an independently trained cycle-GAN generator $G_{i\to j}$ into a counterfactual image x_C of class j, such that the classifier f we wish to interpret predicts $y_R = i$ and $y_C = j$. The Discriminative Attribution method then searches for the minimal mask m, such that copying the most discriminative parts of the real image x_R into the counterfactual x_C (resulting in the hybrid x_H) is again classified as $y_H = i$.

forms all considered attribution methods in identifying key discriminatory features between the classes. Source code and datasets are publicly available at https://dac-method.github.io.

3.2 Related Work

Recent years have seen a large number of contributions addressing the problem of interpretability of DNNs. These can be broadly distinguished by the type of explanation they provide, either *local* or *global*. Methods for local interpretability provide an explanation for every input, highlighting the reasons why a particular input is assigned to a certain class by the DNN. Global methods attempt to distill the DNN in a representation that is easier to understand for humans, such as decision trees. One can further distinguish between interpretability methods that are post-hoc, i.e., applicable to every DNN after it has been trained, and those methods that require modifications to existing architectures to perform interpretable classification as part of the model. In this work we focus on a specific class of local, post-hoc approaches to DNN interpretability for image classification, so-called feature importance estimation methods.

Attribution Methods for Image Classification Even in this restricted class of approaches there is a large variety of methods (Ribeiro et al., 2016; Lundberg and Lee, 2017; Baehrens et al., 2010; Bach et al., 2015; Zintgraf et al., 2017; Selvaraju et al., 2017; Sundararajan et al., 2017; Simonyan et al., 2014; Zeiler and Fergus, 2014; Kindermans et al., 2017; Montavon et al., 2017; Fong and Vedaldi, 2017; Dabkowski and Gal, 2017; Zhang et al., 2016; Shrikumar et al., 2017, 2016; Smilkov et al., 2017). They have in common that they aim

to highlight the most important features that contributed to the output classification score for a particular class, generating a heatmap indicating the influence of input pixels and features on the output classification. Among those, of particular interest to the work presented here are *baseline* feature importance methods, which perform feature importance estimation with reference to a second input. Those methods gained popularity, as they assert *sensitivity* and *implementation invariance* (Sundararajan et al., 2017; Shrikumar et al., 2016). The baseline is usually chosen to be the zero image as it is assumed to represent a neutral input.

Counterfactual Interpretability Another body of literature that is relevant to the presented work are counterfactual interpretability methods first proposed by Martens and Provost (2014). Since then, the standard approach for generating counterfactuals broadly follows the procedure proposed by Wachter et al. (2017), in which the counterfactual is found as a result of an optimization aiming to maximize output differences while minimizing input differences between the real image $\boldsymbol{x}_{\mathrm{R}}$ and the counterfactual $\boldsymbol{x}_{\mathrm{C}}$:

$$\boldsymbol{x}_{\mathrm{C}} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \; L_i(\boldsymbol{x}_{\mathrm{R}}, \boldsymbol{x}) - L_o(f(\boldsymbol{x}_{\mathrm{R}}), f(\boldsymbol{x})) \tag{3.1}$$

with L_i and L_o some loss that measures the distance between inputs and outputs, respectively, and f the classifier in question. However, optimizing this objective can be problematic because it contains competing losses and does not guarantee that the generated counterfactual $\boldsymbol{x}_{\mathrm{C}}$ is part of the data distribution $p(\boldsymbol{x})$. Current approaches try to remedy this by incorporating additional regularizers in the objective (Liu et al., 2019; Verma et al., 2020), such as adversarial losses that aim to ensure that the counterfactual $\boldsymbol{x}_{\mathrm{C}}$ is not distinguishable from a sample $\boldsymbol{x} \sim p(\boldsymbol{x})$ (Barredo-Arrieta and Del Ser, 2020; Liu et al., 2019). However, this does not address the core problem of competing objectives and will result in a compromise between obtaining in-distribution samples, maximizing class differences, and minimizing input differences. Interpreting the presented work in this context, we circumvent this issue by dropping the input similarity loss in the generation of counterfactuals and instead enforce similarity post-hoc, similar to the strategy used by Mothilal et al. (2020).

A closely related work addressing counterfactual interpetability is the method presented by Narayanaswamy et al. (2020). Similar to ours, this method uses a cycle-GAN to generate counterfactuals for DNN interpretability. However, this method differs in that the cycle-GAN is applied multiple times to a particular input in order to increase the visual differences in the real and counterfactual images for hypothesis generation. Subsequently, the found features are confirmed by contrasting the original classifiers performance with one that is trained on the discovered features. Similar to other previous

methods, this does not lead to attribution maps or an objective evaluation of feature importance.

Attribution and Counterfactuals Closest to our approach is the work by Wang and Vasconcelos (2020), which proposes to combine attribution and counterfactual explanations. This work introduces a novel family of so-called *discriminative explanations* that also leverage attribution on a real and counterfactual image in addition to confidence scores from the classifier to derive attributions for the real and counterfactual image that show highly discriminative features. In contrast to our work, this approach requires calculation of three different attribution maps, which are subsequently combined to produce a discriminative explanation. In addition, this method does not generate new counterfactuals using a generative model, but instead selects a real image from a different class. On one hand this is advantageous because it does not depend on the generator's performance, but on the other hand this does not allow creating hybrid images for the evaluation of attribution maps.

Another relevant work is presented by Goyal et al. (2019). Similar to our work, the authors devise a method to generate counterfactual visual explanations by searching for a feature pair in two real images of different classes that, if swapped, influences the classification decision. To this end, they propose an optimization procedure that searches for the best features to swap, utilizing the networks feature representations. In contrast to our work, the usage of real (instead of generated) counterfactuals can lead to more artifacts during the replacement of features. In addtion, our work supports the generation of attribution maps and features a procedure for the quantitative evaluation of the explanations.

Attribution Evaluation Our work differs notably from the current state of the art as it enables quantitative evaluation of the generated attributions by copy-pasting features from a paired image set (the real and the counterfactual). Prior work evaluated the importance of highlighted features by removing them (Samek et al., 2016). However, it has been noted that this strategy is problematic because it is unclear whether any observed performance degradation is due to the removal of relevant features or because the new sample comes from a different distribution. As a result, strategies to remedy this issue have been proposed, for example by retraining classifiers on the modified samples (Hooker et al., 2018). Instead of removing entire features, in this work we replace them with their corresponding counterfactual features.

3.3 Method

The method we propose combines counterfactual interpretability with discriminative attribution methods to find and highlight the most important features between images of two distinct classes i and j, given a pretrained classifier f. For that, we first generate for a given input image $\boldsymbol{x}_{\mathrm{R}}$ of class i a counterfactual image $\boldsymbol{x}_{\mathrm{C}}$ of class j. We then use a *discriminative* attribution method to find the attribution map of the classifier for this pair of images. As we will show qualitatively and quantitatively in Section 3.4, using paired images results in attribution maps of higher quality. Furthermore, the use of a counterfactual image gives rise to an objective evaluation procedure for attribution maps.

In the next sections we describe (1) our choice for generating counterfactual images, (2) the derivation of discriminative attribution methods from existing baseline attribution methods, and (3) how to use counterfactual images to evaluate attribution maps. We denote with f a pretrained classifier with N output classes, input images $\boldsymbol{x} \in \mathbb{R}^{h \times w}$, and output vector $f(\boldsymbol{x}) = \boldsymbol{y} \in [0,1]^N$ with $\sum_i y_i = 1$.

3.3.1 Creation of Counterfactuals

We train a cycle-GAN (Zhu et al., 2017) for each pair of image classes $i \neq j \in \{0, 1, ..., N-1\}$, which enables translation of images of class i into images of class j and vice versa. We perform this translation for each image of class i and each target class $j \neq i$ to obtain datasets of paired images $D_{i \to j} = \left\{(\boldsymbol{x}_{\mathrm{R}}^k, \boldsymbol{x}_{\mathrm{C}}^k) | k = 1, \ldots, n(i)\right\}$, where $\boldsymbol{x}_{\mathrm{R}}^k$ denotes the kth real image of class i and $\boldsymbol{x}_{\mathrm{C}}^k$ its counterfactual of class j. We then test for each image in the dataset whether the translation was successful by classifying the counterfactual image $\boldsymbol{x}_{\mathrm{C}}$ and reject a sample pair whenever $f(\boldsymbol{x}_{\mathrm{C}})_j < \theta$, with θ a threshold parameter (in the rest of this work we set $\theta = 0.8$, except otherwise specified).

This procedure results in a dataset of paired images, where the majority of the differences between an image pair is expected to be relevant for the classifiers decision, i.e., we retain formerly present non-discriminatory distractors such as orientation, lighting, or background. We encourage that the translation makes as little changes as necessary by choosing a Res-Net (He et al., 2016) architecture for the cycle-GAN generator, which is able to trivially learn the identity function.

3.3.2 Discriminative Attribution from Counterfactuals

The datasets $D_{i \to j}$ are already useful to visualize data-intrinsic class differences (see Fig. 3.5 for examples). However, we wish to understand which

input features the classifier f makes use of. Specifically, we are interested in finding the smallest binary mask $\boldsymbol{m}$, such that swapping the contents of $\boldsymbol{x}_{\mathrm{C}}$ with $\boldsymbol{x}_{\mathrm{R}}$ within this mask changes the classification under f.

To find $\boldsymbol{m}$, we repurpose existing attribution methods that are amendable to be used with a reference image. The goal of those methods is to produce attribution maps $\boldsymbol{a}$, which we convert into a binary mask via thresholding. A natural choice for our purposes are so-called *baseline attribution methods*, which derive attribution maps by contrasting an input image with a baseline sample (e.g., a zero image). In the following, we review suitable attribution methods and derive discriminative versions that use the counterfactual image as their baseline. We will denote the discriminative versions with the prefix *D*.

Input * Gradients

One of the first and simplest attribution methods is *Input * Gradients* (INGRADS) (Shrikumar et al., 2016; Simonyan et al., 2014), which is motivated by the first order Taylor expansion of the output class with respect to the input around the zero point:

$$\mathrm{INGRADS}(\boldsymbol{x}) = |\nabla_x f(\boldsymbol{x})_i \cdot \boldsymbol{x}|, \tag{3.2}$$

where i is the class for which an attribution map is to be generated. We derive an explicit baseline version for the discriminatory attribution of the real $\boldsymbol{x}_{\mathrm{R}}$ and its counterfactual $\boldsymbol{x}_{\mathrm{C}}$ by choosing $\boldsymbol{x}_{\mathrm{C}}$ as the Taylor expansion point:

$$D\text{-}\mathrm{INGRADS}(\boldsymbol{x}_{\mathrm{R}}, \boldsymbol{x}_{\mathrm{C}}) = \left|\nabla_x f(\boldsymbol{x})_j\Big|_{\boldsymbol{x}=\boldsymbol{x}_{\mathrm{C}}} \cdot (\boldsymbol{x}_{\mathrm{C}} - \boldsymbol{x}_{\mathrm{R}})\right|, \tag{3.3}$$

where j is the classes of the counterfactual image.

Integrated Gradients

Integrated Gradients (IG) is an explicit baseline attribution method, where gradients are accumulated along the straight path from a baseline input $\boldsymbol{x}_0$ to the input image $\boldsymbol{x}$ to generate the attribution map (Sundararajan et al., 2017). Integrated gradients along the kth dimension are given by:

$$\mathrm{IG}_k(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{x}_0)_k \cdot \int_{\alpha=0}^{1} \frac{\partial f(\boldsymbol{x}_0 + \alpha(\boldsymbol{x} - \boldsymbol{x}_0))_i}{\partial \boldsymbol{x}_k} d\alpha. \tag{3.4}$$

We derive a discriminatory version of IG by replacing the baseline as follows:

$$D\text{-}\mathrm{IG}_k(\boldsymbol{x}_{\mathrm{R}}, \boldsymbol{x}_{\mathrm{C}}) = (\boldsymbol{x}_{\mathrm{C}} - \boldsymbol{x}_{\mathrm{R}})_k \cdot \int_{\alpha=0}^{1} \frac{\partial f(\boldsymbol{x}_{\mathrm{R}} + \alpha(\boldsymbol{x}_{\mathrm{C}} - \boldsymbol{x}_{\mathrm{R}}))_j}{\partial \boldsymbol{x}_k} d\alpha. \tag{3.5}$$

Deep Lift

Deep Lift (DL) is also an explicit baseline attribution method which aims to compare individual neurons activations of an input w.r.t. a reference baseline input (Shrikumar et al., 2016). It can be expressed in terms of the gradient in a similar functional form to IG:

$$\mathrm{DL}(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{x}_0) \cdot F_{DL}, \tag{3.6}$$

where F_{DL} is some function of the gradient of the output (see Ancona et al. (2018) for the full expression). The discriminative attribution we consider is simply:

$$D\text{-DL}(\boldsymbol{x}_\mathrm{R}, \boldsymbol{x}_\mathrm{C}) = (\boldsymbol{x}_\mathrm{C} - \boldsymbol{x}_\mathrm{R}) \cdot F_{DL}. \tag{3.7}$$

GradCAM

GradCAM (GC) is an attribution method that considers the gradient weighted activations of a particular layer, usually the last convolutional layer, and propagates this value back to the input image (Selvaraju et al., 2017). We denote the activation of a pixel (u, v) in layer l with size (h, w) and channel k by $C^l_{k,u,v}$ and write the gradient w.r.t. the output $\boldsymbol{y}$ as:

$$\nabla_{C^l_k}\boldsymbol{y} = \left(\frac{d\boldsymbol{y}}{dC^l_{k,0,0}}, \frac{d\boldsymbol{y}}{dC^l_{k,1,0}}, \frac{d\boldsymbol{y}}{dC^l_{k,2,0}}, ..., \frac{d\boldsymbol{y}}{dC^l_{k,h,w}}\right) \tag{3.8}$$

The original GC is then defined as:

$$\mathrm{GC}(\boldsymbol{x}) = \mathrm{ReLU}\left(\sum_k \nabla_{C_k}\boldsymbol{y} \cdot \mathbf{C}_k\right) = \mathrm{ReLU}\left(\sum_k \sum_{u,v} \frac{d\boldsymbol{y}}{dC_{k,u,v}} C_{k,u,v}\right) \tag{3.9}$$

$$= \mathrm{ReLU}\left(\sum_k \alpha_k C_k\right), \tag{3.10}$$

where we ommitted the layer index l for brevity. Each term $\frac{d\boldsymbol{y}}{dC_{k,u,v}} C_{k,u,v}$ is the contribution of pixel u, v in channel k to the output classification score $\boldsymbol{y}$ under a linear model. GC utilizes this fact and projects the layer attribution from layer l back to the input image, generating the final attribution map.

In contrast to the setting considered by GC, we have access to a matching pair of real and counterfactual images $\boldsymbol{x}_\mathrm{R}$ and $\boldsymbol{x}_\mathrm{C}$. We extend GC to consider both feature maps $C^{\boldsymbol{x}_\mathrm{R}}_k$ and $C^{\boldsymbol{x}_\mathrm{C}}_k$ by treating GC as an implicit zero baseline method similar to INGRADS:

$$D\text{-GC}_k(\boldsymbol{x}_\mathrm{R}, \boldsymbol{x}_\mathrm{C}) = \frac{d\boldsymbol{y}_j}{dC_k}\Big|_{C=C^{\boldsymbol{x}_\mathrm{C}}_k} \left(C_k(\boldsymbol{x}_\mathrm{C}) - C_k(\boldsymbol{x}_\mathrm{R})\right). \tag{3.11}$$

Averaging those gradients over feature maps k, and projecting the activations back to image space then highlights pixels that are most discriminative for a particular pair:

$$D\text{-GC}_{\text{P}}(\boldsymbol{x}_{\text{R}}, \boldsymbol{x}_{\text{C}}) = \left| \mathbb{P} \sum_{k} D\text{-GC}_{k}(\boldsymbol{x}_{\text{R}}, \boldsymbol{x}_{\text{C}}) \right|, \tag{3.12}$$

where $\mathbb{P}$ is the projection matrix from feature space C to input space X. Note that in contrast to GC, we use the absolute value of the output attribution, as we do not apply ReLU activations to layer attributions.

Because feature maps can be of lower resolution than the input space, GC tends to produce coarse attribution maps (Selvaraju et al., 2017). To address this issue it is often combined with *Guided Backpropagation* (GBP), a method that uses the gradients of the output class w.r.t. the input image as the attribution map (Springenberg et al., 2014). During the backwards pass, all values < 0 at each ReLU non-linearity are then discarded to only retain positive attributions.

Guided GradCAM (GGC) uses this strategy to sharpen the attribution of GC via element-wise multiplication of the attribution maps (Selvaraju et al., 2017). For the baseline versions we thus consider multiplication of D-GC with the GBP attribution maps:

$$\text{GBP}(\boldsymbol{x}) = \nabla_{x} f(\boldsymbol{x})_{i} \ \text{, with} \nabla\text{ReLU} > 0 \tag{3.13}$$

$$\text{GGC}(\boldsymbol{x}) = \text{GC}(\boldsymbol{x}) \cdot \text{GBP}(\boldsymbol{x}) \tag{3.14}$$

$$D\text{-GGC}(\boldsymbol{x}_{\text{R}}, \boldsymbol{x}_{\text{C}}) = D\text{-GC}(\boldsymbol{x}_{\text{R}}, \boldsymbol{x}_{\text{C}}) \cdot \text{GBP}(\boldsymbol{x}_{\text{R}}). \tag{3.15}$$

3.3.3 Evaluation of Attribution Maps

The discriminative attribution map $\boldsymbol{a}$ obtained for pair of images $(\boldsymbol{x}_{\text{R}}, \boldsymbol{x}_{\text{C}})$ can be used to quantify the causal effect of the attribution. Specifically, we can copy the area highlighted by $\boldsymbol{a}$ from the real image $\boldsymbol{x}_{\text{R}}$ of class i to the counterfactual image $\boldsymbol{x}_{\text{C}}$ of class j, resulting in a hybrid image $\boldsymbol{x}_{\text{H}}$. If the attribution accurately captures class-relevant features, we would expect that the classifier f assigns a high probability to $\boldsymbol{x}_{\text{H}}$ being of class i.

The ability to create those hybrid images is akin to an intervention, and has two important practical implications: First, it allows us to find a minimal binary mask that captures the most class-relevant areas for a given input image. Second, we can compare the change in classification score for hybrids derived from different attribution maps. This allows us to compare different methods in an objective manner, following the intuition that an attribution map is better, if it changes the classification with less pixels changed.

To find a minimal binary mask $\boldsymbol{m}_{\text{min}}$, we search for a threshold of the attribution map $\boldsymbol{a}$, such that the mask score $\Delta f(\boldsymbol{x}_{\text{H}}) = f(\boldsymbol{x}_{\text{H}})_{i} - f(\boldsymbol{x}_{\text{C}})_{i}$ (i.e.,

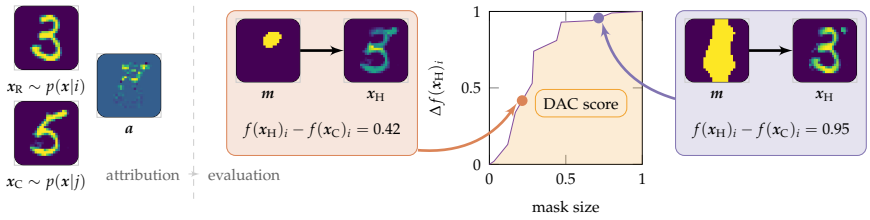


Figure 3.2: Evaluation procedure for discriminative attribution methods: Given the real image $\boldsymbol{x}_R$ of class i and its counterfactual $\boldsymbol{x}_C$ of class j, we generate a sequence of binary masks $\boldsymbol{m}$ by applying different thresholds to the attribution map $\boldsymbol{a}$. Those masks are then used to generate a sequence of hybrid images $\boldsymbol{x}_H$. The plot shows the change in classifier prediction $\Delta f(\boldsymbol{x}_H)_i = f(\boldsymbol{x}_H)_i - f(\boldsymbol{x}_C)_i$ over the size of the mask $\boldsymbol{m}$ (normalized between 0 and 1). The DAC score is the area under the curve, i.e., a value between 0 and 1. Higher DAC scores are better and indicate that a discriminative attribution method found small regions that lead to the starkest change in classification.

the change in classification score) is maximized while the size of the mask is minimized, i.e., $\boldsymbol{m}_{\text{min}} = \arg\min_{\boldsymbol{m}} \ |\boldsymbol{m}| - \Delta f(\boldsymbol{x}_H)$ (where we omitted the dependency of $\boldsymbol{x}_H$ on $\boldsymbol{m}$ for brevity). In order to minimize artifacts in the copying process we also apply a morphological closing operation with a window size of 10 pixels followed by a Gaussian Blur with $\sigma = 11px$. The final masks highlight the relevant class discriminators by showing the user the counterfactual features, the original features they are replaced with, and the corresponding mask score $\Delta f(\boldsymbol{x}_H)$, indicating the quantitative effect of the replacement on the classifier. See Fig. 3.5 for example pairs and corresponding areas $\boldsymbol{m}_{\text{min}}$.

Furthermore, by applying a sequence of thresholds for the attribution map $\boldsymbol{a}$, we derive an objective evaluation procedure for a given attribution map: For each hybrid image $\boldsymbol{x}_H$ in the sequence of thresholds, we consider the change in classifier prediction relative to the size of the mask that has been used to create the hybrid. We accumulate the change in classifier prediction over all mask sizes to derive our proposed DAC score. This procedure is explained in detail in Fig. 3.2 for a single pair of images. When reporting the DAC score for a particular attribution method, we average the single DAC scores over all images, and all distinct pairs of classes.

3.4 Experiments

We evaluate the presented method on four datasets: MNIST (LeCun and Cortes, 2010), SYNAPSES (Eckstein et al., 2020a)[2] and two versions of a syn-

[2]Dataset kindly provided by the authors of Eckstein et al. (2020a).

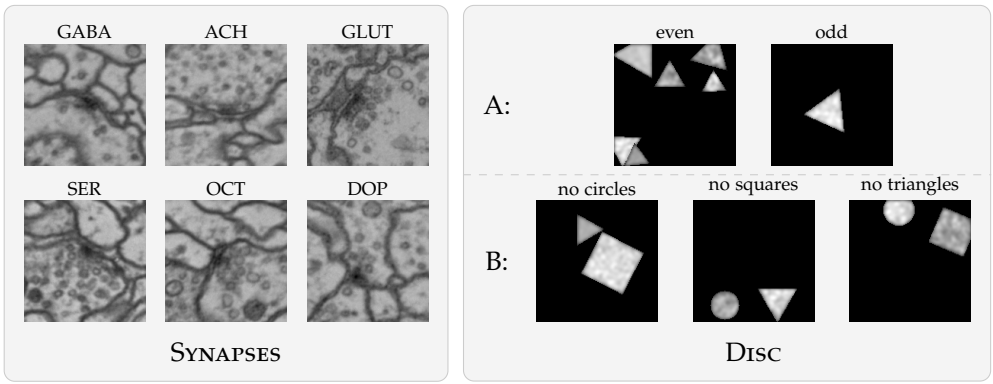


Figure 3.3: Example images of datasets Synapses and Disc. Synapses consists of electron microscopy images of synapses. Each class is defined by the neurotransmitter the synapse releases. Disc is a synthetic dataset we designed in order to highlight failure cases of popular attribution methods. We consider two subsets: Disc-A shows triangles in each image and classes are defined by the parity of the number of triangles. Disc-B consists of images showing triangles, squares, and disks. Each class is one combination of two shapes.

thetic dataset that we call Disc-A and Disc-B (see Fig. 3.3 for an overview).

Synapses A real world biological dataset, consisting of 128×128px electron microscopy images of synaptic sites in the brain of *Drosophila melanogaster*. Each image is labelled with a functional property of the synapse, namely the neurotransmitter it releases (the label was acquired using immunohistochemistry labelling, see Eckstein et al. (2020a) for details). This dataset is of particular interest for interpretability, since a DNN can recover the neurotransmitter label from the images with high accuracy, but human experts are not able to do so. Interpretability methods like the one presented here can thus be used to gain insights into the relation between structure (from the electron microscopy image) and function (the neurotransmitter released).

Disc-A and Disc-B Two synthetic datasets with different discriminatory features of different difficulty. Each image is 128×128px in size and contains spheres, triangles or squares. For Disc-A, the goal is to correctly classify images containing an even or odd number of triangles. Disc-B contains images that show exactly two of the three available shapes and the goal is to predict which shape is missing (e.g., an image with only triangles and squares is to be classified as "does not contain spheres"). This dataset was deliberately designed to investigate attribution methods in a setting where the discrimination depends on the absence of a feature.

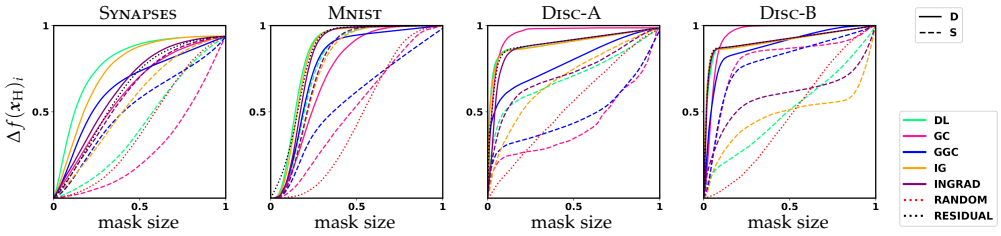


Figure 3.4: Quantitative evaluation of discriminative (D - solid) and corresponding original (S for "single input" - dashed) attribution methods over four datasets. Corresponding D and S versions of the same method are shown in the same color. For each, we plot the average change of classifier prediction $\Delta f(\boldsymbol{x}_H)_i^k = f(\boldsymbol{x}_H)_i - f(\boldsymbol{x}_C)_i$ as a function of mask size $\boldsymbol{m} \in [0,1]$. In addition we show performance of the two considered baselines: masks derived from random attribution maps (random - red, dotted) and mask derived from the residual of the real and counterfactual image (residual - black, dotted). On all considered datasets all versions of D attribution outperform their S counterparts. All experiments are performed with VGG architectures. For ResNet results of MNIST and DISC see supplement.

Training For MNIST and DISC, we train a VGG and ResNet for 100 epochs and select the epoch with highest accuracy on a held out validation dataset. For SYNAPSES we adapt the 3D-VGG architecture from Eckstein et al. (2020a) to 2D and train for 500,000 iterations. We select the iteration with the highest validation accuracy for testing. For each dataset we train one cycle-GAN for 200 epochs, on each class pair and on the same training set the respective classifier was trained on (the full network specifications are given in the supplement).

Results Quantitative results (in terms of the DAC score, see Section 3.3.3) for each investigated attribution method are shown in Fig. 3.4 and Table 3.1. In summary, we find that attribution maps generated from the proposed discriminative attribution methods consistently outperform their original versions in terms of the DAC score. This observation also holds visually: the generated masks from discriminative attribution methods are smaller and more often highlight the main discriminatory parts of a considered image pair (see Fig. 3.5). In particular, the proposed method substantially outperforms the considered random baseline, whereas standard attribution methods sometimes fail to do so (e.g., GC on dataset SYNAPSES). Furthermore, on MNIST and DISC-A, the mask derived from the residual of real and counterfactual image is already competitive with the best considered methods and outperforms standard attribution substantially. However, for more complex datasets such as SYNAPSES the residual becomes less accurate in highlighting discriminative features. Here, the discriminatory attributions outperform all other considered methods.

Dataset	D-IG	D-DL	D-INGR.	D-GC	D-GGC	RES.	IG	DL	INGR.	GC	GGC	RND.
MNIST	0.83	**0.84**	0.82	0.73	0.78	0.84	0.77	0.79	0.77	0.52	0.56	0.46
SYNAPSES	0.75	**0.79**	0.65	0.62	0.65	0.63	0.56	0.43	0.61	0.28	0.52	0.41
DISC-A	0.9	0.9	0.88	**0.95**	0.79	0.9	0.69	0.7	0.72	0.43	0.48	0.54
DISC-B	0.91	0.91	0.91	**0.95**	0.88	0.91	0.48	0.51	0.6	0.8	0.79	0.48

Table 3.1: Summary of DAC scores for each investigated method on the three datasets MNIST, SYNAPSES, and DISC (two versions) corresponding to 3.4. Best results are highlighted. For extended results with ResNet architectures see supplement.

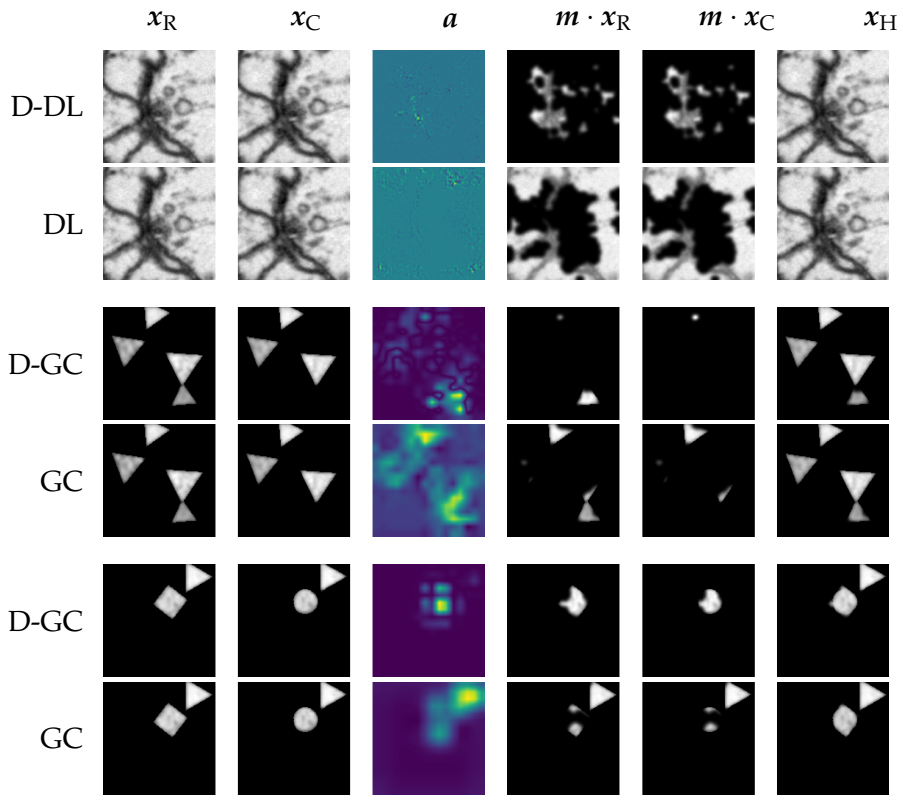


Figure 3.5: Samples from the best performing method pair on SYNAPSES and DISC-A and B. Discriminative attribution methods are able to highlight the key discriminative features while vanilla versions often fail to do so (e.g., a subtle intensity change in the synaptic cleft in the top rows). Further qualitative results (including the other considered datasets) can be found in the supplement.

3.5 Discussion

This work demonstrates that the combination of counterfactual interpretability with suitable attribution methods is more accurate in extracting key discriminative features between class pairs than standard methods. While the method succeeds in the presented experiments, it comes with a number of limitations. It requires the training of cycle-GANs, one for each pair of output classes. Thus training time and compute cost scale quadratically in the number of output classes and it is therefore not feasible for classification problems with a large number of classes. Furthermore, the translation from the real to the counterfactual image could fail for a large fraction of input images, i.e., $f(\boldsymbol{x}_C) \neq j$. In this work, we only consider those image pairs where translation was successful, as we focus on extracting knowledge about class differences from the classifier. For applications that require an attribution for each input image this approach is not suitable. An additional concern is that focusing only on images that have a successful translation may bias the dataset we consider and with it the results. GANs are known to exhibit so called mode collapse (Che et al., 2016; Salimans et al., 2016), meaning they focus on only a small set of modes of the full distribution. As a consequence, the method described here may miss discriminatory features present in other modes. Using a cycle-GAN is not possible in all image domains. Image classes need to be sufficiently similar in appearance for the cycle-GAN to work, and translating, e.g., an image of a mouse into an image of a tree is unlikely to work and produce meaningful attributions. However, we believe that the generation of masks in combination with the corresponding mask score is superior for interpreting DNN decision boundaries than classical attribution maps and suggest the usage of cycle-GAN baselines for attribution in cases where a fine grained understanding of class differences is sought.

Although we present this work in the context of understanding DNNs and the features they make use of, an uncritical adaptation of this and other similar interpretability methods can potentially lead to ethical concerns. As an example, results should be critically evaluated when using this method to interpret classifiers that have been trained to predict human behaviour, or demographic and socioeconomic features. As with any data-heavy method, it is important to realize that results will be reflective of data- and model-intrinsic biases. As such, an interpretability method like the one we present here can at most identify a correlation between input features and labels, but not true causal links. The method presented here should therefore not be used to "proof" that a particular feature leads to a particular outcome. Such claims should be met with criticism to prevent agenda-driven narratives of malicious actors.

Chapter 4

Neurotransmitter Classification from Electron Microscopy Images at Synaptic Sites in Drosophila[1]

4.1 Introduction

In recent years, advances in imaging technology enabled high resolution electron microscopy (EM) imaging of whole brain data sets (Zheng et al., 2018; Ryan et al., 2016; Cook et al., 2019; Ohyama et al., 2015), opening up the possibility of generating cellular level wiring diagrams (connectomes) of nervous systems. Generating a connectome entails identifying all neurons and the synapses that connect them. Due to the size of these data sets manual tracing of all neurons and synapses is not feasible even for comparitively small organisms such as *Drosophila melanogaster*. However, recent advances in automated methods for segmenting neurons (Funke et al., 2018; Januszewski et al., 2018; Lee et al., 2019) and detecting synapses (Kreshuk et al., 2015; Staffler et al., 2017; Buhmann et al., 2019) greatly reduce the time of human involvement in these tasks and have just recently been applied to generate the connectome for a large part of the *Drosophila melanogaster* brain (Xu et al., 2020).

However, EM data does not directly give us information about gene expression and as a result quantities such as neurotransmitter identity, while crucial to determine the function of any given synapse, are unknown for a

[1]This chapter is an extended version of the preprint Eckstein et al. (2020a). It also contains parts of our contribution to the preprint Baker et al. (2021) and the paper Li et al. (2020). Alexander S. Bates helped with the conception of this project, data collection, data analysis and visualization. He also identified the hemilineages used in this work together with Volker Hartenstein. Gregory S.X.E. Jefferis analyzed the whole brain neurotransmitter predictions, provided visualizations and wrote software for accessing the predictions. Michelle Du ran the initial experiments for this study.

majority of the cells in the connectome of *Drosophila melanogaster*. The action a neuron has on its downstream targets is determined by the neurotransmitters it releases into the extracellular space. Before release, neurotransmitters are packaged into different types of vesicles at synaptic sites. The so-called *classical*, fast-acting transmitters GABA, acetylcholine and glutamate are contained in small, clear vesicles, while monoamines such as dopamine, norepinephrine, octopamine and serotonin are packaged into pleomorphic clear-core or small dense-core vesicles (Goyal and Chaudhury, 2013). The large number of various neuropeptides such as cholecystokinin, galanin, neurokinin and oxytocin are contained in large dense-core vesicles. In vertebrates it is generally possible for humans to distinguish between different clear-core and dense-core vesicles (Goyal and Chaudhury, 2013) (indicating neurotransmitter identity) and there are automated methods for classifying symmetric and asymmetric synapses (Dorkenwald et al., 2017). In contrast, for invertebrates such as *Drosophila melanogaster* it is so far unknown whether synaptic phenotype, as seen in EM, is sufficient to consistently determine neurotransmitter identity, especially different varieties of clear-core vesicles.

As a result, adding neurotransmitter identities to connectomic data requires light microscopy (LM) pipelines, in which gene transcripts or proteins involved in the pathway of interest have been made visible using fluorescent probes. Common methods for neurotransmitter detection include sequencing transcriptomics (RNAseq) (Henry et al., 2012; Konstantinides et al., 2015; Davie et al., 2018; Davis et al., 2020), immunolabeling (Hyatt and Wise, 2001) and fluorescent *in situ* hybridization (FISH) (Long et al., 2017; Meissner et al., 2019). Subsequent morphological matching of these neurons to reconstructed neurons in the EM data can then be performed using NBLAST (Costa et al., 2016), providing neurotransmitter identity to connectome data (Bates et al., 2019b,a).

However, this approach is very difficult to scale to an entire connectome of *Drosophila melanogaster* comprising ∼150,000 cells. Although imaging expression patterns for multiple neurotransmitters on a brain scale can be done in a matter of minutes to days depending on the required resolution[2] (Meissner et al., 2019), bridging the gap between LM and EM remains challenging: In addition to imaging expression patterns of neurotransmitter related proteins, it requires a well-characterized, sparse genetic driver line in order to perform accurate morphological matching to EM tracings using tools such as NBLAST (Costa et al., 2016). As a result, transmitter identity is known for only a few hundred types of neurons (Bates et al., 2019a).

[2]Throughput estimated at around one neuron per minute at sufficient spatial resolution for colocalization with single cell labeling - personal communication with authors.

Here, we show that it is possible to determine the primary neurotransmitter of a given neuron in the *Drosophila melanogaster* brain from the phenotype of its synaptic sites in EM alone. For that, we train a deep learning classifier to predict the neurotransmitter of a $640 \times 640 \times 640\text{nm}^3$ 3D EM volume with a synaptic site at the center. We find that this method is able to classify the neurotransmitter of any given synapse with 87% accuracy on average. Furthermore, we show on a large test set that the classifier generalizes across neurons with different developmental histories (*i.e.*, that derive from different 'hemilineages'), brain regions and datasets, indicating that the influence of the neurotransmitter on the phenotype of a synaptic site is largely conserved and that the classifier learned robust features. We use our method to predict the neurotransmitter identity of over 1000 neurons in 89 hemilineages with so far unknown neurotransmitter identities in the *Drosophila melanogaster* brain. In contrast to recent findings in the ventral nervous system (VNS) (Lacin et al., 2019), our results suggest that the neurotransmitter identity of neurons within hemilineages in the brain is not limited to one fast-acting transmitter. Furthermore, we predict the neurotransmitter of all automatically detected synapses (Buhmann et al., 2018) in the *Drosophila melanogaster* brain and make the data publicly available. Given that the relation of synaptic phenotype and neurotransmitter identity is not fully understood in *Drosophila melanogaster*, we use neural network interpretability tools to extract a set of hypothetical discriminatory features between synaptic sites of different neurotransmitter identity, elucidating another piece of the relationship between structure and function in the *Drosophila melanogaster* brain.

In summary, our method circumvents a major bottleneck in neurotransmitter identification, matching LM expression patterns to EM tracings, and is able to assign neurotransmitter identity to individually traced neurons in a matter of seconds. Combined with automated synapse detection methods (Buhmann et al., 2018; Kreshuk et al., 2015; Staffler et al., 2017; Buhmann et al., 2019), we generate a comprehensive neurotransmitter atlas for the connectome of *Drosophila melanogaster* and extract human interpretable features from the classifier, highlighting so far unknown phenotypical differences between neurotransmitter classes in *Drosophila melanogaster*.

[3] See Footnote 6

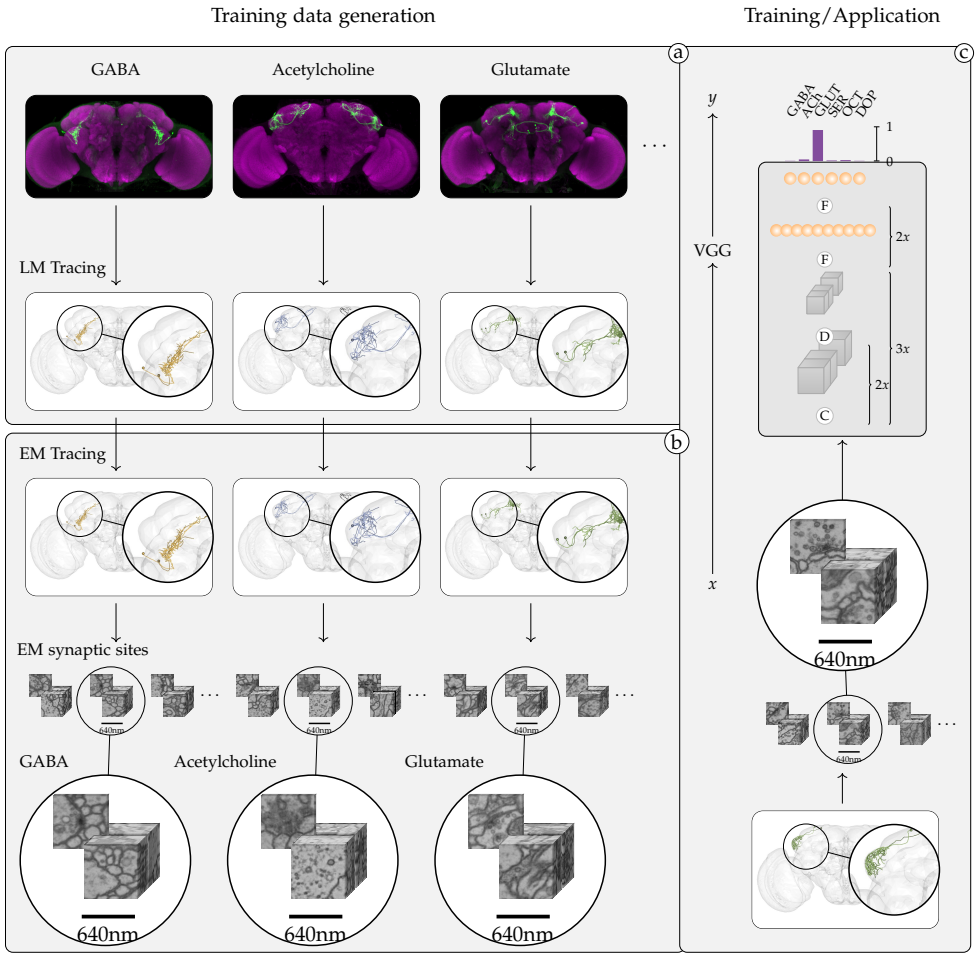


Figure 4.1: Method Overview. We assemble a dataset of neurons with known neurotransmitter in the *Drosophila* whole brain EM dataset (FAFB) (Zheng et al., 2018) from the literature and retrieve corresponding synaptic locations from the subset of skeletons that have been annotated in the CATMAID (Saalfeld et al., 2009; Schneider-Mizell et al., 2016) FAFB collaboration database[3]. (**a**) Typically, neurons are genetically tagged to identify their neurotransmitter identity and to reconstruct their coarse morphology using light microscopy. (**b**) Light microscopy tracings of neurons are then matched to corresponding neuron traces in the FAFB dataset, and synaptic locations are annotated, resulting in a data set of EM volumes of synaptic sites with known neurotransmitter identity. (**c**) We use the resulting pair (x, y), with x a 3D EM volume of a synaptic site and $y \in \{\text{GABA, ACh, GLUT, SER, OCT, DOP}\}$, the neurotransmitter of that synaptic site, to train a 3D VGG-style (Simonyan and Zisserman, 2014) deep neural network to assign a given synaptic site x to one of the six considered neurotransmitters. We use the trained network to predict the neurotransmitter identity of synapses from neurons with hitherto unknown neurotransmitter identity in the *Drosophila* FAFB dataset. C, D, and F denote convolution, downsampling, and fully connected layers respectively.

4.2 Methods

We learn a mapping $f : x \rightarrow y$, where x is a local 3D EM volume with a synaptic site at the center and y the neurotransmitter of the corresponding neuron (see Fig. 4.1). To this end, we need to generate a training dataset of pairs (x,y). This involves light microscopy of genetically tagged neurons to determine their neurotransmitter expression, neuron tracing and synapse annotation in the corresponding EM dataset and matching of the LM neuron morphology to a traced neuron in EM.

4.2.1 Data Acquisition

We acquire the majority of neurotransmitter to neuron assignments used for training and evaluation from published reconstructions in the full adult fly brain (*FAFB*) dataset (Bates et al., 2020; Dolan et al., 2018; Felsenberg et al., 2018; Frechter et al., 2019; Huoviala et al., 2018; Dolan et al., 2019; Marin et al., 2020; Sayin et al., 2019; Turner-Evans et al., 2019; Zheng et al., 2018), as well as unpublished but identified neuron reconstructions offered by the *FAFB* community (see acknowledgements). In these studies, the authors had already linked some of their reconstructed cell types to immunohistochemical data (Aso et al., 2014; Bräcker et al., 2013; Busch et al., 2009; Davis et al., 2018; Dolan et al., 2019; Ito et al., 2013; Lai et al., 2008; Okada et al., 2009; Shinomiya et al., 2015; Tanaka et al., 2012; Wilson and Laurent, 2005). Stainings were typically performed on neurons visualized by GFP expression in a GAL4/split-GAL4 line. Dissected brains are incubated with primary antibodies (*e.g.*, anti-VGlut, anti-GABA, anti-ChAT), followed by secondary antibodies which have a fluorescent tag to visualise the primary antibody. The transcripts/proteins related to certain transmitter expressions are thus labelled across the brain and if they colocalize with the GFP signal for the GAL4/split-GAL4 line of interest, those neurons are considered to express that transmitter[4]. For RNA transcripts, usually the neuron's soma is examined. Other methods involve RNA sequencing and include TAPIN-seq (Davis et al., 2018). Note that individual studies often only test single transmitters and do not show negative staining. As a result, there is limited data for cotransmission of multiple neurotransmitters in a single neuron and we therefore assume no cotransmission of neurotransmitters within one neuron.

We use manually reconstructed neuron annotations from the *FAFB* community CATMAID[5] (Saalfeld et al., 2009; Schneider-Mizell et al., 2016) database[6]. Synapses were annotated at presynaptic sites, defined by T-bars, vesicles

[4] A commonly used, full step-by-step protocol can be found at https://www.janelia.org/project-team/flylight/protocols.

[5] http://www.catmaid.org

[6] https://neuropil.janelia.org/tracing/fafb

	Neuron split	Hemilineage split	Brain Region split
Train	140,565	140,868	138,982
Test	40,104	40,703	39,715
Validation	20,084	19,182	19,858
Avg. Synapse Accuracy	87%	75%	88%
Avg. Neuron Accuracy	94%	92%	95%

Table 4.1: Overview of the three data splits used for evaluation of the classifier. Shown are the number of synapses for training, testing and validation as well as average synapse and neuron classification accuracy on the test set for each data split.

and a thick dark active zone by a synaptic cleft (Prokop and Meinertzhagen, 2006). In total, the assembled dataset contains 153,593 cholinergic synapses (679 neurons), 33881 GABAergic synapses (181 neurons), 7953 glutamatergic synapses (49 neurons), 9526 dopaminergic synapses (89 neurons), 2924 octopaminergic synapses (7 neurons) and 4732 serotonergic synapses (5 neurons).

4.2.2 Train and Test Datasets

For each neurotransmitter y $\in$ {GABA, ACh, GLUT, SER, OCT, DOP}, we divide the data in test, train and validation set by randomly assigning entire neurons, each containing multiple synapses. We refer to this split as *Neuron split* in the following. We use 70% of neurons for training, 10% for validation and the remaining 20% for testing. Splitting the dataset by entire neurons, instead of randomly sampling synapses, mirrors the real world use case in which we typically know the neurotransmitter of an entire neuron and are interested in the neurotransmitter of an entirely different neuron.

In order to test how well the proposed method generalizes across morphologically distinct cells and regions, and to exclude potentially confounding variables, we also generate two additional splits that separate the data by hemilineage (*Hemilineage split* - neurons in a hemilineage are lineally related, see section 4.4 for further details) and brain region (*Brain Region split*) respectively. To this end, we find the optimal split between entire hemilineages and brain regions, such that the fraction of synapses for every neurotransmitter in the train set is close to 80% of all synapses of that neurotransmitter. We further use randomly selected 12.5% of the training synapses (10% of the entire dataset) for validation.

4.2.3 Network Architecture and Training

We use a 3D VGG-style (Simonyan and Zisserman, 2014) network to predict the neurotransmitter identity from a 3D EM input cube of edge length 640 nm with a synaptic site at its center. The network consists of four functional blocks, each consisting of two 3D convolution operations, batch normalization, ReLU non-linearities and subsequent max pooling with a downsample factor of (z=1, y=2, x=2) for the first three blocks and (z=2, y=2, x=2) for the last block and is followed by three fully connected layers with dropout (p=0.5) applied after the first two. We train the network to minimize cross entropy loss over the six classes (GABA, ACh, GLUT, SER, OCT and DOP), using the Adam optimizer (Kingma and Ba, 2014). We train for a total of 500,000 iterations in batches of size eight and select the iteration with highest validation accuracy for testing. A full specification of the network architecture and training pipeline, including data augmentations, can be found in the appendix. For an illustration of the network architecture used see Fig. 4.1c.

4.3 Classifier accuracy

We tested the classifier on our held out test sets. For the *Neuron split*, the test set consists of a total of 40,104 synapses from 185 neurons that the network was not trained on. We achieve an average, per transmitter accuracy of 87% for the neurotransmitter prediction of individual synapses. Since we assume that each neuron expresses the same neurotransmitters at its synaptic sites we can additionally quantify the per neuron accuracy of the presented method. To this end we assign each neuron with more than 30 synapses in the test set a neurotransmitter by majority vote of its synapses, leading to an average accuracy of 94% for the neurotransmitter prediction per neuron. For the *Hemilineage split*, we find an accuracy of 75% for individual synapses and 92% for entire neurons. The *Brain Region split* evaluates to 88% synapse classification accuracy and 95% neuron classification accuracy. A per class overview can be seen in Fig. 4.3, for a summary of the results and data splits see Table 4.1.

4.3.1 Hemibrain

In addition to the *FAFB* dataset we also train a classifier to predict synaptic neurotransmitter identity in the so called *Hemibrain* (Xu et al., 2020), an isotropic 8 x 8 x 8 nm resolution FIB-SEM dataset containing roughly 50% of the full Drosophila brain. We adapt the architecture to match the field of view of 640nm^3 in the *FAFB* dataset and train on a set of 219,971 manually annotated synapses. We test on a held out dataset containing 63,438 synapses in 207 neurons. We achieve an average synaptic accuracy of 79%

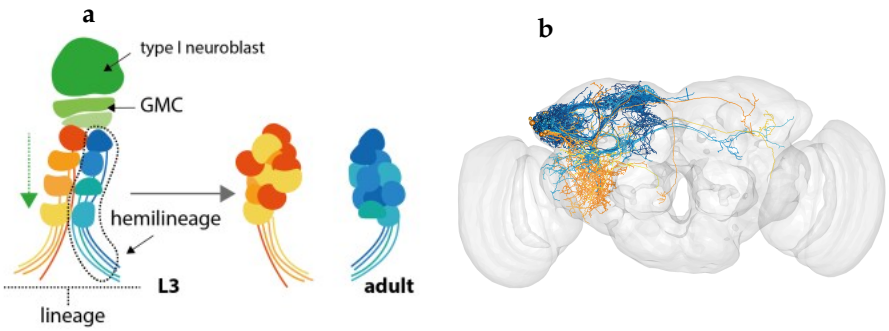


Figure 4.2: Illustration of (**a**) the progression of a Type I neuroblast from third-instar (L3) larva into the adult, GMC, ganglion mother cell and (**b**) breakdown of a single secondary lineage, LHl2 (also known as DPLal2) into its two hemilineages. Neuronal reconstruction data from the FAFB project shown, which covers half the neurons in this lineage.

and average neuron accuracy of 90% for the skeleton split, showing that predicting neurotransmitter identity from EM is possible on diverse EM datasets (Fig.4.4 a). The reduction in average accuracy compared to our results on *FAFB* could be explained by the lower lateral resolution of 8 nm (see Fig.4.4 b for images of synaptic sites from both datasets).

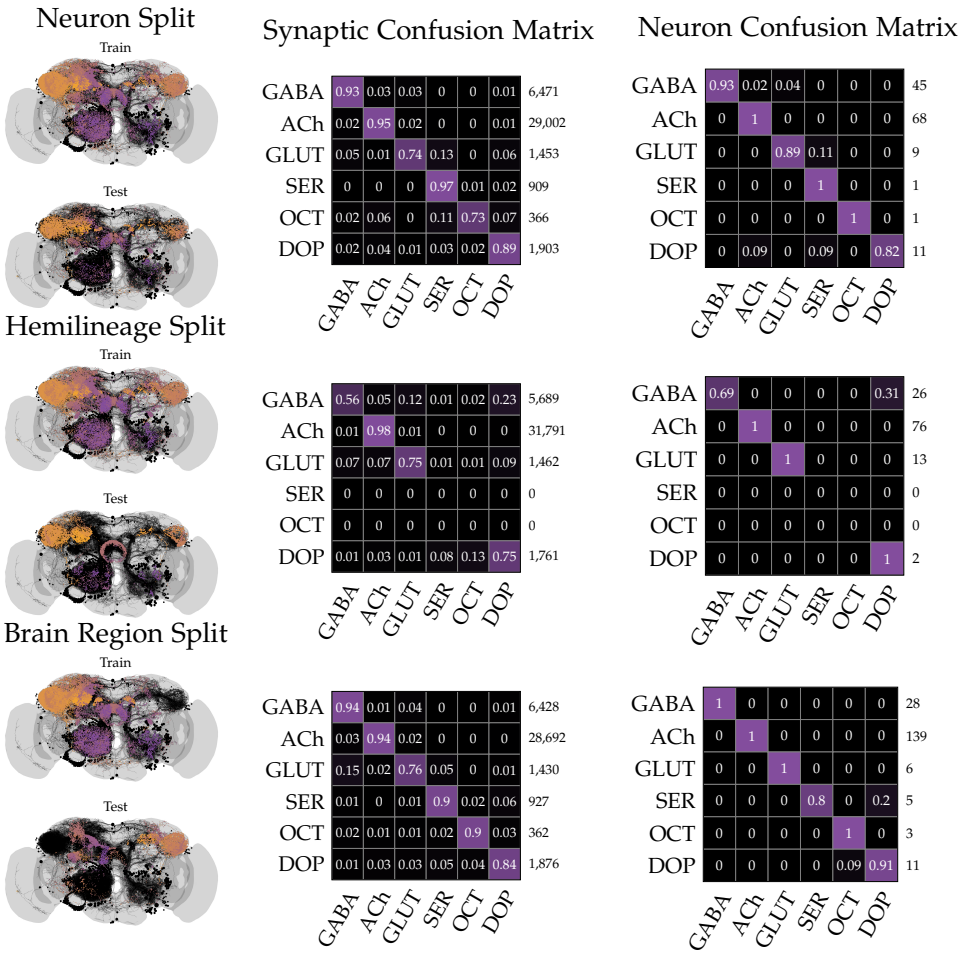


Figure 4.3: Illustration of the spatial distribution of synapses in the considered data splits (left column) and corresponding confusion matrices for synapses (middle column) and entire neurons (right column). **Datasets:** For each split we visualize the synaptic locations used for training (top) and for testing (bottom). Synapse locations are color coded according to their z-depth (perpendicular to viewing plane). **Confusion Matrices:** Rows show labels and columns the predicted neurotransmitter. The total number of test set ground truth synapses and neurons respectively are shown next to each row. In order to be able to have a meaningful majority vote we only consider neurons with more than 30 synapses for the neuron confusion matrices.

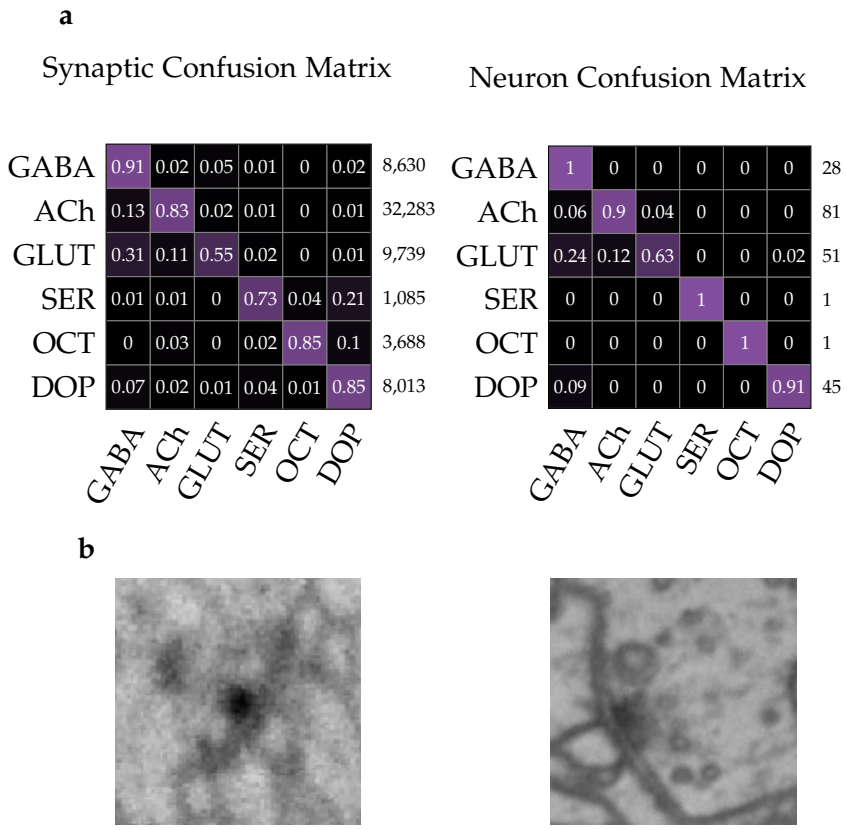


Figure 4.4: **a)** Synaptic (left) and neuronal (right) confusion matrix on the *Hemibrain* skeleton split test set. **b)** Appearance of synaptic sites in the *Hemibrain* (left) and *FAFB* (right) datasets, highlighting differences in resolution. Both synapses are from GABAergic neurons.

4.4 Transmitter prediction for hemilineages

Similar to neurons, which release the same set of neurotransmitters at their synaptic sites (Eccles, 1976; Dale, 1934), it has been found that sets of lineally related neurons in the *Drosophila melanogaster* ventral nervous system (VNS), so-called hemilineages, also show homogeneous neurotransmitter expression patterns (Lacin et al., 2019). If a similar principle holds for the *Drosophila melanogaster* brain, it would enable us to assign neurotransmitter identity to large groups of neurons simultaneously. With the presented method we are able to verify to what extent such a principle holds.

The roughly 45,000 neurons of the central brain of *Drosophila melanogaster* (Croset et al., 2018) are generated by a set of stem cells known as neuroblasts. During division neuroblasts generate two cells, one additional stem cell and

one cell that further divides into two sibling neurons. In only one of these siblings the so called Notch pathway is activated, leading to two different "hemilineages" of neurons within the lineage (Kumar et al., 2009; Sen, 2019; Lacin et al., 2019). Lacin et al. (2019) showed that each hemilineage in the VNS expresses just one of the fast-acting transmitters acetylcholine, glutamate and GABA, even though mRNA transcripts for combinations of these can appear in the nucleus (Lacin et al., 2019). This raises the question whether the same holds true in the adult brain. Using the presented classifier, we predict the neurotransmitter identity of all identified neurons within 89 out of a total of ˜150 identified hemilineages in the *Drosophila melanogaster* brain. The majority of our predictions show homogeneity of neurotransmitter identity within a single hemilineage, in line with findings in the VNS. However, we identify a set of hemilineages which express two fast acting neurotransmitters with high statistical significance. We find no hemilineage that expresses all three. As a result, our predictions are inconsistent with the hypothesis that all hemilineages express the same fast-acting neurotransmitter.

4.4.1 Hemilineage assignments in Drosophila

Cell body fiber tracts for identified hemilineages had previously been identified using TrakEM2 (Cardona et al., 2012) in a light-level atlas for a *Drosophila melanogaster* brain, stained with an antibody against neurotactin (BP104) (Lovick et al., 2013). We extracted these expertly identified tracts and registered them into a common template brain, JFRC2, using CMTK (Rohlfing and Maurer, 2003), and then into *FAFB* space (Bates et al., 2019b). This enabled us to identify cell body fibre tracts in this ssTEM dataset in the vicinity of the transformed hemilineage tracts.

4.4.2 Predictions

We retrain the classifier on 90% of the entire dataset and use the remaining 10% to select the best performing iteration. We predict the neurotransmitter identity of 180,675 synapses within 1,164 neurons with previously unknown neurotransmitter identity. These neurons come from a total of 89 hemilineages, of which 20 have more than one neuron with genetically identified neurotransmitter. Fig. 4.5 shows ground truth neurotransmitter annotations for the subset of neurons N^h_{gt} that have known neurotransmitters and our predictions for the remaining neurons N^h_{pred} ($N^h_{\text{gt}} \cap N^h_{\text{pred}} = \emptyset$) in the hemilineage for five selected hemilineages. In the following, we analyse the results by quantifying how neurotransmitter predictions are distributed over neurons and synapses within hemilineages.

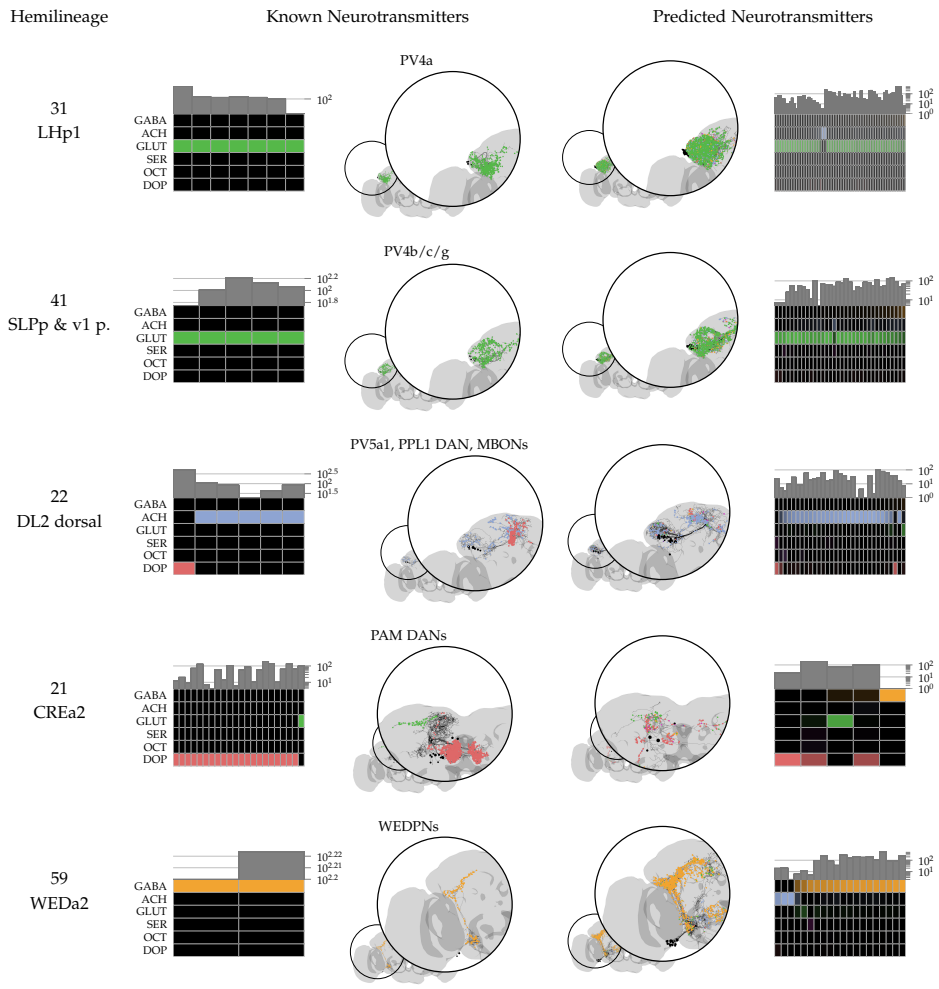


Figure 4.5: Neurotransmitter barcode plots and corresponding renderings of skeletons and synapses (color coded according to their neurotransmitter identity) of five selected hemilineages, for which a subset of neurons N^h_{gt} have genetically determined, known neurotransmitters (left) and our predictions on the remaining neurons N^h_{pred} in the same hemilineage (right). Each column in the neurotransmitter barcode represents one neuron. For each neuron (column), the relative number of synapses with neurotransmitter $y, \hat{y} \in Y = \{\text{GABA, ACh, GLUT, SER, OCT, DOP}\}$ is represented by the color intensity of the respective row. The total number of synapses in each neuron is shown above each row. Note that $N^h_{\text{gt}} \cap N^h_{\text{pred}} = \emptyset$. Neuron classes, shown above the inset, are given for the most numerous cells in our training data, for each hemilineage that we show (Dolan et al., 2019; Otto et al., 2020; Aso et al., 2014). For an overview of all hemilineages that have partially known neurotransmitter identities and our associated predictions, see Fig. A.14.

Neuron Level Entropy

In order to quantify multimodality of neurotransmitter predictions on a neuron level within a hemilineage we calculate the entropy H of the neurotransmitter distribution over neurons in the following way: Let $n \in N_h$ be a neuron in hemilineage h and $\hat{y}_n \in Y = \{\text{GABA, ACh, GLUT, SER, OCT, DOP}\}$ the predicted neurotransmitter of neuron n. Then

$$H(N_h) = -\sum_{y \in Y} p_h(y) \log_6 p_h(y) \text{ , with} \tag{4.1}$$

$$p_h(y) = \frac{1}{|N_h|} \sum_{n \in N_h} \delta(\hat{y}_n = y) \tag{4.2}$$

A value of $H(N_h) = 0$ (minimal entropy) then means that all neurons within hemilineage h have the same predicted neurotransmitter, while a value of $H(N_h) = 1$ (maximal entropy) means that within hemilineage h all predicted neurotransmitters are equally common.

Synapse Level Entropy

Similarly we can quantify the average multimodality over synapses within neurons of a given hemilineage: Let $s \in S_n$ be the synapses in neuron $n \in N_h$ of hemilineage h and $\hat{y}_s$ the predicted neurotransmitter. The entropy of predicted synaptic neurotransmitters $H(s_n)$ in neuron n is then given by:

$$H(S_n) = -\sum_{y \in Y} p_n(y) \log_6 p_n(y) \text{ , with} \tag{4.3}$$

$$p_n(y) = \frac{1}{|S_n|} \sum_{s \in S_n} \delta(\hat{y}_s = y) \tag{4.4}$$

With this, the average synaptic entropy over all neurons within hemilineage h is given by:

$$H(S_h) = \frac{1}{|N_h|} \sum_{n \in N_h} H(S_n) \tag{4.5}$$

A value of $H(S_h) = 0$ (minimal entropy) then means that all synapses of all neurons in hemilineage h have the same predicted neurotransmitter, while a value of $H(S_h) = 1$ (maximal entropy) means that in all neurons within hemilineage h all synaptic neurotransmitter predictions are equally common. Fig. 4.6 shows the distribution of $H(N_h)$ and $H(S_h)$ of all predicted hemilineages with more than ten neurons that have more than 30 synapses each.

On the population level we find relatively lower values of $H(S_h)$ (Synapse level entropy) than $H(N_h)$ (Neuron level entropy). 75% of hemilineages

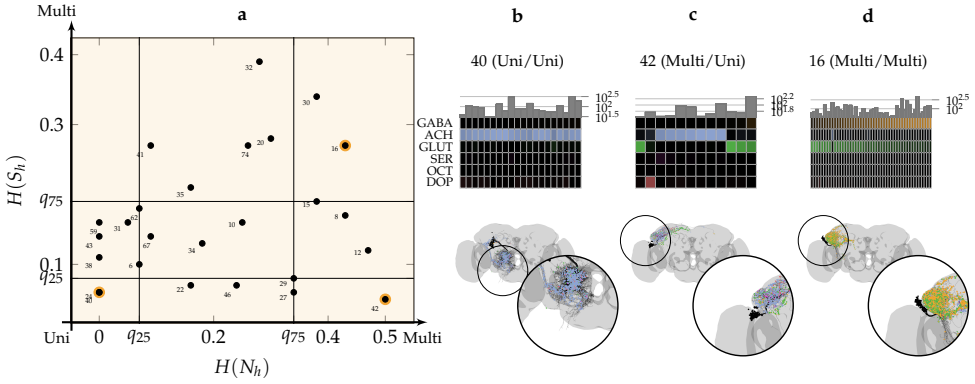


Figure 4.6: (**a**) Neuron level entropy vs average synapse level entropy for all predicted hemilineages with more than 10 neurons and more than 30 synapses per neuron. $q_{25}(H(N_h)) = 0.07$, $q_{25}(H(S_h)) = 0.08$ and $q_{75}(H(N_h)) = 0.34$, $q_{75}(H(S_h)) = 0.19$ are indicating 25% and 75% percentiles respectively. Highlighted hemilineages show the extreme points of the entropy distribution: (**b**) Hemilineage 40 shows low neuron level entropy $H(N_h)$ and low synapse level entropy $H(S_h)$, caused by a unimodal distribution of neurotransmitters on synapse and neuron level in this hemilineage. (**c**) Hemilineage 42 shows high neuron level entropy $H(N_h)$ but low synapse level entropy $H(S_h)$, caused by neurotransmitter predictions that are unimodal within each neuron but multimodal across neurons. (**d**) Hemilineage 16 shows high neuron level entropy $H(N_h)$ and high synapse level entropy $H(S_h)$, as a result of a bimodal distribution of neurotransmitter predictions within most neurons of this hemilineage. For a mapping of hemilineage ID to hemilineage name see Table A.8.

show a synapse level entropy below $q_{75}(H(S_h)) = 0.19$ as compared to $q_{75}(H(N_h)) = 0.34$. This is reassuring as it suggests less variation of neurotransmitter identity predictions within individual neurons compared to variations of neurotransmitter identity of neurons within a hemilineage. However, we also find cases with a high level of synaptic entropy, such as hemilineage 16 and 30. For these hemilineages it is unclear whether neuron level multimodality is only an artifact of uncertain, multimodal predictions on synapse level of individual neurons. In contrast to 16 and 30 hemilineages 29, 27 and 42 show high neuron level entropy $H(N_h) \geq q_{75}$ but low synapse level entropy $H(S_h) \leq q_{25}$, suggesting clear neuron level segregation of predicted neurotransmitters within those hemilineages. Hemilineages such as 24 and 40 with $H(S_h) < q_{25}$ and $H(S_n) < q_{25}$ appear homogeneous within each neuron and within the entire hemilineage.

4.4.3 Number of distinct, fast-acting neurotransmitters in hemilineages of the *Drosophila melanogaster* adult brain

We can now ask the question how likely it is to observe a given prediction of neurotransmitters in a hemilineage under some error rate given by the confusion matrix on the test set, and the assumption that all neurons in the hemilineage have the same underlying neurotransmitter. We can then compare this likelihood to the alternative hypothesis that a hemilineage consists of neurons with more than one neurotransmitter. Out of 26 investigated hemilineages with a sufficient number of predicted neurotransmitters, up to five show strong evidence for expressing two distinct, fast-acting neurotransmitters (Bayes factor $K \geq 10^2$, *decisive*). We find none that expresses all three.

Probability to observe neurotransmitter predictions $\hat{y}$

Given a neuron has true neurotransmitter $y \in Y$, the probability that we predict neurotransmitter $\hat{y} \in Y$ (assuming that each prediction is independent and identically distributed) is given by the categorical distribution

$$p(\hat{y}|y) = C_{y,\hat{y}} \tag{4.6}$$

where C is the neuron confusion matrix obtained on the test data set (see Fig. 4.3).

Let m be the number of different neurotransmitters in hemilineage h. We model the probability $p(\hat{\mathbf{y}}|m)$ of observing neurotransmitter predictions $\hat{\mathbf{y}} = \{\hat{y}_0, \hat{y}_1, ..., \hat{y}_n\}$ under the assumption that hemilineage h contains m different neurotransmitters. Here, $\hat{y}_j$ is the predicted neurotransmitter of neuron j in hemilineage h with n neurons total. Let $\mathbb{P}_c(Y)$ be the set of subsets of true neurotransmitters Y with cardinality c, then:

$$p(\hat{\mathbf{y}}|m) = \sum_{S \in \mathbb{P}_m(Y)} p(\hat{\mathbf{y}}|S) \cdot p(S|m), \tag{4.7}$$

where $p(\hat{\mathbf{y}}|S)$ is the probability to observe predictions $\hat{\mathbf{y}}$ if the hemilineage has true underlying neurotransmitters $y \in S$ and $p(S|m)$ is the probability for the set of true neurotransmitters S given the hemilineage contains m different neurotransmitters. Since we assume i.i.d. predictions $\hat{\mathbf{y}}$, $p(\hat{\mathbf{y}}|S)$ factorizes as follows:

$$p(\hat{\mathbf{y}}|S) = \prod_j p(\hat{y}_j|S) \tag{4.8}$$

and marginalizing over $y \in S$ yields:

$$p(\hat{\mathbf{y}}|S) = \prod_j \sum_{y \in S} p(\hat{y}_j|y) \cdot p(y|S) \tag{4.9}$$

$$= \prod_j \sum_{y \in S} C_{y,\hat{y}_j} \cdot p(y|S) \tag{4.10}$$

Regarding $p(S|m)$ and $p(y|S)$ we assume a flat prior, *i.e.*:

$$p(S|m) = \binom{|Y|}{m}^{-1} \tag{4.11}$$

$$p(y|S) = \frac{1}{|S|} = \frac{1}{m} \tag{4.12}$$

With this, the probability of observing predictions $\hat{\mathbf{y}}$ given m different neurotransmitters becomes:

$$p(\hat{\mathbf{y}}|m) = \binom{|Y|}{m}^{-1} \sum_{S \in \mathbb{P}_m(Y)} \left(\prod_j \sum_{y \in S} C_{y,\hat{y}_j} \cdot \frac{1}{|S|} \right) \tag{4.13}$$

Bayes Factor

With this formalism in place, we can compare hypotheses about the number of true neurotransmitters m in a given hemilineage by using the Bayes Factor $K = \frac{p(D|M_1)}{p(D|M_2)}$, where D is our observed data (predicted neurotransmitters) and M_1, M_2 are two models about the underlying true neurotransmitters that we wish to compare. The Bayes factor for a model M_1 with m_1 true neurotransmitters per hemilineage and model M_2 with m_2 different neurotransmitters is given by:

$$K = \frac{p(\hat{\mathbf{y}}|m_1)}{p(\hat{\mathbf{y}}|m_2)} \tag{4.14}$$

$$= \frac{\binom{|Y|}{m_1}^{-1} \sum_{S \in \mathbb{P}_{m_1}(Y)} \left(\prod_j \sum_{y \in S} C_{y,\hat{y}_j} \cdot \frac{1}{m_1} \right)}{\binom{|Y|}{m_2}^{-1} \sum_{S \in \mathbb{P}_{m_2}(Y)} \left(\prod_j \sum_{y \in S} C_{y,\hat{y}_j} \cdot \frac{1}{m_2} \right)} \tag{4.15}$$

So far, we assumed that $p(\hat{y}_j|y) = C_{y,\hat{y}_j}$, *i.e.*, we estimate this distribution on the test dataset. However, because our test set is finite we cannot expect that the estimated error rates perfectly transfer to other datasets. In order to relax our assumptions about this distribution we simulate additional errors, by incorporating additive smoothing on the counts of neurons $N_{y,\hat{y}}$ that have true neurotransmitter y and were predicted as neurotransmitter $\hat{y}$, *i.e.*:

$$\tilde{C}_{y,\hat{y}} = \frac{N_{y,\hat{y}} + \beta}{\sum_{\hat{y}} N_{y,\hat{y}} + 6\beta}, \tag{4.16}$$

where $\beta \in \mathbb{N}_0$ is the smoothing parameter. With $C_{y,\hat{y}} = \frac{N_{y,\hat{y}}}{\sum_{\hat{y}} N_{y,\hat{y}}}$ we then have

$$\tilde{C}_{y,\hat{y}} = \frac{C_{y,\hat{y}} + \frac{\beta}{\sum_{\hat{y}} N_{y,\hat{y}}}}{1 + 6\frac{\beta}{\sum_{\hat{y}} N_{y,\hat{y}}}} = \frac{C_{y,\hat{y}} + \alpha}{1 + 6\alpha} \tag{4.17}$$

and $\alpha \in \mathbb{R}_{\geq 0}$ the count normalized smoothing parameter. In the limit of $\alpha \to \infty$, $\tilde{C}_{y,\hat{y}}$ approaches the uniform distribution with probability $1/6$ for each neurotransmitter, whereas a value of $\alpha = 0$ means we recover the observed confusion matrix C. With this, our distributions are now parametrized by α and the Bayes factor becomes:

$$K = \frac{\int_\alpha p(\hat{\mathbf{y}}, \alpha | m_1) p(\alpha) d\alpha}{\int_\alpha p(\hat{\mathbf{y}}, \alpha | m_2) p(\alpha) d\alpha} \tag{4.18}$$

$$= \frac{\tilde{p}(\hat{\mathbf{y}}|m_1)}{\tilde{p}(\hat{\mathbf{y}}|m_2)} \tag{4.19}$$

$$\tag{4.20}$$

where $\tilde{p}(\hat{\mathbf{y}}|m)$ is as defined in (4.13) but with $C_{y,\hat{y}_j}$ replaced with its expected value $\mathbb{E}_{p(\alpha)}[\tilde{C}_{y,\hat{y}_j}]$.

The prior distribution on α, $p(\alpha)$, allows us to encode our prior knowledge about α and use it to weight the likelihood of the corresponding model. Given the data, a value of $\alpha = \epsilon$ with epsilon small ($0 < \epsilon \ll 1$), should be most probable, while the probability of values $\alpha > \epsilon$ should monotonically decrease as we deviate more from the observed confusion matrix. Values of $\alpha < \epsilon$ should have probability zero, because they correspond to the unsmoothed confusion matrix with zero entries, *i.e.*, a probability of zero for misclassification of certain neurotransmitters. While these probabilities may be small, they are likely greater than zero and an artifact caused by the finite test set. Many distributions fulfill these criteria, in particular the family of exponential distributions with rate parameter λ:

$$p(\alpha) = \begin{cases} \lambda e^{-\lambda(\alpha - \epsilon)} & \alpha \geq \epsilon \\ 0 & \alpha < \epsilon \end{cases}$$

Thus, λ controls the weight for smoothing parameter α in the integral $\mathbb{E}_{p(\alpha|M)}[\tilde{C}(\alpha)_{y,\hat{y}_j}] = \int_\alpha \tilde{C}_{y,\hat{y}_j} p(\alpha) d\alpha$. For $\lambda \to 0$, the expected confusion matrix approaches the unweighted average of all $C(\alpha)$ in the integration range. For $\lambda \to \infty$, the expected confusion matrix approaches the ϵ-smoothed confusion matrix $\tilde{C}_{y,\hat{y}} = \frac{C_{y,\hat{y}} + \epsilon}{1 + 6\epsilon}$. The rate parameter λ can also be understood via its influence on the expected average accuracy $\tilde{c}_{\text{exp}} = \frac{1}{6} \sum_i \mathbb{E}_{p(\alpha|M)}[\tilde{C}]_{i,i}$. For values of $\lambda \to 0$, the expected accuracy approaches chance level while for values of $\lambda \to \infty$, the expected accuracy approaches the ϵ-smoothed, observed accuracy on the test set.

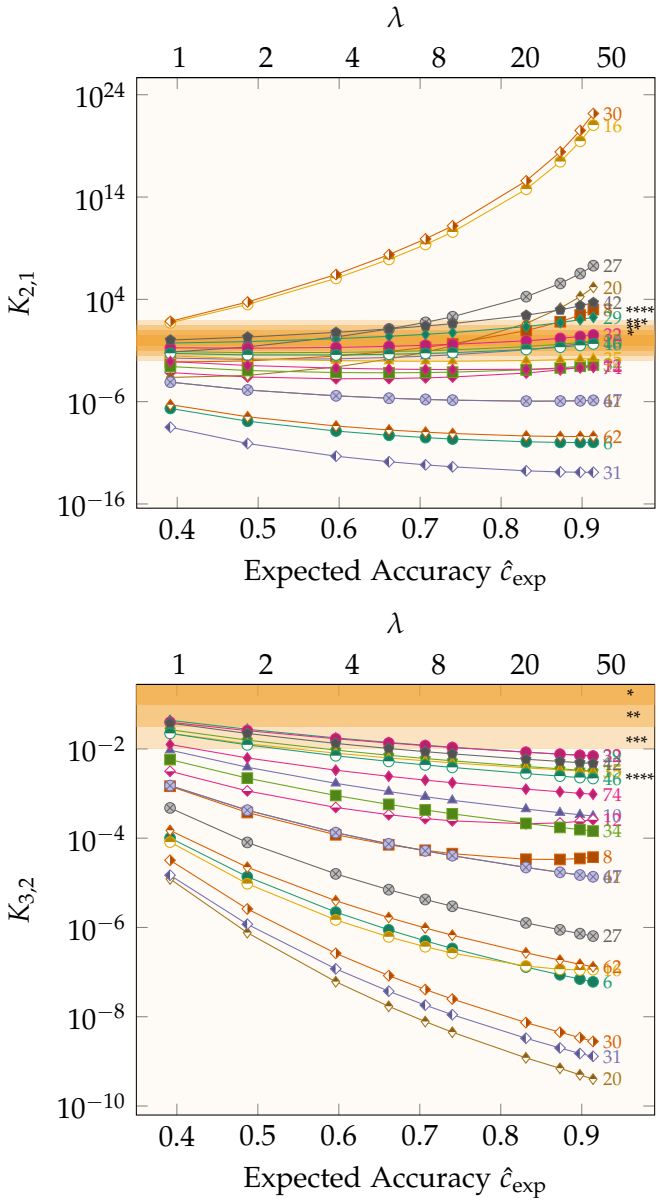


Figure 4.7: Bayes factor K over a range of expected average predictor accuracy $\tilde{c}_{\text{exp}}$. Shown are hemilineages with more than ten neurons that have more than 30 synapses each and more than one predicted, fast-acting neurotransmitter. Stars indicate regions of evidence for model M_1 ($K > 1$) or model M_2 ($K < 1$), respectively: * - substantial, ** - strong, *** - very strong, **** - decisive (Jeffreys, 1998). **Top:** M_1: $m_1 = 2$ and M_2: $m_2 = 1$. **Bottom:** M_1: $m_1 = 3$ and M_2: $m_2 = 2$. For a mapping of hemilineage ID to hemilineage name see Table A.8.

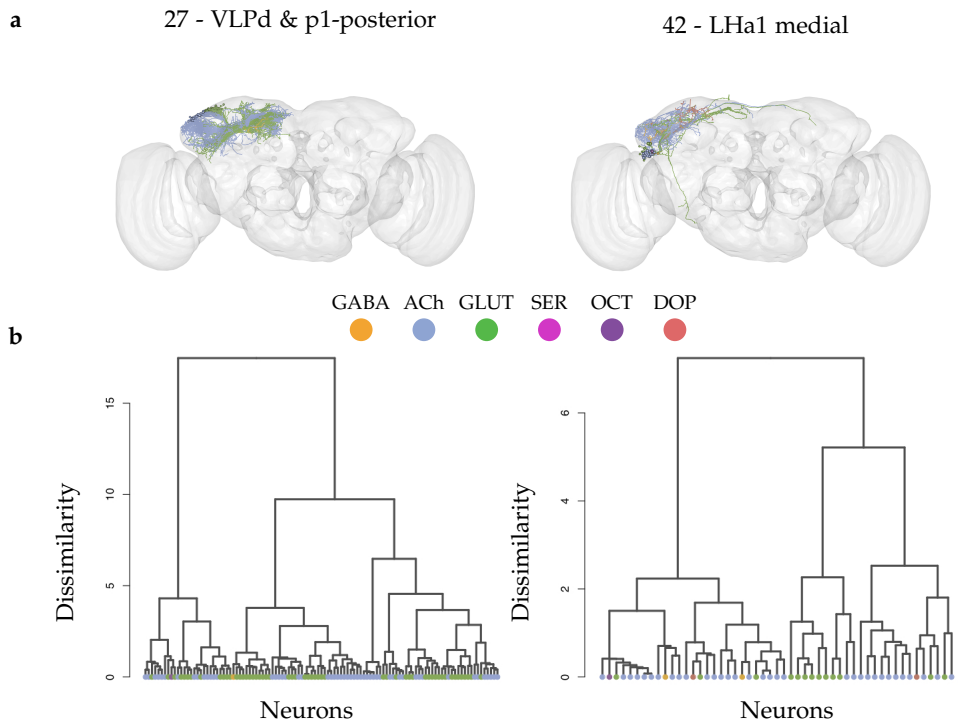


Figure 4.8: Renderings of neurotransmitter predictions of all neurons within two selected hemilineages that show a high Bayes factor $K_{2,1}$ (see Fig. 4.7) in addition to low synaptic entropy $H(S_h)$ (see Fig. 4.6) (**a**), and corresponding NBLAST dendrograms (**b**). Y-axis shows the morphological dissimilarity between clusters, based on Ward's method. Each neuron is color coded according to its predicted, majority vote neurotransmitter. The dendrograms show that neurotranmsitter predictions correlate strongly with neuron morphology for hemilineage 27 and to a lesser degree for hemilineage 42. For renderings of all hemilineages and corresponding barcode plots see Fig. A.13 and Fig. A.12.

Evidence for two fast acting transmitters in a single hemilineage

We calculate the Bayes factor $K_{2,1} = \frac{p(\hat{y}|m=2)}{p(\hat{y}|m=1)}$ and $K_{3,2} = \frac{p(\hat{y}|m=3)}{p(\hat{y}|m=2)}$ for the set of three classical neurotransmitters $Y_{\text{cl}} = \{\text{GABA, ACh, GLUT}\}$ for those hemilineages that have more than ten annotated neurons and 30 annotated synapses each with neurotransmitter predictions in the set Y_{cl}. For this analysis, we ignore all other neurons with predicted neurotransmitter identity $\hat{y} \notin Y_{\text{cl}}$. Fig. 4.7 shows $K_{2,1}$ and $K_{3,2}$ for a range of rate parameters λ and corresponding expected average accuracy $\tilde{c}_{\text{exp}}$. For hemilineages 30, 16, 27, 20, and 42 there is decisive evidence ($K_{2,1} \geq 10^2$) for the presence of two distinct fast acting neurotransmitters for a large range of expected accuracies $\tilde{c}_{\text{exp}}$. However, note that hemilineage 30 and 16 show high synaptic entropy $H(S_h)$ (see Fig. 4.6), indicating that individual neurons within the

hemilineage already show bimodal neurotransmitter predictions. As such, strong bimodality at the neuron level is at least partially explained by uncertain predictions at the individual synaptic level. This is in contrast to hemilineages 27 and 42, who show synaptic entropies below the 25% percentile. In these hemilineages, large Bayes factor values $K_{2,1}$ directly stem from neuron level segregation of the predicted neurotransmitters within the hemilineage. See Fig. 4.8 for a rendering of the neurotransmitter predictions of these hemilineages and corresponding NBLAST dendrograms, indicating that the two fast acting neurotransmitters in some of these hemilineages are divided between morphologically distinct neurons. The remaining 13 hemilineages show no strong evidence for either hypothesis ($K_{2,1} \approx 1$, n=5) or favor the hypothesis of expressing only one fast acting neurotransmitter ($K_{2,3} \leq 10^{-2}$, n=8). No hemilineage shows evidence for expressing all three fast acting neurotransmitters ($K_{2,3} < 10^{-2}$).

4.5 Whole Brain Predictions

We use the classifier to predict the neurotransmitter identity of all automatically detected synapses in *FAFB* (Buhmann et al., 2019). We use the same network as described in Section 4.4. We validate whether neurotransmitter predictions from automatically detected pre-synaptic sites are robust by predicting the neurotransmitter of neurons within 6 cell types with known neurotransmitters and find perfect agreement for all neurons with conclusive neurotransmitter predictions (Fig. 4.9 b). A neuron is defined to have conclusive predictions if the predicted, synaptic, majority neurotransmitter constitutes more than 65% of all synapses of that neuron. This threshold corresponds to a >95% accuracy for all three fast acting transmitters on the test set (see Fig. 4.9 a). Prediction of all $\approx$ 220 Million pre-synaptic sites took 3 days on 20 GPUs and 100 CPUs. 22% of all synapses are predicted as GABA, 47.5% as ACh, 19.2% GLUT, 3.9% SER, 2.2% OCT and 5.2% DOP, consistent with experimental results from large scale RNA-seq (Croset et al., 2018). Furthermore we recover known anatomical features of neurotransmitter distribution such as a GABA rich ellipsoid body, dopamine rich mushroom body lobes and expression symmetry in the hemispheres. For an overview see Fig. 4.9 c.

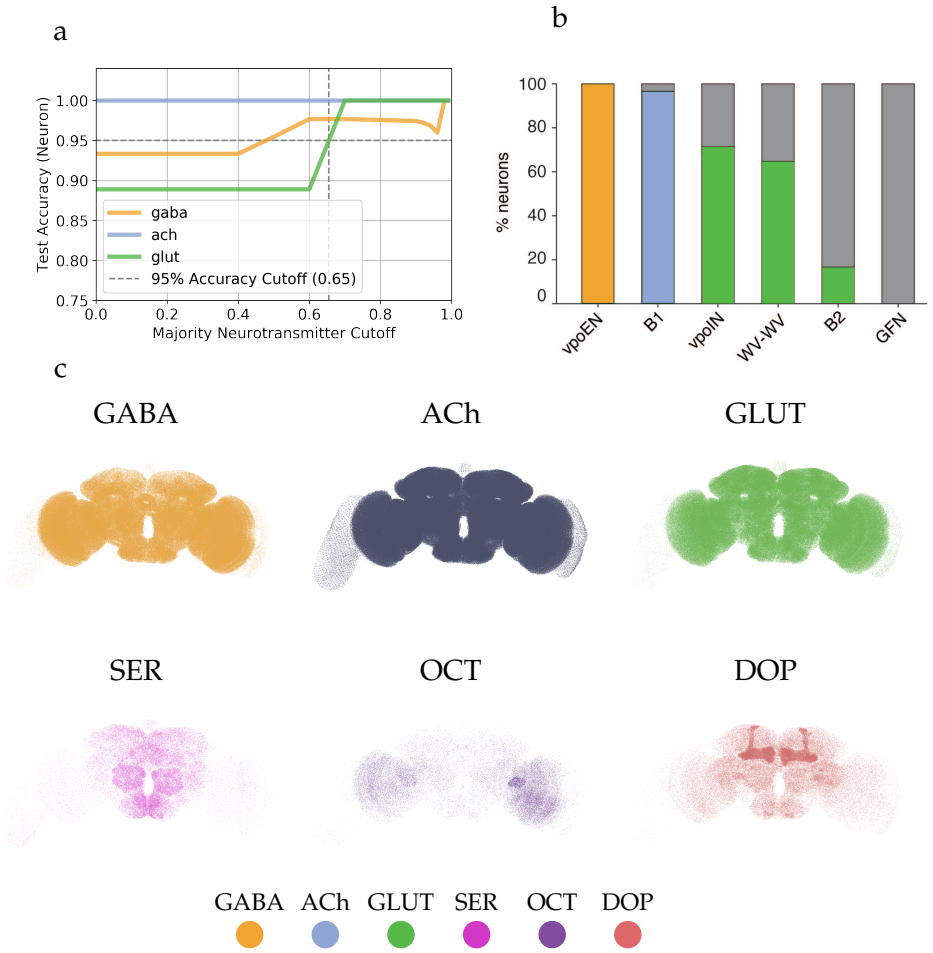


Figure 4.9: Whole brain neurotransmitter prediction from automatically detected synapses. **a)** Neurotransmitter neuron classification accuracy on the test set as a function of the fraction of synapses in a neuron below which we consider the neuron inconclusive. For whole brain prediction accuracy evaluation we use a cutoff of 0.65 (grey line), corresponding to a >95% neuron classification accuracy on the test set for the three fast acting transmitters. **b)** Shown is the fraction of neurons that have been predicted as a particular neurotransmitter using automatically detected synapses for 6 different cell types in *FAFB* with known neurotransmitters. Neuron classes are colored according to their known neurotransmitter identity. **c)** Frontal view of whole brain neurotransmitter distributions for all six considered neurotransmitters. We only show 10 million randomly chosen synapses from the total of 220 million for visualization purposes. Note that the predictions recover already known morphological features such as a dopamine rich mushroom body lobes, a GABA enriched ellipsoid body and neurotransmitter expression symmetry between the two hemispheres. Panels a) and b) adapted from Baker et al. (2021). Visualizations in panel c) were created by Gregory S.X.E. Jefferis.

4.6 Data Availability

For on demand predictions, we make the data and the classifier available to the community by providing a public website that allows a user to request predictions. The service supports coordinate queries from *FAFB* and the *Hemibrain* datasets in all three major service coordinates: CATMAID (Saalfeld et al., 2009), FlyWire (Dorkenwald et al., 2020) and Neuprint (Xu et al., 2020). In addition we provide the possibility to query from skeleton- (CATMAID) or body-ids (Neuprint) and request predictions for traced neurons that have annotated synapses in either one of these services. Because FlyWire neuron segmentation data is not public we cannot support query by neuron identifier from FlyWire. After processing of a users request, we send out a neurotransmitter prediction report, detailing the predicted neurotransmitter fractions for each neuron as well as provide the raw input and output data of the classifier for further analysis. In addition, we provide a SQL database dump of the *FAFB* whole brain predictions and make it accessible via a query module available within the natverse library[7].

4.7 Interpretability

In *Drosophila melanogaster* humans can not generally distinguish different neurotransmitter containing vesicles from EM alone, thus we would like to know how the presented classifier is doing it and which features it relies on. Despite the fact that we have complete information about the artificial neural system that is able to assign images of synaptic sites to neurotransmitters, it is difficult to extract the rules under which it operates. Here we use *DAC*, a novel neural network interpretability method introduced in the prior chapter, to generate a hypothesis matrix, detailing the feature differences between each pair of neurotransmitter classes as learned by the classifier.

4.7.1 Hypothetical Discriminators

In order to simplify the problem and to be able to use *DAC*, we train a 2D VGG f_{2D} to perform neurotransmitter classification from a 128 x 128 pixel large cropout around the synapse. While accuracy is reduced compared to using a 3D classifier, we still achieve 77% average synaptic classification accuracy. This implies that the classifier picked up on the most important features of each class. For each input image x_r (real) of a synaptic site with neurotransmitter nt_A, *DAC* generates another image of a synaptic site x_f (Fake), showing how the synaptic site would look if it would express neurotransmitter nt_B. In addition, *DAC* outputs a binary mask A_{min} of pixels in the image, that are most important for the classifier's decision. In or-

[7] http://natverse.org

der to understand the decision boundaries learned by the classifier, we thus look for features in the set of the top 40 image pairs with smallest area of importance A_{min} for each pair of neurotransmitters nt_A, nt_B. We only accept sample pairs if the real image x_r and the fake image x_f are classified correctly, i.e. $f_{2D}(x_r) > 0.9$ and $f_{2D}(x_f) > 0.9$. For each pair we consider both directions $\text{nt}_A^r \rightarrow \text{nt}_B^f$ and $nt_B^r \rightarrow nt_A^f$ and note an observed change in features as a hypothetical discriminator, if the feature consistently changes in one direction, and is symmetrically reversed when going in the other. For example, if we observe a darkening of the cleft from $\text{nt}_A^r \rightarrow \text{nt}_B^f$, we require the cleft to broaden going from $\text{nt}_B^r \rightarrow \text{nt}_A^f$. This ensures that we do not pick up on features that are not present in real synapses. All symmetric, hypothetical discriminators can be seen in Fig. 4.10.

The most notable findings are that the classical transmitters GABA, glutamate and acetylcholine look different in very subtle ways. For example we observe a consistent brightening of the inside of the synaptic cleft going from GABA to acetylcholine and slightly enlarged vesicles going from GABA to glutamate. Changes from acetylcholine to glutamate are a darker T-bar and a darker cleft. Other notable features are the apparent removal of post synaptic densities going from acetylcholine and glutamate to dopamine, in line with findings in mammalian cells (Uchigashima et al., 2016). A known discriminator we were able to rediscover is the addition of dense core vesicles when going from the classical transmitters to serotonin and octopamine. We leave confirmation of these hypotheses for future work.

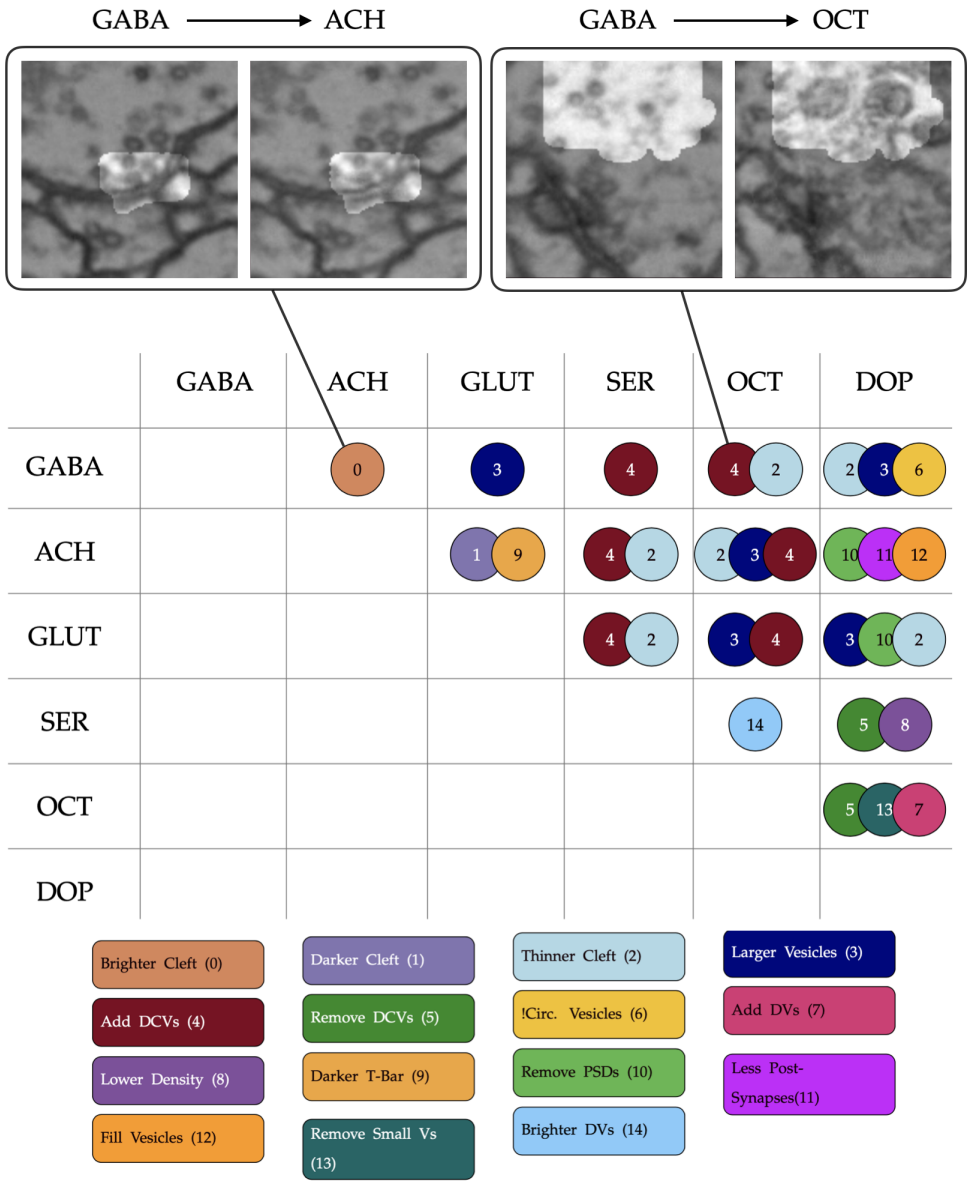


Figure 4.10: Hypothetical Discriminator Matrix of features that change between images of two different neurotransmitter classes. We only show the direction from rows to columns in the upper triangle of the matrix, as we only consider symmetrically reversed features as valid hypotheses. Each box shows the feature change from the respective row title to the column title. *Brighter Cleft* refers to less electron density inside the synaptic cleft. *Add DCVs* is the addition of dense core vesicles, *Lower Density* means an overall reduction of content in the pre-synaptic site. *Fill Vesicles* means the darkening of the inside of vesicles. *Darker Cleft* refers to more electron density inside the synaptic cleft. *Remove DCVs* is the removal of dense core vesicles. *Remove small Vs* means the removal of vesicles with a small diameter. *Brighter DVs* refers to the brightening of the inside of medium size vesicles that we call *Dense Vesicles*. *Thinner Cleft* means that the distance between pre and post-synaptic partners is reduced. *!Circ. Vesicles* refers to a change of shape from circular to non-circular vesicles. *Remove PSDs* is the removal of post-synaptic densities. *Larger Vesicles* means the increase of vesicle diameter. *Add DVs* means there is an addition of *Dense Vesicles*. *Less Post-Synapses* describes a reduction in the number of post-synaptic partners. Inlets above the hypothesis matrix show example image pairs and highlighted regions lead to a change of classification decision in the indicated direction if swapped.

4.8 Discussion

4.8.1 Results

We presented a classifier which is able to predict the neurotransmitter identity of a synapse from a local 3D EM volume with high accuracy. We showed that the method generalizes across neurons, brain regions, hemilineages and datasets with different resolutions, staining and microscopes. This is strong evidence that the classifier is indeed learning how neurotransmitter identity affects the phenotype of synaptic sites, as opposed to other spurious correlations that may be present in small homogeneous datasets. We predicted the neurotransmitter identity of 180,675 manually annotated synapses within 1,164 neurons from 89 hemilineages with heretofore unknown neurotransmitter identity. We analyzed the neurotransmitter distribution of 26 hemilineages that have a sufficient amount of annotated neurons and synapses and showed that most of them homogeneously express one fast acting neurotransmitter. However, we also identified a set of five hemilineages that, according to our predictions, express two distinct fast-acting neurotransmitters with high statistical significance. Two of those five, 27 and 42, also show low synaptic entropy $H(S_h)$, indicating that the observed effect is a result of neuron-level neurotransmitter segregation within the hemilineage. Furthermore we predicted all automatically detected synapses in the *FAFB* dataset and made the data publicly available. We also support on demand, custom neurotransmitter prediction queries via a publicly available web-interface. Finally, we use neural network interpretability tools to identify previously unknown hypothetical discriminative features between all six considered neurotransmitters.

4.8.2 Limitations

A potential source of neurotransmitter misclassification is the possibility that a given neuron releases more than one neurotransmitter at its synaptic sites. Due to a lack of known and annotated neurons with cotransmission of the considered neurotransmitters our current model ignores this possibility. However, single cell transcriptomic data of the *Drosophila melanogaster* brain shows that neurotransmitter gene expression is largely exclusive for the fast acting transmitters ACh, GABA and GLUT (Croset et al., 2018), excluding widespread cotransmission of these transmitters. For the considered monoamines (SER, OCT, and DOP), coexpression with another fast acting transmitter is more probable. In particular Croset et al. (2018) suggests that a large fraction of octopaminergic neurons likely corelease GLUT, while SER and DOP show less evidence for coexpression with fast acting transmitters. If a particular neuron in the dataset were to corelease a fast acting transmitter and a monoamine the presented classifier would predict

only one of the two. However, this is not a fundamental limitation of the presented approach and could be remedied if coexpression training data becomes available.

Another current limitation is the fact that we only consider the set of six neurotransmitters {GABA, ACh, GLUT, DOP, OCT, SER} and due to our use of a softmax normalization at the network output layer, the model is forced to select one of the six classes, even if there is no evidence for any of them. As a result, the current model is not able to identify synapses or neurons that release something other than the considered neurotransmitters, notably histamine (Nässel, 2018), tyramine and a vast number of neuropeptides (Croset et al., 2018). Similar to coexpression, we expect an extension to further neurotransmitters to be possible if training data becomes available.

Regarding our analysis of the number of distinct fast-acting neurotransmitters in a hemilineage, a potential source of error is misassignments of neurons to hemilineages. If neurons are erroneously assigned to a particular hemilineage any observed effect of multimodal neurotransmitter distributions on neuron-level within a hemilineage could be an artifact. Furthermore, for hemilineages 16 and 30 the high synaptic entropy $H(S_h)$ suggests that the phenotype of synapses is ambiguous in these hemilineages. Although coexpression of fast-acting neurotransmitters is unlikely (Croset et al., 2018), the neurotransmitter distribution would be consistent with cotransmission of GABA and glutamate in the neurons of these hemilineages.

An apparent issue of our analysis of neurotransmitter discriminatory features is that here we only extract relative differences between the classes, not absolute descriptions of the features of each class. While this might appear as a limitation on first sight, relative differences between all classes lead to descriptive features of each class if we have one reference point. For example, we could measure the average intensity of the synaptic cleft of GABA neurons and thus infer that e.g. ACH must have an average intensity strictly higher as we found a brightening of the cleft going from GABA to ACH. Similarly, we would expect the same average intensity in GABA and GLUT neurons, as we observe no change of synaptic cleft brightness. We hope to address this point in future work by manual annotation of the discovered features for a set of presynaptic sites.

4.8.3 Hypothetical Discriminators

In mammalian brains, it is known that there are two distinct synaptic types, namely asymmetric and symmetric ones. Symmetric synaptic sites are characterized by a comparatively small post synaptic density of the same size as the pre-synaptic density and a thin cleft, while asymmetric synaptic sites

show a broad cleft and a dominant post-synaptic density (Peters and Folger, 2013). Interestingly we find similar features for some neurotransmitter classes in Drosophila. In particular, dopamine shows features of symmetric synaptic sites, as it has a small post synaptic density and a thin cleft compared to the classical transmitters. This is in line with findings in mammalian brains (Uchigashima et al., 2016). However, in contrast to mammalian cells we did not find a relation between inhibitory and excitatory synapses and their appearance as symmetric and asymmetric synaptic sites. In particular the classical transmitters all fit the description of asymmetric synaptic sites, independent of sign.

4.8.4 Generalization

We showed that our network is able to generalize across brain regions and hemilineages. However, although the performance on the brain region split (88% average accuracy) outperforms even the baseline neuron split (87% average accuracy), the hemilineage split suffers a performance decrease of more than 10% (75% average accuracy), suggesting that the influence of the neurotransmitter on the phenotype of a synaptic site is influenced by its hemilineage. This is partially remedied when averaging over multiple synapses: Neuron-level neurotransmitter classification for the hemilineage split is robust with an average accuracy of 92%. Note that the presented data splits already exclude neuron identity, hemilineage identity and brain region as potential confounding variables for the prediction of neurotransmitter identity as performance is far above chance level. Similarly, we show that the classifier generalizes across datasets by training and predicting on the *Hemibrain* dataset. The reduction in accuracy compared to the *FAFB* dataset could be explained by the halved lateral resolution of 8x8x8nm. We observe the strongest performance decrease on the *Hemibrain* test set in the classification of GLUT synaptic sites as the classifier wrongly predicts GABA instead of GLUT. Given the close metabolic relationship between these two transmitters the discriminative features are likely subtle and our analysis suggests that there is a slight difference in vesicle size, a feature that could be invisible in lower resolution datasets. Given the importance of neurotransmitter knowledge for circuit analysis, we recommend imaging future datasets at the highest lateral resolution possible.

Appendix A

Appendix

A.1 Microtubule Tracking

Operation	Size	Feature Maps
Conv	(3,3,3)	12
Conv (1)	(3,3,3)	12
MaxPool	(1,3,3)	12
Conv	(3,3,3)	60
Conv (2)	(3,3,3)	60
MaxPool	(1,3,3)	60
Conv	(3,3,3)	300
Conv (3)	(3,3,3)	300
MaxPool	(1,3,3)	300
Conv	(3,3,3)	1500
Conv	(3,3,3)	1500
TConv	(1,3,3)	300
Concat (3)		600
Conv	(3,3,3)	300
Conv	(3,3,3)	300
TConv	(1,3,3)	60
Concat (2)		120
Conv	(3,3,3)	60
Conv	(3,3,3)	60
TConv	(1,3,3)	12
Concat (1)		24
Conv	(3,3,3)	12
Conv	(3,3,3)	12
Conv	(1,1,1)	$1/10^*$

Table A.1: 3D-UNet architecture used for all models. "TConv" denotes a transposed convolution, "Concat (i)" concatenates features maps from "Conv (i)". The final convolution (denoted by *) produces 1 or 10 feature maps for models NMS_SM and NMS_GRAD, respectively.

Parameter	Value
Input Shape	(32, 322, 322)
Loss	MSE
Optimizer	Adam Kingma and Ba (2014)
Learning Rate	5E-05
β_1	0.95
β_2	0.999
Iterations	300,000

Augmentation	Parameter	Value
Elastic	control point spacing	(4,40,40)
	jitter sigma	(0, 2, 2)
	subsample	8
Rotation	axis	z
	angle	in $[0, \frac{\pi}{2}]$
Section Defects	slip probability	0.05
	shift probability	0.05
	max misalign	10
Mirror	n/a	
Transpose	axes	x, y
Intensity	scale	in $[0.9, 1.1]$
	shift	in $[-0.1, 0.1]$

Table A.2: Training parameters used for all models. Augmentations were performed using our augmentation library (https://github.com/funkey/gunpowder), see online documentation for details.

Model	θ_S	θ_P	θ_D	θ_E	θ_C	θ_d	Block Size b	Context Size $\overline{b}$
NMS_GRAD	180	-80	0	12	14	90	(30, 250, 250)	(50,450,450)
CC_GRAD	200	-70	0	14	14	120	(30, 250, 250)	(50,450,450)
NMS_SM	180	-70	0	14	16	120	(30, 250, 250)	(50,450,450)
NMS_RFC	180	-90	0	12	14	90	(30, 250, 250)	(50,450,450)
Baseline	60	-100	0	12	10	140	(30, 250, 250)	(50,450,450)

Table A.3: ILP validation best parameters for all considered models.

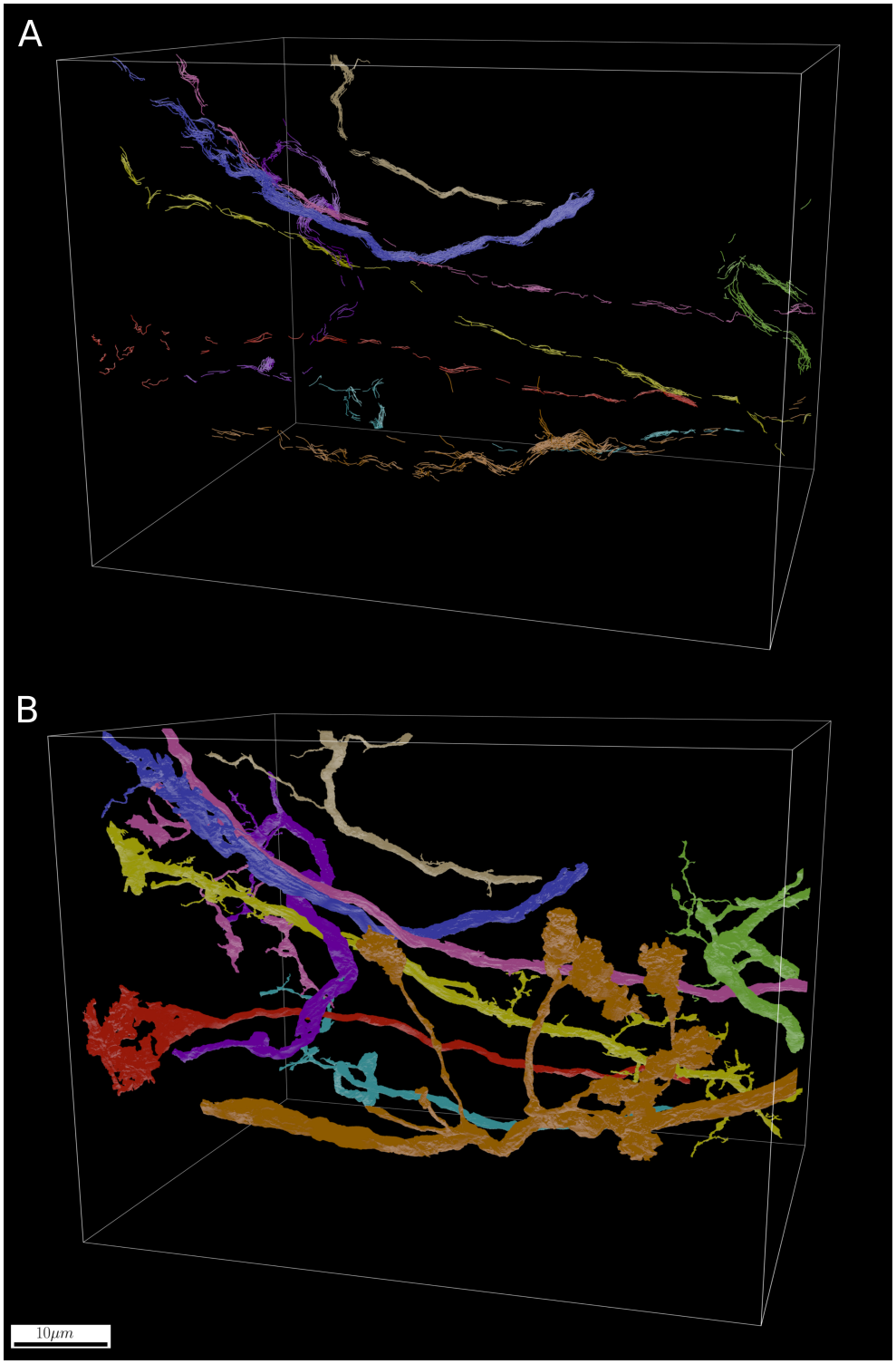


Figure A.1: Rendering of (A) automatically reconstructed microtubules in (B) selected, automatically segmented neurons in the Calyx, a 76 x 52 x 64 μm region of the *Drosophila Melanogaster* brain. Microtubules of the same color belong to the same neuron.

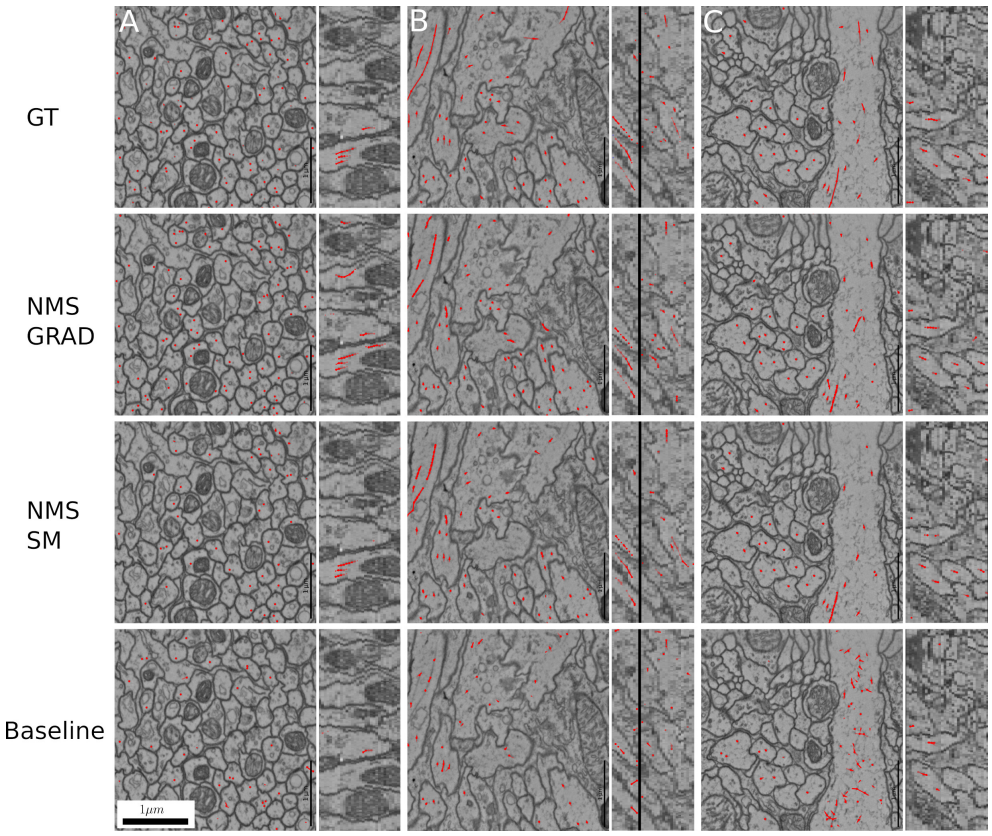


Figure A.2: x-y and x-z view for selected slices of test data sets A, B, C with reconstructed microtubules (red) from groundtruth, our method (NMS_GRAD & NMS_SM) and the baseline. Best viewed on screen.

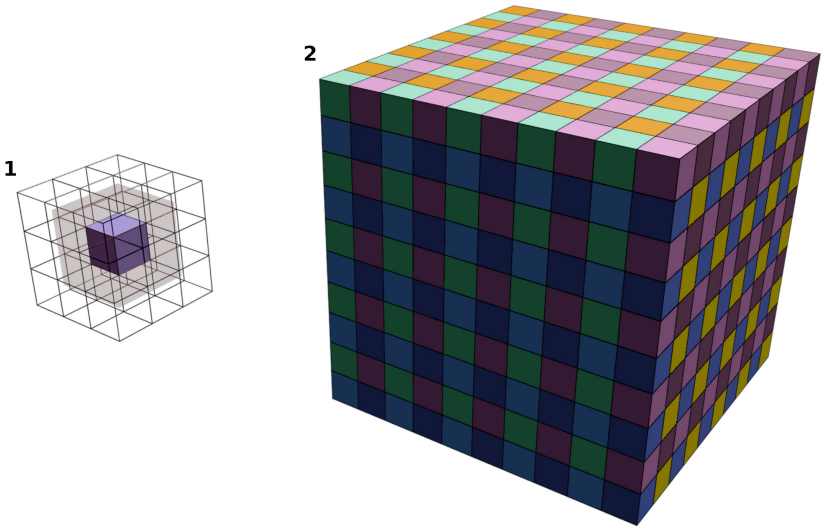


Figure A.3: Illustration of the proposed blockwise processing scheme for distributed ILP solving. **1**: Block region $b \in B$ (purple) and associated context region $\overline{b}$ (ligth purpe). **2**: Conflict free subsets $S_i \subset B$ induced by **1**. Blocks of the same color can be solved in parallel without conflicts.

A.2 Discriminative Attribution from Counterfactuals

A.2.1 Training Details

Network Architectures

Cycle-GAN We extend the cylce-GAN implementation from https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix for our purposes. For all experiments we use a 9-block ResNet generator and a 70×70 PatchGAN (Isola et al., 2017) discriminator. For training we use a least squares loss (LS-GAN (Mao et al., 2017)), a batch size of one, instance normalization and normal initialization. We use the Adam optimizer (Kingma and Ba, 2014) with momentum $\beta_1 = 0.5$ and a learning rate of 0.0002 with a linear decay to zero after the first 100 epochs.

Classifiers The classifiers used for attribution are either VGG (for datasets Synapses, Mnist, and Disc) or ResNet (for datasets Mnist and Disc) architectures, trained using a cross-entropy loss. Individual layers are shown in Table A.4 and Table A.5.

For the training of the VGG network on the Synapses dataset, we use the same strategy (including augmentations) as described in Eckstein et al. (2020a), with the only difference being that we consider 2D images instead of 3D volumes. We did not attempt to train a ResNet on the Synapses dataset.

For the training of the VGG and ResNet architectures on the Mnist and Disc datasets we did not make use of augmentations and trained each network for 100 epochs with a batch size of 32 using the Adam optimizer (learning rate 10^{-4}).

Compute

The most significant part of the compute costs come from training the cycle-GANs. For each experiment, cycle-GAN training for 200 epochs took around 5 days on a single RTX 2080Ti GPU. For Mnist experiments we trained a total of 45 cycle GANs, 15 for Synapses, and 4 for Disc. In total this results in roughly 320 GPU-days for cycle-GAN training. In contrast, attribution and mask generation is comparatively cheap and takes between 1-3 hours on 20 RTX 2080Ti GPUs for each dataset, resulting in 60 GPU hours for each experiment and 15 GPU days in total.

Operation	Tensor Size
input image	$(128, 128)$
`Conv2d`, size $(3, 3)$	$(12, 128, 128)$
`BatchNorm2d`	$(12, 128, 128)$
`ReLU`	$(12, 128, 128)$
`Conv2d`, size $(3, 3)$	$(12, 128, 128)$
`BatchNorm2d`	$(12, 128, 128)$
`ReLU`	$(12, 128, 128)$
`MaxPool2d`, size $(2, 2)$	$(12, 64, 64)$
`Conv2d`, size $(3, 3)$	$(24, 64, 64)$
`BatchNorm2d`	$(24, 64, 64)$
`ReLU`	$(24, 64, 64)$
`Conv2d`, size $(3, 3)$	$(24, 64, 64)$
`BatchNorm2d`	$(24, 64, 64)$
`ReLU`	$(24, 64, 64)$
`MaxPool2d`, size $(2, 2)$	$(24, 32, 32)$
`Conv2d`, size $(3, 3)$	$(48, 32, 32)$
`BatchNorm2d`	$(48, 32, 32)$
`ReLU`	$(48, 32, 32)$
`Conv2d`, size $(3, 3)$	$(48, 32, 32)$
`BatchNorm2d`	$(48, 32, 32)$
`ReLU`	$(48, 32, 32)$
`MaxPool2d`, size $(2, 2)$	$(48, 16, 16)$
`Conv2d`, size $(3, 3)$	$(96, 16, 16)$
`BatchNorm2d`	$(96, 16, 16)$
`ReLU`	$(96, 16, 16)$
`Conv2d`, size $(3, 3)$	$(96, 16, 16)$
`BatchNorm2d`	$(96, 16, 16)$
`ReLU`	$(96, 16, 16)$
`MaxPool2d`, size $(2, 2)$	$(96, 8, 8)$
`Linear`	(4096)
`ReLU`	(4096)
`Dropout`	(4096)
`Linear`	(4096)
`ReLU`	(4096)
`Dropout`	(4096)
`Linear`	(k)

(a) VGG architecture used for the SYNAPSES ($k = 6$), DISC-A ($k = 2$), and DISC-B ($k = 3$) dataset.

Operation	Tensor Size
input image	$(28, 28)$
`Conv2d`, size $(3, 3)$	$(12, 28, 28)$
`BatchNorm2d`	$(12, 28, 28)$
`ReLU`	$(12, 28, 28)$
`Conv2d`, size $(3, 3)$	$(12, 28, 28)$
`BatchNorm2d`	$(12, 28, 28)$
`ReLU`	$(12, 28, 28)$
`MaxPool2d`, size $(2, 2)$	$(12, 14, 14)$
`Conv2d`, size $(3, 3)$	$(24, 14, 14)$
`BatchNorm2d`	$(24, 14, 14)$
`ReLU`	$(24, 14, 14)$
`Conv2d`, size $(3, 3)$	$(24, 14, 14)$
`BatchNorm2d`	$(24, 14, 14)$
`ReLU`	$(24, 14, 14)$
`MaxPool2d`, size $(2, 2)$	$(24, 7, 7)$
`Conv2d`, size $(3, 3)$	$(48, 7, 7)$
`BatchNorm2d`	$(48, 7, 7)$
`ReLU`	$(48, 7, 7)$
`Conv2d`, size $(3, 3)$	$(48, 7, 7)$
`BatchNorm2d`	$(48, 7, 7)$
`ReLU`	$(48, 7, 7)$
`Conv2d`, size $(3, 3)$	$(96, 7, 7)$
`BatchNorm2d`	$(96, 7, 7)$
`ReLU`	$(96, 7, 7)$
`Conv2d`, size $(3, 3)$	$(96, 7, 7)$
`BatchNorm2d`	$(96, 7, 7)$
`ReLU`	$(96, 7, 7)$
`Linear`	(4096)
`ReLU`	(4096)
`Dropout`	(4096)
`Linear`	(4096)
`ReLU`	(4096)
`Dropout`	(4096)
`Linear`	(10)

(b) VGG architecture used for the MNIST dataset.

Table A.4: VGG classifier network architectures.

Operation	Tensor Size
input image	$(128, 128)$
`Conv2d`, size $(3, 3)$	$(12, 128, 128)$
`BatchNorm2d`	$(12, 128, 128)$
`ReLU`	$(12, 128, 128)$
`ResBlock`, stride $(2, 2)$	$(12, 64, 64)$
`ResBlock`	$(12, 64, 64)$
`ResBlock`, stride $(2, 2)$	$(24, 32, 32)$
`ResBlock`	$(24, 32, 32)$
`ResBlock`, stride $(2, 2)$	$(48, 16, 16)$
`ResBlock`	$(48, 16, 16)$
`ResBlock`, stride $(2, 2)$	$(96, 8, 8)$
`ResBlock`	$(96, 8, 8)$
`Linear`	(4096)
`ReLU`	(4096)
`Dropout`	(4096)
`Linear`	(4096)
`ReLU`	(4096)
`Dropout`	(4096)
`Linear`	(k)

(a) ResNet architecture used for the Disc-A ($k = 2$) and Disc-B ($k = 3$) dataset.

Operation	Tensor Size
input image	$(28, 28)$
`Conv2d`, size $(3, 3)$	$(12, 28, 28)$
`BatchNorm2d`	$(12, 28, 28)$
`ReLU`	$(12, 28, 28)$
`ResBlock`, stride $(2, 2)$	$(12, 14, 14)$
`ResBlock`	$(12, 14, 14)$
`ResBlock`, stride $(2, 2)$	$(24, 7, 7)$
`ResBlock`	$(24, 7, 7)$
`ResBlock`, stride $(2, 2)$	$(48, 3, 3)$
`ResBlock`	$(48, 3, 3)$
`ResBlock`, stride $(2, 2)$	$(96, 1, 1)$
`ResBlock`	$(96, 1, 1)$
`Linear`	(4096)
`ReLU`	(4096)
`Dropout`	(4096)
`Linear`	(4096)
`ReLU`	(4096)
`Dropout`	(4096)
`Linear`	(10)

(b) ResNet architecture used for the Mnist dataset.

Table A.5: ResNet classifier network architectures.

A.3 Extended Results for ResNet Architectures

In addition to the results using VGG architectures in the main text, below we show additional results for RESNET architectures on MNIST and DISC-B (see Fig. A.4 and Table A.6). We do not show results for DISC-A, because all considered RESNET architectures failed to achieve more than chance level accuracy on the validation dataset. Since our goal is to understand what the classifier learned about class differences, using a network that did not successfully learn to classify will not produce meaningful results.

The shown results for RESNET architectures follow the same pattern as observed in the main VGG results: All discriminative attribution methods outperform their counterparts in terms of DAC-score. For MNIST, the overall best performing method is the residual, which already performed well for VGG experiments. This is a consequence of the sparsity and simplicity of MNIST and can be observed less drastically for DISC as well. The changes the cycle-GAN introduces are often minimal, and thus the residual is already an accurate attribution. However, in general, the residual is not a good choice for an attribution map as intensity differences between classes do not generally correlate with feature importance. This is particularly noticeable in the experiments on the more challenging SYNAPSES experiments (see main text).

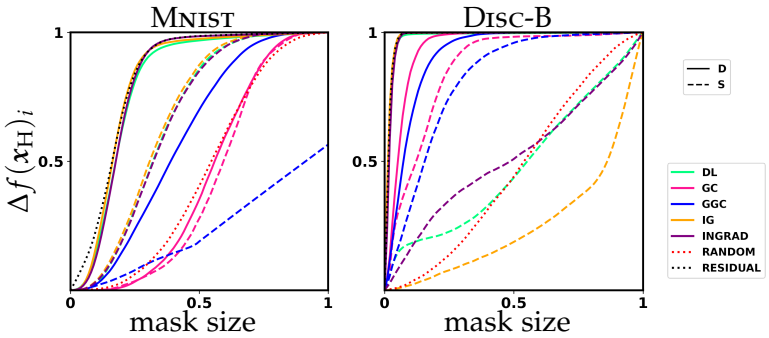


Figure A.4: Quantitative evaluation of discriminative (*D* - solid) and corresponding original (*S* for "single input" - dashed) attribution methods for MNIST and DISC-B using a **ResNet** architecture. Corresponding *D* and *S* versions of the same method are shown in the same color. For each, we plot the average change of classifier prediction $\Delta f(x_H)_i^k = f(x_H)_i - f(x_C)_i$ as a function of mask size $m \in [0,1]$. In addition we show performance of the two considered baselines: masks derived from random attribution maps (random - red, dotted) and mask derived from the residual of the real and counterfactual image (residual - black, dotted). On all considered datasets all versions of *D* attribution outperform their *S* counterparts.

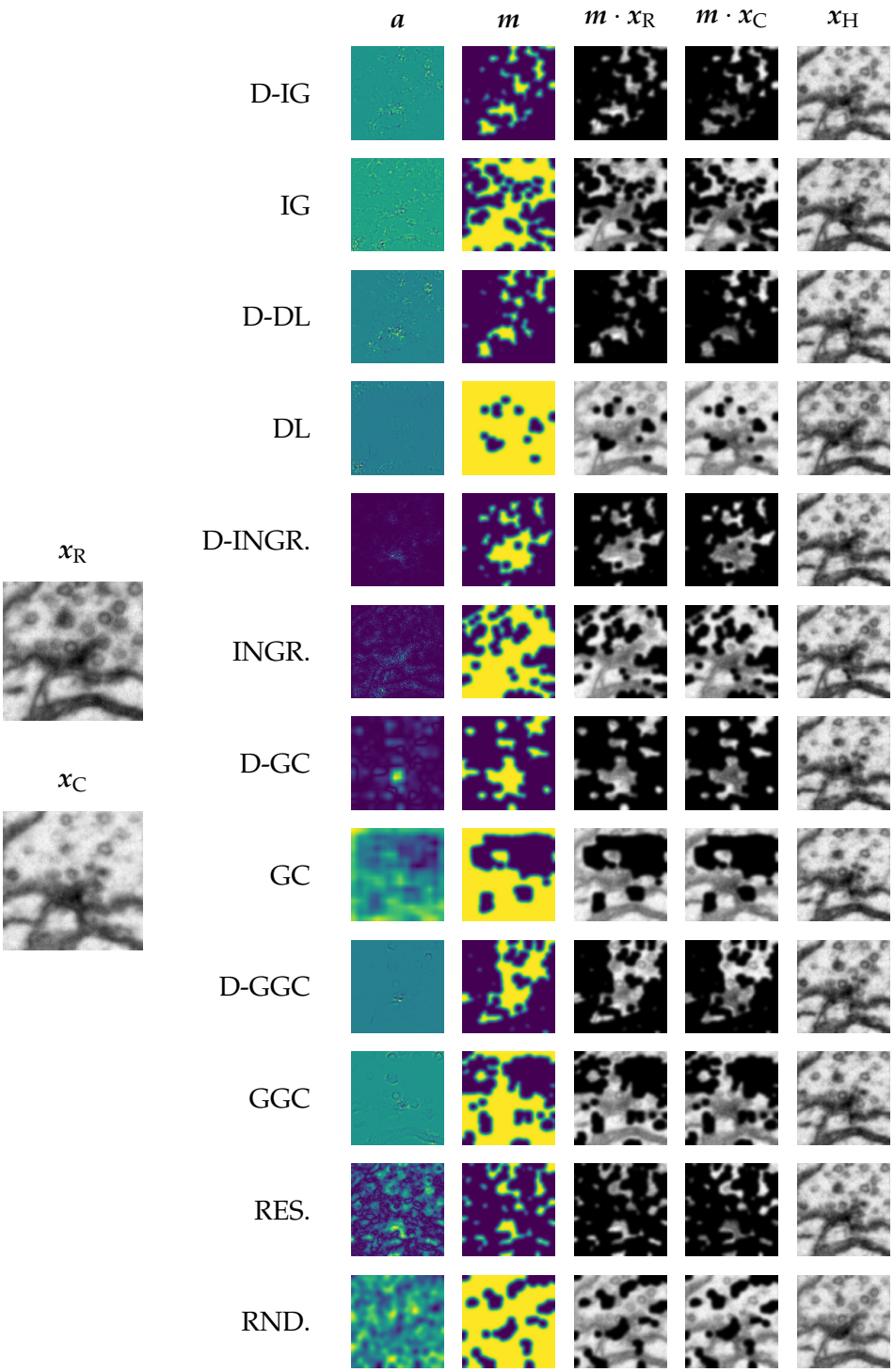


Figure A.5: Qualitative samples from the Synapses dataset for all considered methods. x_R shows a synapse from class Serotonin, x_C shows a synapse from class Acetylcholine.

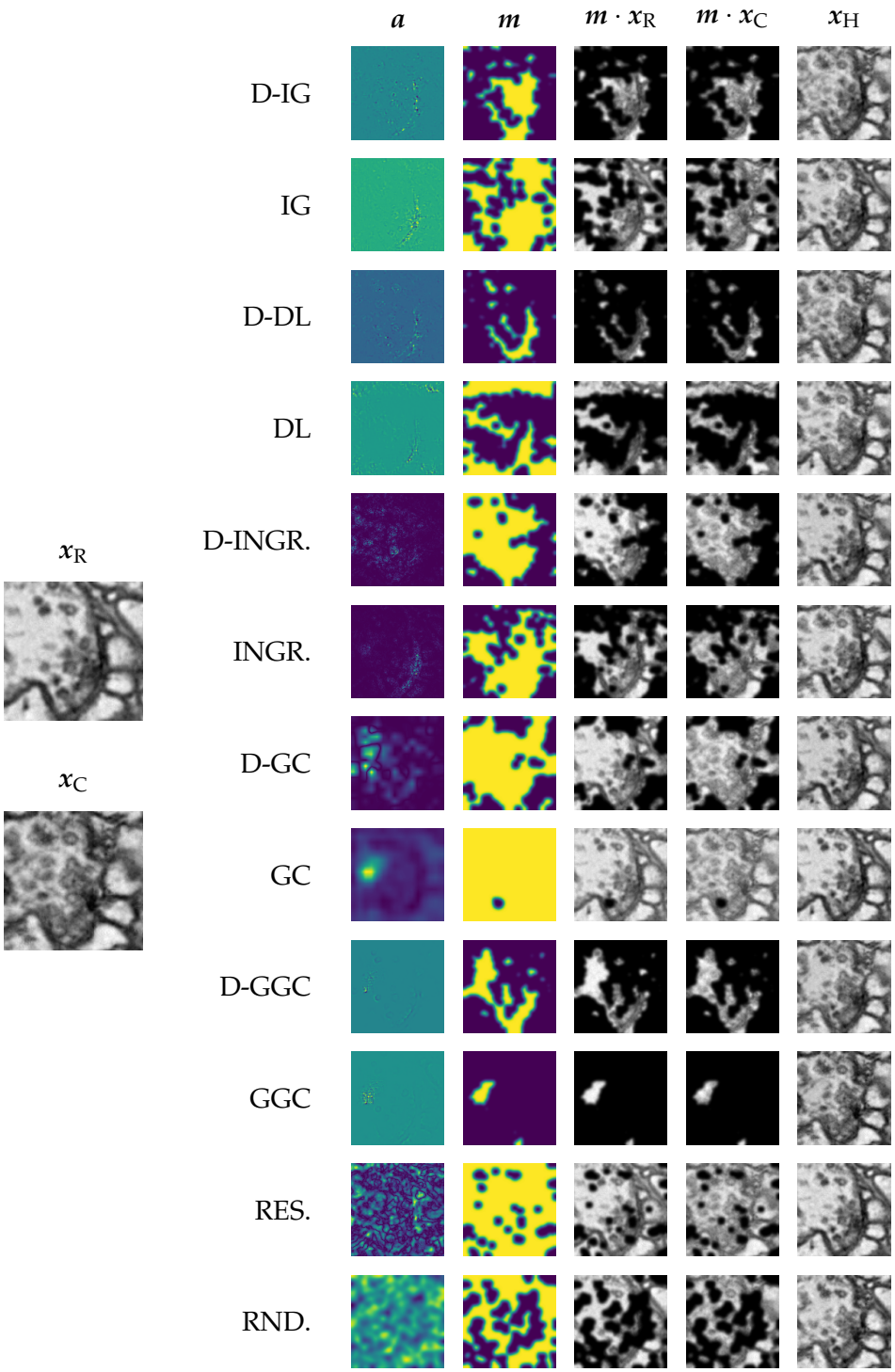


Figure A.6: Qualitative samples from the SYNAPSES dataset for all considered methods. x_R shows a synapse from class Acetylcholine, x_C shows a synapse from class Octopamine.

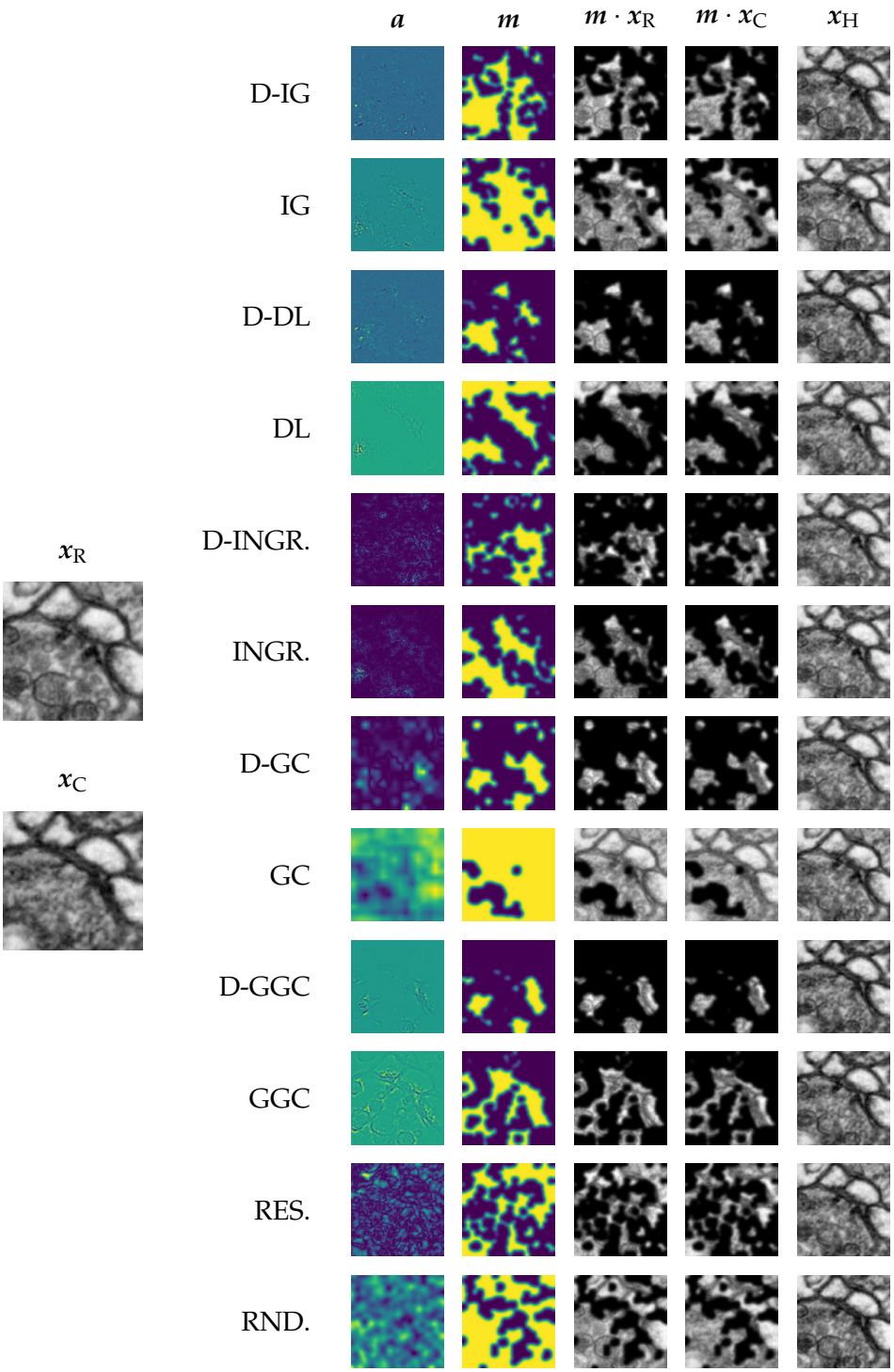


Figure A.7: Qualitative samples from the Synapses dataset for all considered methods. x_R shows a synapse from class Serotonin, x_C shows a synapse from class Glutamate.

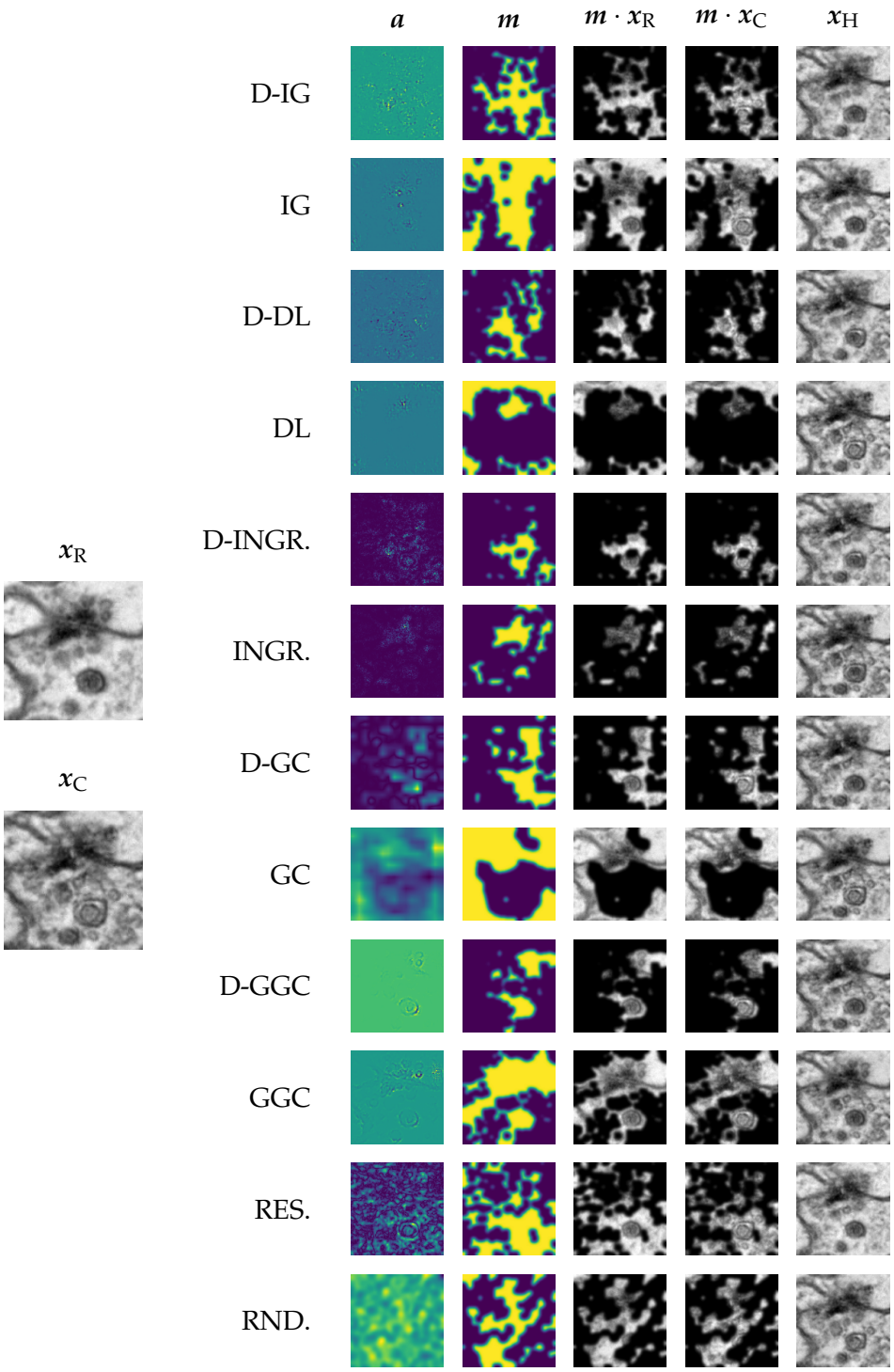


Figure A.8: Qualitative samples from the SYNAPSES dataset for all considered methods. x_R shows a synapse from class Dopamine, x_C shows a synapse from class Serotonin.

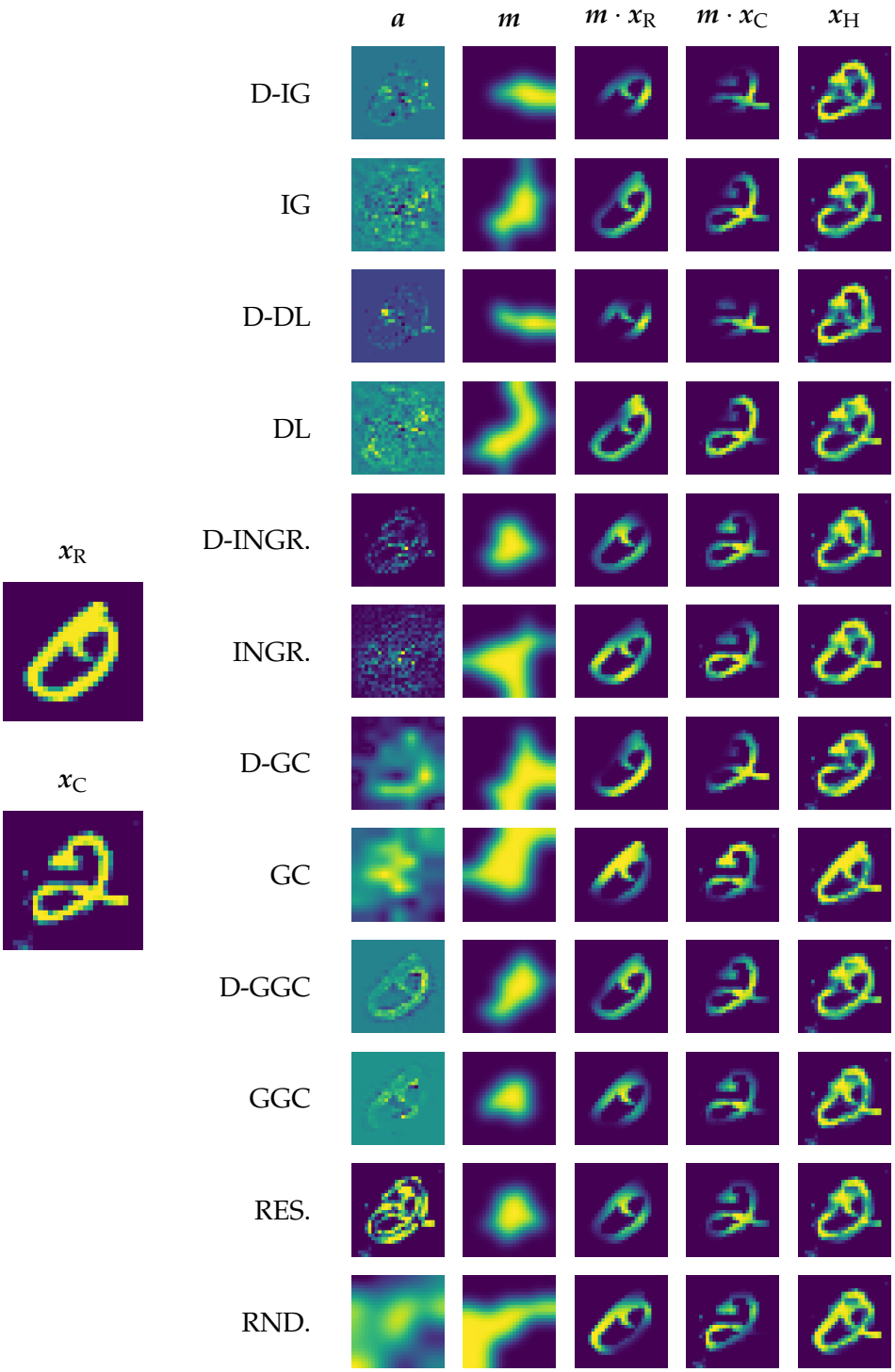


Figure A.9: Qualitative sample from the Mnist dateset for all considered methods.

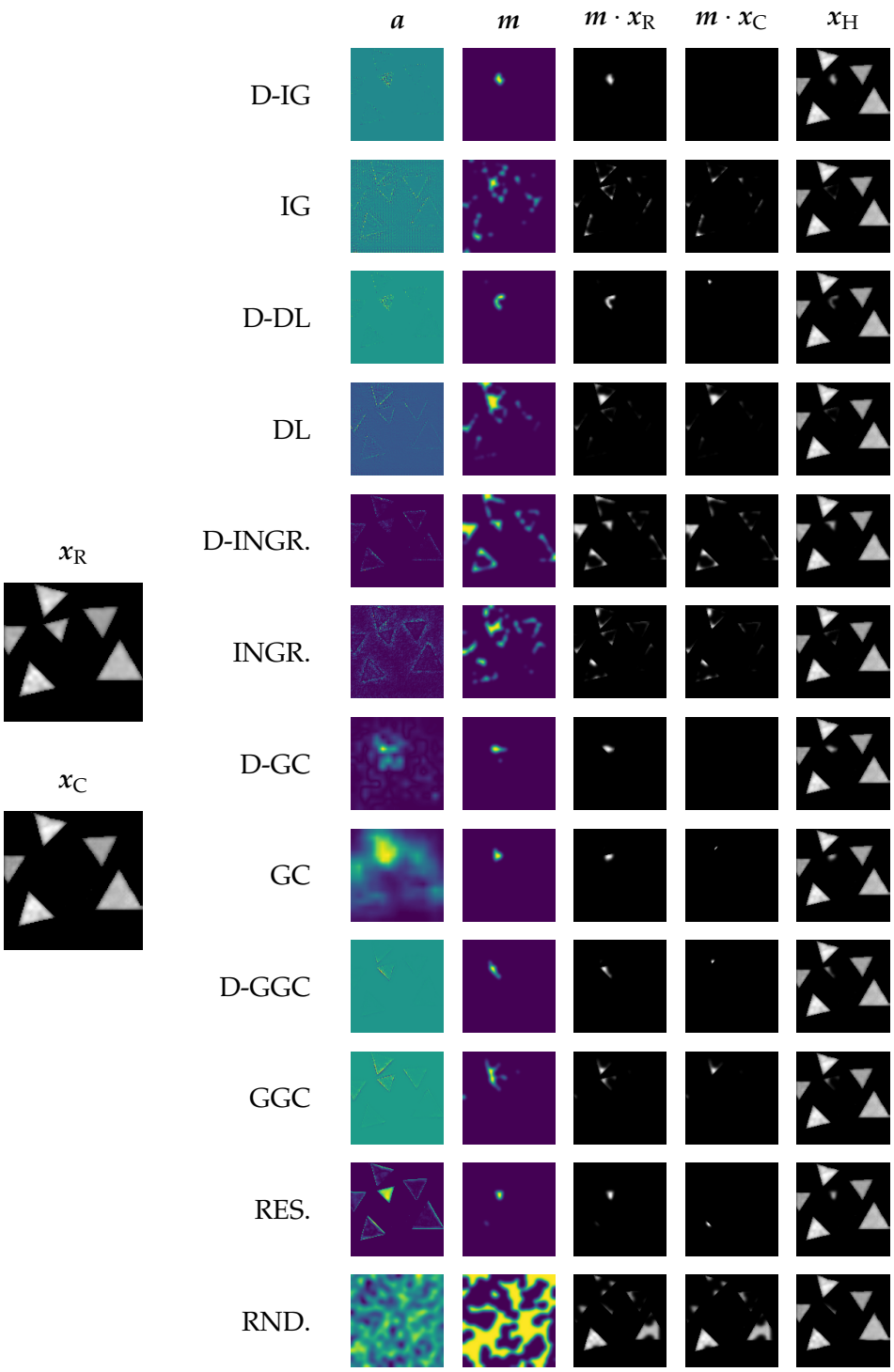


Figure A.10: Qualitative sample from the DISC-A dateset for all considered methods.

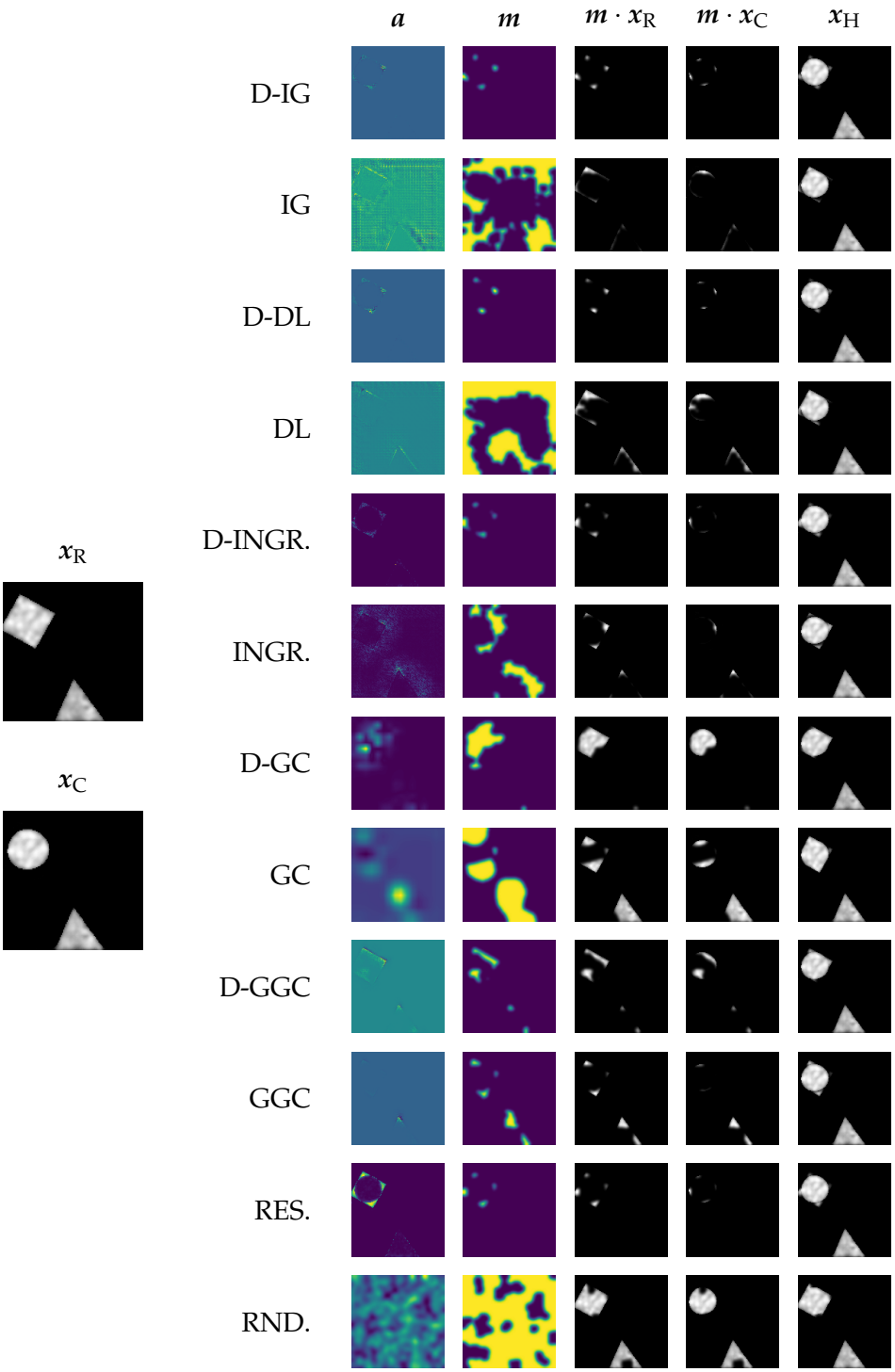


Figure A.11: Qualitative sample from the Disc-B dateset for all considered methods.

Dataset	D-IG	D-DL	D-INGR.	D-GC	D-GGC	RES.	IG	DL	INGR.	GC	GGC	RND.
MNIST	0.82	0.8	0.81	0.44	0.6	**0.83**	0.68	0.66	0.66	0.42	0.23	0.46
DISC-B	**0.98**	0.98	0.98	0.94	0.91	0.98	0.26	0.48	0.52	0.86	0.81	0.47

Table A.6: Summary of DAC scores for RESNET architectures on DISC and MNIST corresponding to Fig. A.4. Best results are highlighted.

A.4 Disc Dataset

The DISC dataset was specifically designed to highlight the advantage of discriminative attribution over vanilla attribution. In particular, the discriminatory feature of DISC-A is the parity of the number of triangles in the image. This feature is non-local and it is unclear what vanilla attribution is supposed to highlight. In DISC-B the classes are defined by the absence of a feature, another situation where vanilla attribution is not designed to give a sensible answer and will often highlight all objects in the image, providing little information to the user.

Disc-A For each image we randomly draw an even (class 0) or odd (class 1) number between one and six, indicating the number of triangles to generate. Each triangle has a random size between 20 and 40% of the image size of 128 pixels and a random position. In addition we draw a random intensity value between 120 and 200, a random rotation angle, and additive noise strength before applying Gaussian smoothing to generate different textures. We reject a sample if the fraction of foreground pixels and the total expected area of all shapes (assuming no overlap) is below 90%, thus avoiding strongly overlapping configurations.

Disc-B Similar to DISC-A, we draw a random position, intensity value, rotation and additive noise strength to generate images showing pairs of a triangle and a square, a disk and a square or a disk and a triangle. We reject a sample if the fraction of foreground pixels and the total expected area of all shapes (assuming no overlap) is below 90%.

A.5 Code and Data Availability

All code, datasets, checkpoints, and instructions needed to reproduce the presented results are available at https://dac-method.github.io.

A.6 Neurotransmitter Classification

Parameter	Value
Input Shape	(16, 160, 160)
Loss	CrossEntropy
Optimizer	Adam
Learning Rate	1E-04
β_1	0.95
β_2	0.999
Iterations	315,000

Augmentation	Parameter	Value
Elastic	control point spacing	(4,40,40)
	jitter sigma	(0, 2, 2)
	subsample	8
Rotation	axis	z
	angle	in $[0, \frac{\pi}{2}]$
Section Defects	slip probability	0.05
	shift probability	0.05
	max misalign	10
Mirror	n/a	
Transpose	axes	x, y
Intensity	scale	in $[0.9, 1.1]$
	shift	in $[-0.1, 0.1]$

Table A.7: Training parameters for best performing FAFB model. Augmentations from http://funkey.science/gunpowder.

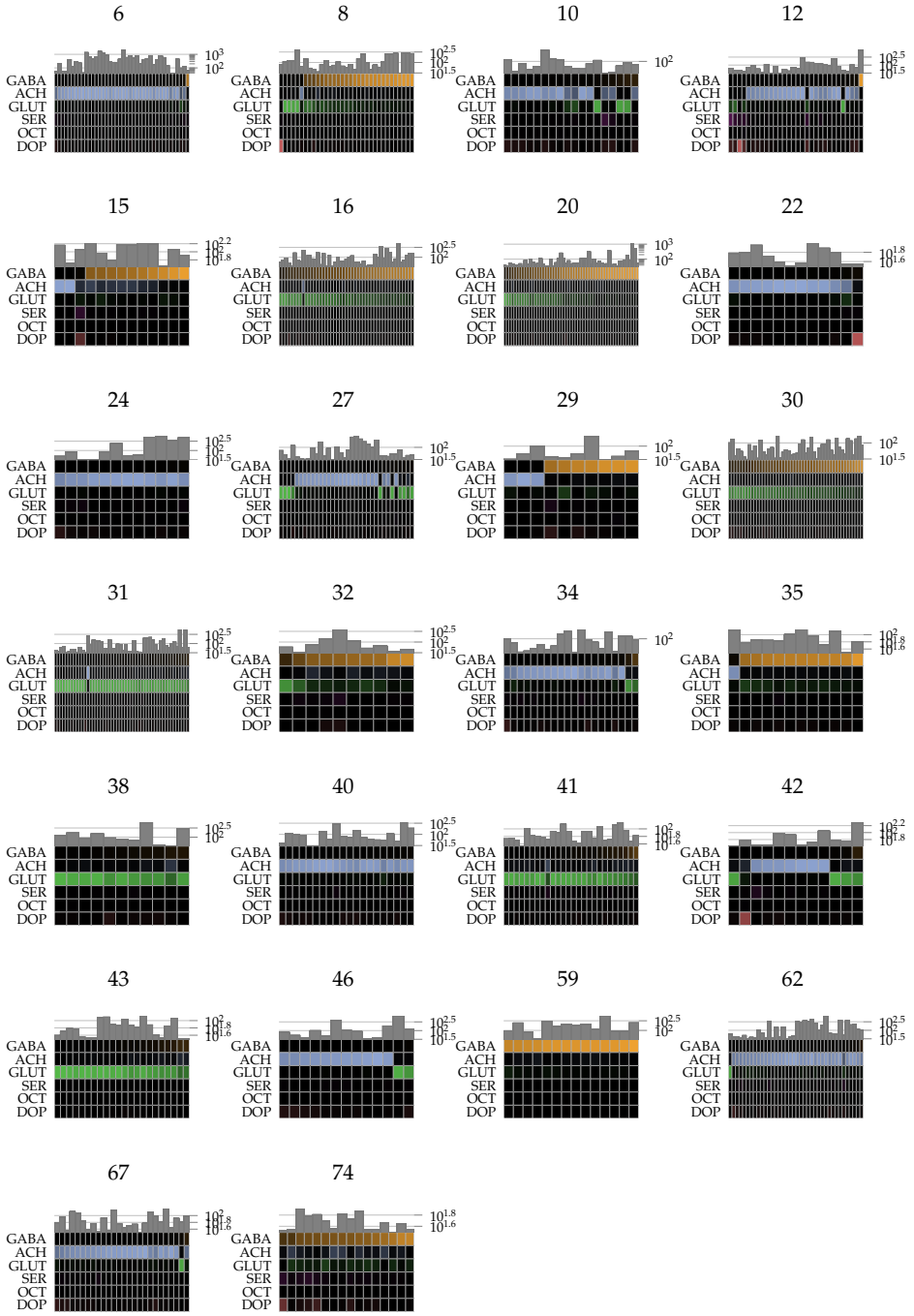


Figure A.12: Neurotransmitter barcode plots of our predictions for all hemilineages that have more than 10 neurons with more than 30 synapses each. Each column represents one neuron. The relative number of synapses predicted as neurotransmitter $\hat{y} \in Y = \{\text{GABA, ACh, GLUT, SER, OCT, DOP}\}$ is represented by the color intensity of the respective row. Corresponding renderings of neurons and predicted neurotransmitters are shown in Fig. A.13. For a mapping of hemilineage ID to hemilineage name see Table A.8

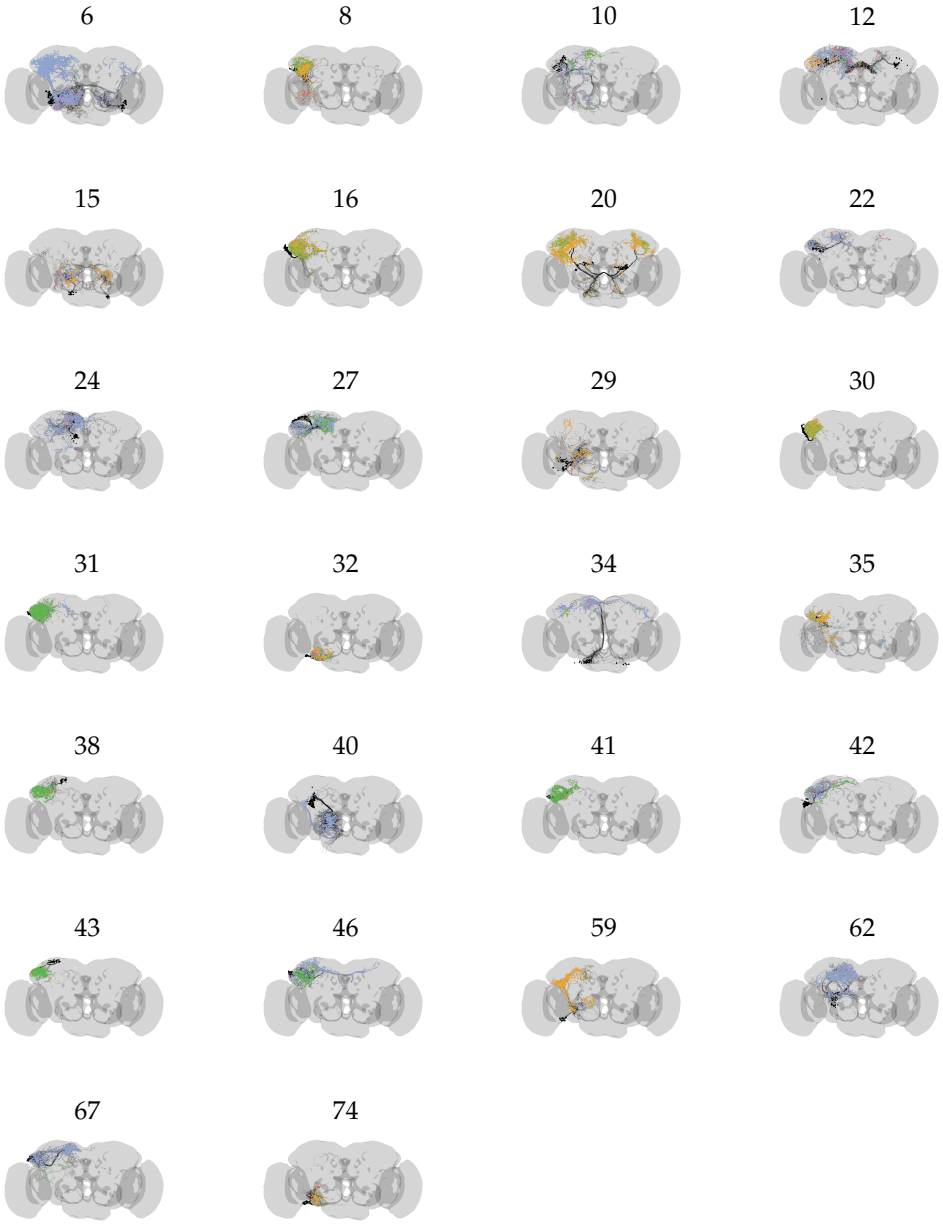


Figure A.13: Renderings of neurotransmitter predictions for all neurons within hemilineages that have more than 10 neurons with more than 30 synapses each. Corresponding neurotransmitter barcode plots are shown in Fig. A.12. For a mapping of hemilineage ID to hemilineage name see Table A.8

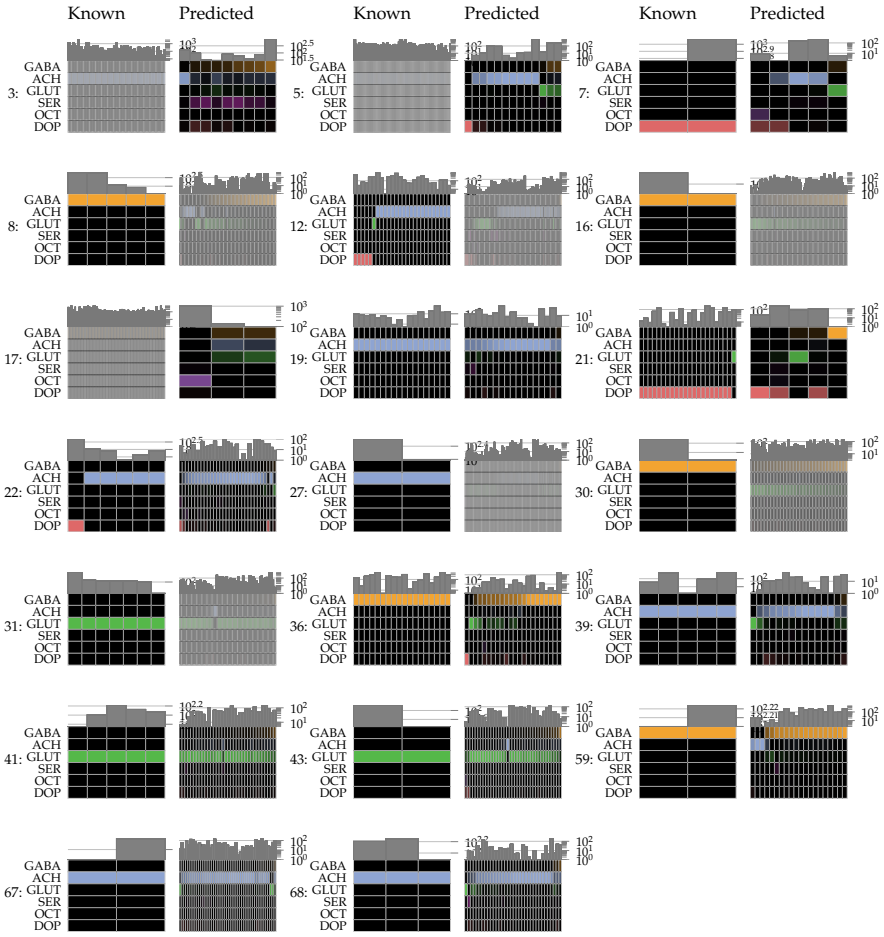


Figure A.14: Neurotransmitter barcode plots of all hemilineages, for which a subset of neurons N^h_{gt} have genetically determined, known neurotransmitters (Known) and our predictions (Predicted) on the remaining neurons N^h_{pred} in the same hemilineage. Each column in the neurotransmitter barcode represents one neuron. The relative number of synapses with neurotransmitter $y, \hat{y} \in Y = \{$GABA, ACh, GLUT, SER, OCT, DOP$\}$ is represented by the color intensity of the respective row. Note that $N^h_{\text{gt}} \cap N^h_{\text{pred}} = \emptyset$. For a mapping of hemilineage ID to hemilineage name see Table A.8

Hemilineage ID	ItoLee Name	Hartenstein Name
1	ALad1	BAmv3
2	AOTUv1	DALcm2
3	ALl1 dorsal	BAlc dorsal
4	SLPav1 lateral	BLAl lateral
5	ALlv1	BAlp4
6	ALl1 ventral	BAlc ventral
7	unnamed	unnamed
8	VPNp and v1 posterior	BLP1 posterior
9	LHl4 posterior	BLD1 posterior
10	VLPd1	DPLam
11	FLAa3	BAmas2
12	DL1 dorsal	CP2 dorsal
13	unnamed	BLP3 ventral
14	DM1	DPMm1
15	LALv1	BAmv1
16	LHl1 lateral	BLD4 lateral
17	ALv1	BAla1
18	LHd1	DPLd
19	SLPal2 ventral	DPLal3 ventral
20	CREa1 ventral	BAmd1 ventral
21	CREa2	DALcm1
22	DL2 dorsal	CP3 dorsal
23	SMPpv1	DPMpl1
24	SIPp1	DPMpl2
25	DL1 ventral	CP2 ventral
26	SLPal1	DPLal1
27	VLPd and p1 posterior	DPLl2 posterior
28	SMPpv2	CP1
29	unnamed	unnamed
30	LHl4 lateral	BLD1 lateral
31	LHp1	BLP4
32	WEDa1 ventral	BAlv
33	SLPad1 anterior	DPLl3 anterior
34	FLAa2	BAmas1
35	VLPp and l1 dorsal	DPLpv dorsal
36	EBa1	DALv2
37	SMPad2	DAMd2/3
38	SLPpm1	DPLm1
39	SLPal2 dorsal	DPLal3 dorsal
40	WEDd1	DALd
41	SLPp and v1 posterior	DPLp2 posterior
42	LHa1 medial	BLAd1 medial
43	LHd2 dorsal	DPLm2 dorsal
44	VPNl and d1 dorsal	BLAvm2 dorsal

Table A.8: Mapping of hemilineage ids to ItoLee and Hartenstein hemilineage names (1/2).

Hemilineage ID	ItoLee Name	Hartenstein Name
45	LHa3	BLVa2
46	LHp2 medial	DPLp1 medial
47	VLPl2 dorsal	BLAv2 dorsal
48	SLPad1 posterior	DPLl3 posterior
49	VLPl2 ventral	BLAv2 ventral
50	SLPav3	BLVa2a
51	LHl2 lateral	DPLal2 lateral
52	VESa1	BAla3
53	SMPpd1	DPLc1
54	SLPal3 dorsal	BLAd3 dorsal
55	VLPl1	BLVa3/4
56	VLPl and p1 posterior	BLVp2 posterior
57	SIPa1 ventral	BLAd2 ventral
58	LHa2 ventral	BLVa1 ventral
59	WEDa2	BAlp3
60	SIPa1 dorsal	BLAd2 dorsal
61	SLPav2 dorsal	BLD2 dorsal
62	VLPl and p2 posterior	BLVp1 posterior
63	SLPpl1	DPLl1
64	LHp2 lateral	DPLp1 lateral
65	SLPpl3 lateral	unnamed
66	VLPl and d1 lateral	BLAv1 lateral
67	LHl2 medial	DPLal2 medial
68	VLPl and p1 anterior	BLVp2 anterior
69	VLPl and d1 dorsal	BLAv1 dorsal
70	AOTUv2	DALl1
71	AOTUv3 dorsal	DALcl1 dorsal
72	VLPp and l1 ventral	DPLpv ventral
73	VPNd1	BLD6
74	WEDa1 dorsal	BAlv
75	VLPd and p1 anterior	DPLl2 anterior
76	unnamed	unnamed
77	VLPp1	BLP2
78	PSa1	BAlp1
79	Primary	Primary
80	VPNp1 posterior	BLD5 posterior
81	unnamed	unnamed
82	DL2 ventral	CP3 ventral
83	LHa2 dorsal	BLVa1 dorsal
84	SLPpl3 posterior	unnamed
85	PSp3	DPMl1/2
86	SLPa and l1 lateral	BLAl lateral
87	SLPa and l1 anterior	BLAvm1 anterior
88	AOTUv3 ventral	DALcl1 ventral
89	CLp1	DPLc2/4

Table A.9: Mapping of hemilineage ids to ItoLee and Hartenstein hemilineage names (2/2).

Operation	Size	Feature Maps
Conv (1)	(3,3,3)	8
BatchNorm		8
ReLU		8
Conv (2)	(3,3,3)	8
BatchNorm		8
MaxPool	(1,2,2)	8
Conv (3)	(3,3,3)	16
BatchNorm		16
ReLU		16
Conv (4)	(3,3,3)	16
BatchNorm		16
MaxPool	(1,2,2)	16
Conv (5)	(3,3,3)	32
BatchNorm		32
ReLU		32
Conv (6)	(3,3,3)	32
BatchNorm		32
MaxPool	(1,2,2)	32
Conv (7)	(3,3,3)	64
BatchNorm		64
ReLU		64
Conv (8)	(3,3,3)	64
BatchNorm		64
MaxPool	(1,2,2)	64
Linear	4096	1
ReLU		1
Dropout		1
Linear	4096	1
ReLU		1
Dropout		1
Linear	6	1

Table A.10: Best performing 3D-VGG-type architecture used for FAFB predictions.

Bibliography

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). Sanity checks for saliency maps. *arXiv preprint arXiv:1810.03292*.

Ahmad, M. A., Eckert, C., and Teredesai, A. (2018). Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pages 559–560.

Alvarez-Melis, D. and Jaakkola, T. S. (2018). On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*.

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. (2018). Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*.

Aso, Y., Hattori, D., Yu, Y., Johnston, R. M., Nirmala, A., Ngo, T.-T., Dionne, H., Abbott, L. F., Axel, R., Tanimoto, H., and Rubin, G. M. (2014). The neuronal architecture of the mushroom body provides a logic for associative learning. *Elife*, pages 1–47.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and Müller, K.-R. (2010). How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831.

Baker, C. A., McKellar, C., Nern, A., Dorkenwald, S., Pacheco, D. A., Pang, R., Eckstein, N., Funke, J., Dickson, B. J., and Murthy, M. (2021). Neural network organization for courtship song feature detection in drosophila. *bioRxiv*.

Barnes, C. L., Bonnéry, D., and Cardona, A. (2020). Synaptic counts approximate synaptic contact area in drosophila. *bioRxiv*.

Barredo-Arrieta, A. and Del Ser, J. (2020). Plausible counterfactuals: Auditing deep learning classifiers with realistic adversarial examples. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Bates, A. S., Janssens, J., Jefferis, G. S., and Aerts, S. (2019a). Neuronal cell types in the fly: single-cell anatomy meets single-cell genomics. *Curr. Opin. Neurobiol.*, 56:125–134.

Bates, A. S., Manton, J. D., Jagannathan, S. R., Costa, M., Schlegel, P., Rohlfing, T., and Gregory S X (2019b). The natverse: a versatile computational toolbox to combine and analyse neuroanatomical data.

Bates, A. S., Schlegel, P., Roberts, R. J. V., Drummond, N., Tamimi, I. F. M., Turnbull, R., Zhao, X., Marin, E. C., Popovici, P. D., Dhawan, S., Jamasb, A., Javier, A., Li, F., Rubin, G. M., Waddell, S., Bock, D. D., Costa, M., and G S X (2020). Complete connectomic reconstruction of olfactory projection neurons in the fly brain.

Beier, T., Pape, C., Rahaman, N., Prange, T., Berg, S., Bock, D. D., Cardona, A., Knott, G. W., Plaza, S. M., Scheffer, L. K., et al. (2017). Multicut brings automated neurite segmentation closer to human performance. *Nature methods*, 14(2):101–102.

Berning, M., Boergens, K. M., and Helmstaedter, M. (2015). Segem: efficient image analysis for high-resolution connectomics. *Neuron*, 87(6):1193–1206.

Boergens, K. M., Berning, M., Bocklisch, T., Bräunlein, D., Drawitsch, F., Frohnhofen, J., Herold, T., Otto, P., Rzepka, N., Werkmeister, T., et al. (2017). webknossos: efficient online 3d data annotation for connectomics. *nature methods*, 14(7):691–694.

Bräcker, L. B., Siju, K. P., Varela, N., Aso, Y., Zhang, M., Hein, I., Vasconcelos, M. L., and Grunwald Kadow, I. C. (2013). Essential role of the mushroom body in context-dependent CO2 avoidance in *Drosophila*. *Curr. Biol.*, 23(13):1228–1234.

Breitenfeld, T., Jurasic, M., and Breitenfeld, D. (2014). Hippocrates: the forefather of neurology. *Neurological Sciences*, 35(9):1349–1352.

Buhmann, J., Krause, R., Ceballos Lentini, R., Eckstein, N., Cook, M., Turaga, S., and Funke, J. (2018). Synaptic partner prediction from point annotations in insect brains. *Medical Image Computing and Computer Assisted Intervention MICCAI 2018*, 11071.

Buhmann, J., Sheridan, A., Gerhard, S., Krause, R., Nguyen, T., Heinrich, L., Schlegel, P., Lee, W.-C. A., Wilson, R., Saalfeld, S., Jefferis, G., Bock, D., Turaga, S., Cook, M., and Funke, J. (2019). Automatic detection of synaptic partners in a whole-brain drosophila em dataset. *bioRxiv*.

Buhmann, J. M., Gerhard, S., Cook, M., and Funke, J. (2016). Tracking of microtubules in anisotropic volumes of neural tissue. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*.

Busch, S., Selcho, M., Ito, K., and Tanimoto, H. (2009). A map of octopaminergic neurons in the *Drosophila* brain. *J. Comp. Neurol.*, 513(6):643–667.

Cardona, A., Saalfeld, S., Schindelin, J., Arganda-Carreras, I., Preibisch, S., Longair, M., Tomancak, P., Hartenstein, V., and Douglas, R. J. (2012). TrakEM2 software for neural circuit reconstruction. *PLoS One*, 7(6):e38011.

Che, T., Li, Y., Jacob, A. P., Bengio, Y., and Li, W. (2016). Mode regularized generative adversarial networks. *arXiv preprint arXiv:1612.02136*.

Cheng, H.-C. and Varshney, A. (2017). Volume segmentation using convolutional neural networks with limited training data. In *2017 IEEE international conference on image processing (ICIP)*, pages 590–594. IEEE.

Ciresan, D., Giusti, A., Gambardella, L., and Schmidhuber, J. (2012). Deep neural networks segment neuronal membranes in electron microscopy images. *Advances in neural information processing systems*, 25:2843–2851.

Cook, S. J., Jarrell, T. A., Brittin, C. A., Wang, Y., Bloniarz, A. E., Yakovlev, M. A., Nguyen, K. C., Tang, L. T.-H., Bayer, E. A., Duerr, J. S., et al. (2019). Whole-animal connectomes of both caenorhabditis elegans sexes. *Nature*, 571(7763):63–71.

Costa, M., Manton, J. D., Ostrovsky, A. D., Prohaska, S., and Jefferis, G. S. X. E. (2016). NBLAST: Rapid, sensitive comparison of neuronal structure and construction of neuron family databases. *Neuron*, 91(2):293–311.

Cranmer, M., Sanchez-Gonzalez, A., Battaglia, P., Xu, R., Cranmer, K., Spergel, D., and Ho, S. (2020). Discovering symbolic models from deep learning with inductive biases. *arXiv preprint arXiv:2006.11287*.

Croset, V., Treiber, C., and Waddell, S. (2018). Cellular diversity in the *Drosophila* midbrain revealed by single-cell transcriptomics. *Elife*, 7:e34550.

Cruz-Roa, A. A., Ovalle, J. E. A., Madabhushi, A., and Osorio, F. A. G. (2013). A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 403–410. Springer.

Dabkowski, P. and Gal, Y. (2017). Real time image saliency for black box classifiers. *arXiv preprint arXiv:1705.07857*.

Dale, H. (1934). Pharmacology and nerve endings. *Proc. R. Soc. Med*, 28:319–332.

Davie, K., Janssens, J., Koldere, D., De Waegeneer, M., Pech, U., Kreft, Ł., Aibar, S., Makhzami, S., Christiaens, V., González-Blas, C. B., et al. (2018). A single-cell transcriptome atlas of the aging drosophila brain. *Cell*, 174(4):982–998.

Davis, F. P., Nern, A., Picard, S., Reiser, M. B., Rubin, G. M., Eddy, S. R., and Henry, G. L. (2018). A genetic, genomic, and computational resource for exploring neural circuit function.

Davis, F. P., Nern, A., Picard, S., Reiser, M. B., Rubin, G. M., Eddy, S. R., and Henry, G. L. (2020). A genetic, genomic, and computational resource for exploring neural circuit function. *Elife*, 9.

Dolan, M.-J., Belliart-Guérin, G., Bates, A. S., Frechter, S., Lampin-Saint-Amaux, A., Aso, Y., Roberts, R. J. V., Schlegel, P., Wong, A., Hammad, A., Bock, D., Rubin, G. M., Preat, T., Plaçais, P.-Y., and Jefferis, G. S. X. E. (2018). Communication from learned to innate olfactory processing centers is required for memory retrieval in *Drosophila*. *Neuron*.

Dolan, M.-J., Frechter, S., Bates, A. S., Dan, C., Huoviala, P., Roberts, R. J., Schlegel, P., Dhawan, S., Tabano, R., Dionne, H., et al. (2019). Neurogenetic dissection of the drosophila lateral horn reveals major outputs, diverse behavioural functions, and interactions with the mushroom body. *eLife*, 8:e43079.

Dorkenwald, S., McKellar, C., Macrina, T., Kemnitz, N., Lee, K., Lu, R., Wu, J., Popovych, S., Mitchell, E., Nehoran, B., Jia, Z., Bae, J. A., Mu, S., Ih, D., Castro, M., Ogedengbe, O., Halageri, A., Ashwood, Z., Zung, J., Brittain, D., Collman, F., Schneider-Mizell, C., Jordan, C., Silversmith, W., Baker, C., Deutsch, D., Encarnacion-Rivera, L., Kumar, S., Burke, A., Gager, J., Hebditch, J., Koolman, S., Moore, M., Morejohn, S., Silverman, B., Willie, K., Willie, R., Yu, S.-c., Murthy, M., and Seung, H. S. (2020). Flywire: Online community for whole-brain connectomics. *bioRxiv*.

Dorkenwald, S., Schubert, P. J., Killinger, M. F., Urban, G., Mikula, S., Svara, F., and Kornfeld, J. (2017). Automated synaptic connectivity inference for volume electron microscopy. *Nature methods*, 14(4):435–442.

Eccles, J. C. (1976). From electrical to chemical transmission in the central nervous system: the closing address of the sir henry dale centennial symposium cambridge, 19 september 1975. *Notes and records of the Royal Society of London*, 30(2):219–230.

Eckstein, N., Bates, A. S., Du, M., Hartenstein, V., Jefferis, G. S., and Funke, J. (2020a). Neurotransmitter classification from electron microscopy images at synaptic sites in drosophila. *BioRxiv*.

Eckstein, N., Buhmann, J., Cook, M., and Funke, J. (2020b). Microtubule tracking in electron microscopy volumes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 99–108. Springer.

Eichler, K., Li, F., Litwin-Kumar, A., Park, Y., Andrade, I., Schneider-Mizell, C. M., Saumweber, T., Huser, A., Eschbach, C., Gerber, B., et al. (2017). The complete connectome of a learning and memory centre in an insect brain. *Nature*, 548(7666):175–182.

Eschbach, C., Fushiki, A., Winding, M., Schneider-Mizell, C. M., Shao, M., Arruda, R., Eichler, K., Valdes-Aleman, J., Ohyama, T., Thum, A. S., et al. (2020). Recurrent architecture for adaptive regulation of learning in the insect brain. *Nature neuroscience*, 23(4):544–555.

Felsenberg, J., Jacob, P. F., Walker, T., Barnstedt, O., Edmondson-Stait, A. J., Pleijzier, M. W., Otto, N., Schlegel, P., Sharifi, N., Perisse, E., Smith, C. S., Lauritzen, J. S., Costa, M., Jefferis, G. S. X. E., Bock, D. D., and Waddell, S. (2018). Integration of parallel opposing memories underlies memory extinction. *Cell*, 175(3):709–722.e15.

Fong, R. C. and Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3429–3437.

Frechter, S., Bates, A. S., Tootoonian, S., Dolan, M.-J., Manton, J. D., Jamasb, A. R., Kohl, J., Bock, D., and Jefferis, G. S. (2019). Functional and anatomical specificity in a higher olfactory centre. *Elife*, 8.

Funke, J., Tschopp, F. D., Grisaitis, W., Sheridan, A., Singh, C., Saalfeld, S., and Turaga, S. C. (2018). Large scale image segmentation with structured loss based deep learning for connectome reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.

Ghorbani, A., Abid, A., and Zou, J. (2019). Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688.

Gittes, F., Mickey, B., Nettleton, J., and Howard, J. (1993). Flexural rigidity of microtubules and actin filaments measured from thermal fluctuations in shape. *The Journal of cell biology*, 120(4):923–934.

Goyal, R. K. and Chaudhury, A. (2013). Structure activity relationship of synaptic and junctional neurotransmission. *Autonomic Neuroscience*, 176(1-2):11–31.

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., and Lee, S. (2019). Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384. PMLR.

Guillery, R. W. (2005). Observations of synaptic structures: origins of the neuron doctrine and its current status. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1458):1281–1307.

Hampel, S., Chung, P., McKellar, C. E., Hall, D., Looger, L. L., and Simpson, J. H. (2011). Drosophila brainbow: a recombinase-based fluorescence labeling technique to subdivide neural expression patterns. *Nature methods*, 8(3):253–259.

Hanslovsky, P., Bogovic, J. A., and Saalfeld, S. (2015). Post-acquisition image based compensation for thickness variation in microscopy section series. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 507–511. IEEE.

Hanslovsky, P., Bogovic, J. A., and Saalfeld, S. (2017). Image-based correction of continuous and discontinuous non-planar axial distortion in serial section microscopy. *Bioinformatics*, 33(9):1379–1386.

Hayworth, K., Kasthuri, N., Schalek, R., and Lichtman, J. (2006). Automating the collection of ultrathin serial sections for large volume tem reconstructions. *Microscopy and Microanalysis*, 12(S02):86.

Hayworth, K. J., Xu, C. S., Lu, Z., Knott, G. W., Fetter, R. D., Tapia, J. C., Lichtman, J. W., and Hess, H. F. (2015). Ultrastructurally smooth thick partitioning and volume stitching for large-scale connectomics. *Nature methods*, 12(4):319–322.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Heinrich, L., Bennett, D., Ackerman, D., Park, W., Bogovic, J., Eckstein, N., Petruncio, A., Clements, J., Xu, C. S., Funke, J., Korff, W., Hess, H. F., Lippincott-Schwartz, J., Saalfeld, S., Weigel, A. V., and Team, C. P. (2020). Automatic whole cell organelle segmentation in volumetric electron microscopy. *bioRxiv*.

Heinrich, L., Funke, J., Pape, C., Nunez-Iglesias, J., and Saalfeld, S. (2018). Synaptic cleft segmentation in non-isotropic volume electron microscopy of the complete drosophila brain. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 317–325. Springer.

Heisenberg, M. (2003). Mushroom body memoir: from maps to models. *Nature Reviews Neuroscience*, 4(4):266–275.

Helmstaedter, M., Briggman, K. L., Turaga, S. C., Jain, V., Seung, H. S., and Denk, W. (2013). Connectomic reconstruction of the inner plexiform layer in the mouse retina. *Nature*, 500(7461):168.

Henry, G. L., Davis, F. P., Picard, S., and Eddy, S. R. (2012). Cell type–specific genomics of drosophila neurons. *Nucleic acids research*, 40(19):9691–9704.

Hooker, S., Erhan, D., Kindermans, P.-J., and Kim, B. (2018). A benchmark for interpretability methods in deep neural networks. *arXiv preprint arXiv:1806.10758*.

Huang, G. B., Scheffer, L. K., and Plaza, S. M. (2018). Fully-automatic synapse prediction and validation on a large data set. *Frontiers in neural circuits*, 12:87.

Huoviala, P., Dolan, M.-J., Love, F. M., Frechter, S., Roberts, R. J. V., Mitrevica, Z., Schlegel, P., Bates, A. S., Aso, Y., Rodrigues, T., Cornwall, H., Stensmyr, M., Bock, D., Rubin, G. M., Costa, M., and Gregory S X (2018). Neural circuit basis of aversive odour processing in *Drosophila* from sensory input to descending output.

Hyatt, A. D. and Wise, T. G. (2001). Immunolabeling. In *Immunocytochemistry and In Situ Hybridization in the Biomedical Sciences*, pages 73–107. Springer.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Ito, M., Masuda, N., Shinomiya, K., Endo, K., and Ito, K. (2013). Systematic analysis of neural projections reveals clonal composition of the *Drosophila* brain. *Curr. Biol.*, 23(8):644–655.

Januszewski, M., Kornfeld, J., Li, P. H., Pope, A., Blakely, T., Lindsey, L., Maitin-Shepard, J., Tyka, M., Denk, W., and Jain, V. (2018). High-precision automated reconstruction of neurons with flood-filling networks. *Nature methods*, page 1.

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.

Jennings, B. H. (2011). Drosophila–a versatile model in biology & medicine. *Materials today*, 14(5):190–195.

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. (2019). The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 267–280. Springer.

Kindermans, P.-J., Schütt, K. T., Alber, M., Müller, K.-R., Erhan, D., Kim, B., and Dähne, S. (2017). Learning how to explain neural networks: Patternnet and patternattribution. *arXiv preprint arXiv:1705.05598*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Knott, G., Marchman, H., Wall, D., and Lich, B. (2008). Serial section scanning electron microscopy of adult brain tissue using focused ion beam milling. *Journal of Neuroscience*, 28(12):2959–2964.

Konstantinides, N., Rossi, A. M., and Desplan, C. (2015). Common temporal identity factors regulate neuronal diversity in fly ventral nerve cord and mouse retina. *Neuron*, 85(3):447–449.

Kornfeld, J. and Denk, W. (2018). Progress and remaining challenges in high-throughput volume electron microscopy. *Current opinion in neurobiology*, 50:261–267.

Kreshuk, A., Funke, J., Cardona, A., and Hamprecht, F. A. (2015). Who is talking to whom: Synaptic partner detection in anisotropic volumes of insect brain. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 661–668, Cham. Springer International Publishing.

Kumar, A., Bello, B., and Reichert, H. (2009). Lineage-specific cell death in postembryonic brain development of drosophila. *Development*, 136(20):3433–3442.

Lacin, H., Chen, H.-M., Long, X., Singer, R. H., Lee, T., and Truman, J. W. (2019). Neurotransmitter identity is acquired in a lineage-restricted manner in the *Drosophila* CNS. *Elife*, 8.

Lai, S.-L., Awasaki, T., Ito, K., and Lee, T. (2008). Clonal analysis of *Drosophila* antennal lobe neurons: diverse neuronal architectures in the lateral neuroblast lineage. *Development*, 135(17):2883–2893.

Larson, S. D., Gleeson, P., and Brown, A. E. (2018). Connectome to behaviour: modelling caenorhabditis elegans at cellular resolution.

LeCun, Y. and Cortes, C. (2010). MNIST handwritten digit database.

Lee, K., Lu, R., Luther, K., and Seung, H. S. (2019). Learning Dense Voxel Embeddings for 3D Neuron Reconstruction. *arXiv e-prints*, page arXiv:1909.09872.

Lee, K., Zung, J., Li, P., Jain, V., and Seung, H. S. (2017). Superhuman accuracy on the snemi3d connectomics challenge. *arXiv preprint arXiv:1706.00120*.

Li, F., Lindsey, J. W., Marin, E. C., Otto, N., Dreher, M., Dempsey, G., Stark, I., Bates, A. S., Pleijzier, M. W., Schlegel, P., et al. (2020). The connectome of the adult drosophila mushroom body provides insights into function. *Elife*, 9:e62576.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Liu, S., Kailkhura, B., Loveland, D., and Han, Y. (2019). Generative counterfactual introspection for explainable deep learning. *arXiv preprint arXiv:1907.03077*.

Liu, T., Jurrus, E., Seyedhosseini, M., Ellisman, M., and Tasdizen, T. (2012). Watershed merge tree classification for electron microscopy image segmentation. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 133–137. IEEE.

Long, X., Colonell, J., Wong, A. M., Singer, R. H., and Lionnet, T. (2017). Quantitative mrna imaging throughout the entire drosophila brain. *nature methods*, 14(7):703.

Lovick, J. K., Ngo, K. T., Omoto, J. J., Wong, D. C., Nguyen, J. D., and Hartenstein, V. (2013). Postembryonic lineages of the *Drosophila* brain: I. development of the lineage-associated fiber tracts. *Dev. Biol.*, 384(2):228–257.

Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Paul Smolley, S. (2017). Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802.

Marin, E. C., Roberts, R. J. V., Büld, L., Theiss, M., Pleijzier, M. W., Sarkissian, T., Laursen, W. J., Turnbull, R., Schlegel, P., Bates, A. S., Li, F., Landgraf, M., Costa, M., Bock, D. D., Garrity, P. A., and Gregory S X (2020). Connectomics analysis reveals first, second, and third order thermosensory and hygrosensory neurons in the adult *Drosophila* brain.

Martens, D. and Provost, F. (2014). Explaining data-driven document classifications. *Mis Quarterly*, 38(1):73–100.

McKellar, C. E. and Wyttenbach, R. A. (2017). A protocol demonstrating 60 different drosophila behaviors in one assay. *Journal of Undergraduate Neuroscience Education*, 15(2):A110.

Meinertzhagen, I. A. and O'neil, S. (1991). Synaptic organization of columnar elements in the lamina of the wild type in drosophila melanogaster. *Journal of comparative neurology*, 305(2):232–263.

Meissner, G. W., Nern, A., Singer, R. H., Wong, A. M., Malkesman, O., and Long, X. (2019). Mapping neurotransmitter identity in the whole-mount drosophila brain using multiplex high-throughput fluorescence in situ hybridization. *Genetics*, 211(2):473–482.

Menzel, R. (2012). The honeybee as a model for understanding the basis of cognition. *Nature Reviews Neuroscience*, 13(11):758–768.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222.

Mothilal, R. K., Sharma, A., and Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617.

Narayanaswamy, A., Venugopalan, S., Webster, D. R., Peng, L., Corrado, G. S., Ruamviboonsuk, P., Bavishi, P., Brenner, M., Nelson, P. C., and Varadarajan, A. V. (2020). Scientific discovery by generating counterfactuals using image translation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 273–283. Springer.

Nässel, D. R. (2018). Substrates for neuronal cotransmission with neuropeptides and small molecule neurotransmitters in *Drosophila*. *Front. Cell. Neurosci.*, 12:83.

Nogales, E. (2000). Structural insights into microtubule function. *Annual Review of Biochemistry*, 69(1):277–302.

Ohyama, T., Schneider-Mizell, C. M., Fetter, R. D., Aleman, J. V., Franconville, R., Rivera-Alba, M., Mensh, B. D., Branson, K. M., Simpson, J. H., Truman, J. W., et al. (2015). A multilevel multimodal circuit enhances action selection in drosophila. *Nature*, 520(7549):633–639.

Okada, R., Awasaki, T., and Ito, K. (2009). Gamma-aminobuyric acid (GABA)-mediated neural connections in the *Drosophila* antennal lobe. *J. Comp. Neurol.*, 514(1):74–91.

Otto, N., Pleijzier, M. W., Morgan, I. C., Edmonson-Stait, A. J., Heinz, K. J., Stark, I., Dempsey, G., Ito, M., Kapoor, I., Hsu, J., et al. (2020). Input connectivity reveals additional heterogeneity of dopaminergic reinforcement in drosophila. *bioRxiv*.

Owald, D. and Waddell, S. (2015). Olfactory learning skews mushroom body output pathways to steer behavioral choice in drosophila. *Current opinion in neurobiology*, 35:178–184.

Palay, S. L. (1956). Synapses in the central nervous system. *The Journal of Cell Biology*, 2(4):193–202.

Pearce, J. M. (2009). Marie-jean-pierre flourens (1794–1867) and cortical localization. *European neurology*, 61(5):311–314.

Peters, A. and Folger, C. (2013). A website entitled the fine structure of the aging brain.

Plaza, S. M. (2014). Focused proofreading: efficiently extracting connectomes from segmented em images. *arXiv preprint arXiv:1409.1199*.

Prokop, A. and Meinertzhagen, I. A. (2006). Development and structure of synaptic contacts in *Drosophila*. *Semin. Cell Dev. Biol.*, 17(1):20–30.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Rohlfing, T. and Maurer, C. R., J. (2003). Nonrigid image registration in shared-memory multiprocessor environments with application to brains, breasts, and bees. *IEEE Trans. Inf. Technol. Biomed.*, 7(1):16–25.

Rolnick, D., Meirovitch, Y., Parag, T., Pfister, H., Jain, V., Lichtman, J. W., Boyden, E. S., and Shavit, N. (2017). Morphological error detection in 3d segmentations. *arXiv preprint arXiv:1705.10882*.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216.

Ryan, K., Lu, Z., and Meinertzhagen, I. A. (2016). The cns connectome of a tadpole larva of ciona intestinalis (l.) highlights sidedness in the brain of a chordate sibling. *Elife*, 5:e16962.

Saalfeld, S., Cardona, A., Hartenstein, V., and Tomančák, P. (2009). Catmaid: collaborative annotation toolkit for massive amounts of image data. *Bioinformatics*, 25(15):1984–1986.

Saalfeld, S., Fetter, R., Cardona, A., and Tomancak, P. (2012). Elastic volume reconstruction from series of ultra-thin microscopy sections. *Nature methods*, 9(7):717–720.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Müller, K.-R. (2016). Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673.

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., and Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278.

Sayin, S., De Backer, J.-F., Siju, K. P., Wosniack, M. E., Lewis, L. P., Frisch, L.-M., Gansen, B., Schlegel, P., Edmondson-Stait, A., Sharifi, N., Fisher, C. B., Calle-Schuler, S. A., Lauritzen, J. S., Bock, D. D., Costa, M., Jefferis, G. S. X. E., Gjorgjieva, J., and Grunwald Kadow, I. C. (2019). A neural circuit arbitrates between persistence and withdrawal in hungry *Drosophila*. *Neuron*, 104(3):544–558.e6.

Scheffer, L. K., Xu, C. S., Januszewski, M., Lu, Z., Takemura, S.-y., Hayworth, K. J., Huang, G. B., Shinomiya, K., Maitlin-Shepard, J., Berg, S., et al. (2020). A connectome and analysis of the adult drosophila central brain. *Elife*, 9:e57443.

Schneider-Mizell, C. M., Gerhard, S., Longair, M., Kazimiers, T., Li, F., Zwart, M. F., Champion, A., Midgley, F. M., Fetter, R. D., Saalfeld, S., et al. (2016). Quantitative neuroanatomy for connectomics in drosophila. *eLife*, 5:e12059.

Schubert, P. J., Dorkenwald, S., Januszewski, M., Jain, V., and Kornfeld, J. (2019). Learning cellular morphology with neural networks. *Nature communications*, 10(1):1–12.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Sen, S. (2019). Neurotransmitter identity: A question of lineage. *eLife*, 8:e47162.

Shen, F. Y., Harrington, M. M., Walker, L. A., Cheng, H. P. J., Boyden, E. S., and Cai, D. (2020). Light microscopy based approach for mapping connectivity with molecular specificity. *Nature communications*, 11(1):1–12.

Sheridan, A., Nguyen, T., Deb, D., Lee, W.-C. A., Saalfeld, S., Turaga, S., Manor, U., and Funke, J. (2021). Local shape descriptors for neuron segmentation. *bioRxiv*.

Shinomiya, K., Takemura, S.-Y., Rivlin, P. K., Plaza, S. M., Scheffer, L. K., and Meinertzhagen, I. A. (2015). A common evolutionary origin for the ON- and OFF-edge motion detection pathways of the *Drosophila* visual system. *Front. Neural Circuits*, 9:33.

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR.

Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. (2016). Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Sommer, C., Straehle, C. N., Koethe, U., Hamprecht, F. A., et al. (2011). Ilastik: Interactive learning and segmentation toolkit. In *ISBI*, volume 2, page 8.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

Staffler, B., Berning, M., Boergens, K. M., Gour, A., van der Smagt, P., and Helmstaedter, M. (2017). Synem, automated synapse detection for connectomics. *Elife*, 6:e26414.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328. JMLR. org.

Takemura, S.-y., Aso, Y., Hige, T., Wong, A., Lu, Z., Xu, C. S., Rivlin, P. K., Hess, H., Zhao, T., Parag, T., et al. (2017). A connectome of a learning and memory center in the adult drosophila brain. *Elife*, 6:e26975.

Takemura, S.-y., Bharioke, A., Lu, Z., Nern, A., Vitaladevuni, S., Rivlin, P. K., Katz, W. T., Olbris, D. J., Plaza, S. M., Winston, P., et al. (2013). A visual motion detection circuit suggested by drosophila connectomics. *Nature*, 500(7461):175.

Takemura, S.-y., Xu, C. S., Lu, Z., Rivlin, P. K., Parag, T., Olbris, D. J., Plaza, S., Zhao, T., Katz, W. T., Umayam, L., et al. (2015). Synaptic circuits and their variations within different columns in the visual system of drosophila. *Proceedings of the National Academy of Sciences*, 112(44):13711–13716.

Tan, S. (2018). Interpretable approaches to detect bias in black-box models. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 382–383.

Tanaka, N. K., Endo, K., and Ito, K. (2012). Organization of antennal lobe-associated neurons in adult *Drosophila* melanogaster brain. *J. Comp. Neurol.*, 520(18):4067–4130.

Turaga, S. C., Briggman, K. L., Helmstaedter, M., Denk, W., and Seung, H. S. (2009). Maximin affinity learning of image segmentation. *arXiv preprint arXiv:0911.5372*.

Turner-Evans, D. B., Jensen, K., Ali, S., Paterson, T., Sheridan, A., Ray, R. P., Lauritzen, S., Bock, D., and Jayaraman, V. (2019). The neuroanatomical ultrastructure and function of a biological ring attractor.

Uchigashima, M., Ohtsuka, T., Kobayashi, K., and Watanabe, M. (2016). Dopamine synapse is a neuroligin-2–mediated contact between dopaminergic presynaptic and gabaergic postsynaptic structures. *Proceedings of the National Academy of Sciences*, 113(15):4206–4211.

Verma, S., Dickerson, J., and Hines, K. (2020). Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*.

Von Ardenne, M. (1938). Das elektronen-rastermikroskop. *Zeitschrift für Physik*, 109(9):553–572.

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841.

Waddell, S. (2013). Reinforcement signalling in drosophila; dopamine does it all after all. *Current opinion in neurobiology*, 23(3):324–329.

Wang, P. and Vasconcelos, N. (2020). Scout: Self-aware discriminant counterfactual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8981–8990.

White, J. G., Southgate, E., Thomson, J. N., and Brenner, S. (1986). The structure of the nervous system of the nematode caenorhabditis elegans. *Philos Trans R Soc Lond B Biol Sci*, 314(1165):1–340.

Williams, D. B. and Carter, C. B. (1996). The transmission electron microscope. In *Transmission electron microscopy*, pages 3–17. Springer.

Wilson, R. I. and Laurent, G. (2005). Role of GABAergic inhibition in shaping odor-evoked spatiotemporal patterns in the *Drosophila* antennal lobe. *J. Neurosci.*, 25(40):9069–9079.

Witvliet, D., Mulcahy, B., Mitchell, J. K., Meirovitch, Y., Berger, D. K., Wu, Y., Liu, Y., Koh, W. X., Parvathala, R., Holmyard, D., et al. (2020). Connectomes across development reveal principles of brain maturation in c. elegans. *BioRxiv*.

Xiao, C., Chen, X., Li, W., Li, L., Wang, L., Xie, Q., and Han, H. (2018). Automatic mitochondria segmentation for em data using a 3d supervised convolutional network. *Frontiers in Neuroanatomy*, 12:92.

Xu, C. S., Januszewski, M., Lu, Z., Takemura, S.-y., Hayworth, K. J., Huang, G., Shinomiya, K., Maitin-Shepard, J., Ackerman, D., Berg, S., Blakely, T., Bogovic, J., Clements, J., Dolafi, T., Hubbard, P., Kainmueller, D., Katz, W., Kawase, T., Khairy, K. A., Leavitt, L., Li, P. H., Lindsey, L., Neubarth, N., Olbris, D. J., Otsuna, H., Troutman, E. T., Umayam, L., Zhao, T., Ito, M., Goldammer, J., Wolff, T., Svirskas, R., Schlegel, P., Neace, E. R., Knecht, C. J., Alvarado, C. X., Bailey, D. A., Ballinger, S., Borycz, J. A., Canino, B. S., Cheatham, N., Cook, M., Dreher, M., Duclos, O., Eubanks, B., Fairbanks, K., Finley, S., Forknall, N., Francis, A., Hopkins, G. P., Joyce, E. M., Kim, S., Kirk, N. A., Kovalyak, J., Lauchie, S. A., Lohff, A., Maldonado, C., Manley, E. A., McLin, S., Mooney, C., Ndama, M., Ogundeyi, O., Okeoma, N., Ordish, C., Padilla, N., Patrick, C., Paterson, T., Phillips, E. E., Phillips, E. M., Rampally, N., Ribeiro, C., Robertson, M. K., Rymer, J. T., Ryan, S. M., Sammons, M., Scott, A. K., Scott, A. L., Shinomiya, A., Smith, C., Smith, K., Smith, N. L., Sobeski, M. A., Suleiman, A., Swift, J., Takemura, S., Talebi, I., Tarnogorska, D., Tenshaw, E., Tokhi, T., Walsh, J. J., Yang, T., Horne, J. A., Li, F., Parekh, R., Rivlin, P. K., Jayaraman, V., Ito, K., Saalfeld, S., George, R., Meinertzhagen, I., Rubin, G. M., Hess, H. F., Scheffer, L. K., Jain, V., and Plaza, S. M. (2020). A connectome of the adult drosophila central brain. *bioRxiv*.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.

Zhao, T., Olbris, D. J., Yu, Y., and Plaza, S. M. (2018). Neutu: software for collaborative, large-scale, segmentation-based connectome reconstruction. *Frontiers in Neural Circuits*, 12:101.

Zheng, Z., Lauritzen, J. S., Perlman, E., Robinson, C. G., Nichols, M., Milkie, D., Torrens, O., Price, J., Fisher, C. B., Sharifi, N., et al. (2018). A complete electron microscopy volume of the brain of adult drosophila melanogaster. *Cell*, 174(3):730–743.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

Zintgraf, L. M., Cohen, T. S., Adel, T., and Welling, M. (2017). Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*.