Ben-Gurion University of the Negev

The Faculty of Natural Sciences

The Department of Computer Science

Applying in silico labeling via transfer learning as a tool to dissect organelle - organelle spatial dependencies

Thesis submitted in partial fulfillment of the requirements
for the Master of Sciences degree

Kathrine Smoliansky

Under the supervision of **Assaf Zaritsky and Andrei Sharf**

August 2022

Ben-Gurion University of the Negev

The Faculty of Natural Sciences

The Department of Computer Science

Applying in silico labeling via transfer learning as a tool to dissect organelle - organelle spatial dependencies

Thesis submitted in partial fulfillment of the requirements
for the Master of Sciences degree

Kathrine Smoliansky

Under the supervision of **Assaf Zaritsky and Andrei Sharf**

Signature of student: __________ Date: __________

Signature of supervisors: __________ __________ Date: 5/9/2022

Signature of chairperson of the
committee for graduate studies: __________ Date: __________

August 2022

Applying in silico labeling via transfer learning as a tool to dissect organelle - organelle spatial dependencies

Kathrine Smoliansky

Master of Sciences Thesis

Ben-Gurion University of the Negev

2022

Abstract

How organelles act in concert to shape and enable cell function is a fundamental question in cell biology. Progress in the field is hindered by lack of systematic tools to dissect spatiotemporal interorganelle interactions, mainly due to technical limitations in simultaneous labeling of multiple organelles within the same living cell. We combine emerging computational techniques and a publicly available comprehensive dataset of 3D cell imaging to validate known relations and to raise new hypotheses regarding inter-organelle spatial dependencies.

Modern machine-learning techniques were recently shown to successfully extract the locations of organelles within the cell from label-free images, a technique called “in silico labeling”. Transfer learning is a machine learning technique where computational models trained for one task are used as a starting point to train a new model to a different but related task. The intuition for the success of transfer learning stems from the understanding that re-training models using better initial states can provide benefits when training data are limited in volume. We use this property as means to measure spatial dependencies between different organelles. If an in silico labeling model trained to localize an organelle is improved by transferring a model trained to localize a different organelle, than we say that the information encoded in the mapping to the latter organelle is useful for the prediction of the former organelle. Such comparison between two models defines an asymmetric link that encodes a spatial dependency between the latter and the former organelles. We applied this methodology to construct an inter-organelle spatial dependencies network by integrating the pairwise relations between 13 organelles in 3D imaged endogenously labeled human-induced pluripotent stem cells available by the Allen Institute of Cell Science.

Our preliminary results validate several known relations between organelles and propose new predictions that should be validated experimentally. For example, both the nuclear envelope and the endoplasmic reticulum (ER) contribute to the prediction of the golgi apparatus, and the nuclear envelope contributes to the prediction of the ER. The golgi and the tight junctions are linked bi-directionally. The plasma membrane contributes to the prediction of adherens junctions. No organelle enhances the prediction of desmosomes.

Ultimately, our project will provide a comprehensive and validated methodology for predicting organelle-organelle interactions, bypassing long-standing technical barriers in microscopy. The establishment of a rich resource of predicted organelle-organelle interaction networks can provide a major leap towards the “holy grail” of cell biology - inclusive understanding of cells as integrated complex systems.

Acknowledgements

First and foremost, I want to express my gratitude to my supervisors Assaf Zaritsky and Andrei Sharf for their constructive criticism and patience. They provided me with direction and help throughout all the phases of my project.

Additionally, I want to express my gratitude to Mazal, Roni, and Mila for their excellent assistance and thorough knowledge of all administrative inquiries.

I would like to thank all the excellent professors I got the chance to speak with and learn from.

I am also grateful to my classmates and lab-mates.

An enormous thank you to my beloved friends, family, and significant other, who have been by my side and believed in me even when I didn't.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

“Organelle- a specialized cellular part (such as a mitochondrion, chloroplast, or nucleus) that has a specific function and is considered analogous to an organ”, -definition by Merriam Webster dictionary.

Organelles, meaning ‘small organ’ are the specialized subcellular structures which allow it to perform its functions. Proper cell function requires multiple organelles to act in concert involving physical and chemical interactions between different organelles. These interactions can be reflected by specific localization patterns and spatial dependencies between organelles [2–9]. Perturbed organelle-organelle interactions can lead to impaired functionality at the cellular level and disease [10, 11]. While many studies focus on the role and structure of individual organelles, much less is known about how different organelles coordinate their organization and function within the cell, a fundamental requirement for the observed integrated complex cell function [2–9].

Microscopy is the only technology that can enable determination of the organelles' intracellular localization. However, imaging and monitoring the dynamics of multiple organelles simultaneously in the same living cell is a challenging task. [5], [12]. Progress in this area has been confounded by the difficulty to achieve an integrated representation of cellular organization due to the limited number of fluorophores that can be simultaneously tagged and distinguished in a microscopy image [13].

To bridge these gaps [1, 14] presented that modern machine learning techniques can computationally infer specific organelle localization patterns from label-free microscopy images. The method presented a convolutional neural networks (CNN)-based tool, employing a U-Net architecture [15] [Methods] to model relationships between distinct but correlated imaging modalities, predicting corresponding fluorescence images directly from three-dimensional (3D) transmitted-light microscopy live cell images. The label-free prediction tool learns each relationship between 3D transmitted-light and fluorescence live cell images for several major subcellular structures. A resultant model can then predict a 3D fluorescence image from a new transmitted-light input. A single transmitted-light input can be applied to multiple subcellular structure models, enabling multi-channel, integrated fluorescence imaging.

Importantly, the method considers prediction of each type of organelle as a separate task which requires its own trained model. At inference, the network generates an image reconstructing the localization of an organelle within the cell from a label free image. Using this method, it becomes possible to annotate many organelle in the living cell by overlaying each network's prediction of the same input image. This process is non-destructive to the cellular structure. These results were replicated in other cellular systems and label-free microscopy techniques [1, 14]. While, in principle, "in silico labeling" can provide an integrated view of a live cell, all of these studies did not go beyond a proof of principle and did not provide practical tools for direct application toward biological discovery.

In my dissertation I developed a method that uses in silico labeling to indirectly dissect organelle-organelle interactions by computationally integrating single-organelle data from multiple experiments.Since the various networks were trained independently of one another on images with only one type of organelle labeled, it is hard to extrapolate information regarding holistic, multi-organelle relations from those trained models. The general concept relied on leveraging information learned by a model trained to localize one organelle to a second model trained to localize another organelle. When information learned to map one organelle enhances the mapping to a different organelle we can hypothesize that these two organelles are spatially dependent.

Transfer learning is a method used to transfer knowledge acquired from one task to resolve another. By transferring the knowledge found in various but related source domains, it is possible to increase the performance of target learners on target domains [16, 17]. Protein folding, speech recognition, object detection, labeling cancer cells, object tracking and recommender systems, are a few examples of tasks that have benefited from transfer learning [18–21]. In my thesis I demonstrated a proof of principle study showing that transfer learning applied to in silico labeling can be used as a tool to dissect organelle - organelle spatial dependencies. This technology will change the way we look at biology and give biologists specific ideas for fresh experimental setups to better understand how cells are structured inside.

Chapter 2

Related Work

Due to the multidisciplinary nature of my study, I have covered three main topics: organelle-organelle interactions, in-silico labeling and transfer learning.

2.1 Organelle-Organelle interactions

The cell is divided into distinct membrane-bound organelles to enable the simultaneous orchestration of complex and frequently incompatible biochemical processes. Organelle interactions are now understood to be essential for a variety of cellular processes [2, 22, 23]. However, due to the finite number of labels that fluorescence imaging techniques can detect in a single image, the spatial and temporal organization of organelles within the cell is still poorly understood [5, 24]. An important feature to notice when talking about Organelle-Organelle interactions is contact sites between organelles, contact sites have many important roles such as communication, regulation and transport of molecules are only a few described in [5, 9, 25]. Valm Et Al [5] presented a systems-level analysis of the organelle interactome using a multispectral image acquisition method that overcomes the challenge of spectral overlap in the fluorescent protein palette. Using confocal and lattice light sheet instrumentation and an imaging informatics pipeline of five steps to achieve mapping of organelle numbers, volumes, speeds, positions and dynamic inter-organelle contacts in live cells. Describing the frequency and locality of two-, three-, four- and five- way interactions among six different membrane-bound organelles (endoplasmic reticulum, Golgi, lysosome, peroxisome, mitochondria and lipid droplet) and showing how these relationships change over time. The paper demonstrates that each organelle has a characteristic distribution and dispersion pattern in three-dimensional space and that there is a reproducible pattern of contacts among the six organelles that are affected by microtubule and cell nutrient status. Another recent attempt to elucidate the subcellular organization of multiple proteins was done by integrating and aligning localization of separately imaged mitotic proteins and constructing a comprehensive and quantitative canonical (i.e., “average”) 4D model of mitotic protein localization network [26]. While this spatio-temporal standardization of a dynamic cellular process holds great promise, it reduces a complex heterogeneous process to a single plausible trajectory that does not capture the entire variability in the diverse phenotypic landscape. Cho et al. [27] Constructed a library of 1310 fluorescently tagged cell lines, and generated a large dataset that maps the cellular localization and physical interactions of the corresponding 1310 proteins. Applying a combination of unsupervised clustering and machine learning for image analysis allowed to objectively identify proteins that share spatial or interaction signatures. The data provides insights into the function of individual proteins, but also enables to derive some general principles of human cellular organization. In particular, it shows that proteins that bind RNA form a separate subgroup defined by specific localization and interaction signatures. Important finding is that the precise spatial distribution of a given protein is very strongly correlated with its cellular function. Viana et al. applied new methods to determine how a subset of expressed genes dictate cellular phenotypes. To address this challenge [28] created novel methods to transform raw cellular image data of cells and their structures into dimensionally reduced information in a form that accommodates the vast cell-to-cell diversity and both summarizes the raw data. They determined, in an integrated, holistic way, where, how much and how variable the various cellular structures are. The initial objective was to identify quantitative relationships, or rules of organization, and then used these data to develop benchmarks for understanding and predicting cellular organization in a wide variety of biological contexts. “OpenOrganelle” is a new data portal containing electron microscopy (FIB-SEM) cell images, their corresponding automated segmentation of 35 organelles, and the code and models for analyzing these data[25, 29], a high-quality resource to study inter-organelle organization in high resolution. However this approach suffers from limited throughput, difficulties in following dynamic processes and in testing the effects of perturbations.

2.2 In-Silico Labeling and Label-free prediction

In the life sciences, microscopy is a crucial technique. Physical fluorescent labels are added to particular cellular components using a variety of techniques, including antibody labeling. However, these methods suffer from a number of important flaws, such as consistency issues, spectrum overlap restrictions on the number of simultaneous labels, and the need to disrupt the experiment in order to get the measurement, such as by fixing the cells. To get around these issues, computational machine-learning methods were developed that can accurately predict some fluorescent labels from transmitted-light images of biological samples that aren't labeled. While labeling whole cells [30–32] or even explicitly the nucleus [33, 34] is a relatively widely researched topic , however labeling finer details in a cell is a relatively novel field of research.

Recent advancements in this field include two publications demonstrating a label-free method for predicting (three-dimensional) fluorescence labels of multiple organelles directly from transmitted light images by training deep neural networks. This was possible by virtue of the recently available dataset from numerous experiments across various labs consisting of pairs of transmitted-light z-stack images and fluorescence images that are pixel registered.[1, 14, 35]. The results in [1] were remarkable for larger structures with more structured characteristics, like the nucleus, mitochondria, and actin filaments. However smoller structures including the golgi apparatus and desmosomes had less significant results. While [14] addressed the prediction of the location and texture of cell nuclei, the health of a cell, and the type of cells in a mixture, [1] concentrated on using their method on a variety of different structures. The work by Christiansen et al. contained data integrated from 3 labs, whereas ounkomol et al. had only one data source. Christiansen et al., also experimented with transfer learning: once trained to predict a set of labels, the network could learn new labels with a small number of additional data, resulting in a highly generalizable algorithm, adaptable across experiments. It has been noticed previously that when learning structure shape, using a combination of other related structures can benefit the results [36].

In a similar manner [37] demonstrate a digital staining method they call PhaseStain that aims to convert quantitative phase images (QPI) of label-free tissue slices into images comparable to brightfield microscopy images of the same samples that have been histologically stained. They used skin, kidney, and liver tissue to train three deep neural network models. reporting the GAN-based staining framework's high-fidelity performance. Worth mentioning Quantitative label-free imaging with phase and polarization (QLIPP), a new computational imaging technique for label-free determination of density and anisotropy from 3D polarization-resolved data, has been published by the authors. [38] demonstrate how different organelles' density and anisotropy can be used to identify them. They also demonstrate the capability of label-free identification of numerous brain tissue areas.

2.3 Transfer learning

Transfer learning (TL) is a method used to transfer knowledge acquired from one task to resolve another[16, 17]. For instance, before learning to fly a real airplane, pilots use a variety of simulations, such as video games, to gain prior skills, knowledge, and experience under the assumption that knowledge gained while learning one task can be a good starting point when learning another task.

In the context of machine learning, By transferring the knowledge found in various but related source domains, TL seeks to increase the performance of target learners on target domains [16, 39–43]. This is very useful when data in the target domain is scarce and it is not possible to train a good model with it, so TL gives the model a head start. Sometimes TL is also used to direct the model in a specific direction when the target task is not easily deducible from the target data.

The idea of transfer learning was first discussed in the context of education in the last century, it was first considered in a computational context in 1995, and went through conceptual refinement in 2005 to the manner that we comprehend it today [44].

It is most common to categorize transfer learning under three subsettings, inductive transfer learning, transductive transfer learning, and unsupervised transfer learning, based on different situations between the source and target domains and tasks [43, 44]. Approaches to transfer learning in the above three different settings can be summarized into four cases based on “What to transfer". The first context can be referred to as instance-based transfer learning (or instance transfer) approach [43, 45, 46] which assumes that certain parts of the data in the source domain can be reused for learning in the target domain by reweighting. Instance reweighting and importance sampling are two major techniques in this context.

Employing transfer learning with convolutional neural networks (CNNs), well-trained on non-medical ImageNet dataset, has shown promising results for medical image analysis in recent years. MA Morid et al. conduct a scoping review to identify these studies and summarize their characteristics in terms of the problem description, input, methodology, and outcome [47].

Other recent studies incorporate transfer learning to boost research on biological data, Wang et al. [48] presented BERMUDA (Batch Effect ReMoval Using Deep Autoencoders), a transfer-learning-based method for batch effect correction in scRNA-seq data. BERMUDA effectively combines different batches of scRNA-seq data with vastly different cell population compositions and amplifies biological signals by transferring information among batches. BERMUDA outperformed existing methods for removing batch effects and distinguishing cell types in multiple simulated and real scRNA-seq datasets.

Widmer, C., and Rätsch, G. [49] present two problems from sequence biology, where multitask learning was successfully applied.

Stumpf, Patrick S., et al. [50]shows that transfer learning can be used to efficiently map bone marrow biology between species, using data obtained from single-cell RNA sequencing.

Sevakula, Rahul K., et al. [18] presented a transfer learning procedure for cancer classification, which uses feature selection and normalization techniques in conjunction with s sparse auto-encoders on gene expression data. While classifying any two tumor types, data of other tumor types were used in an unsupervised manner to improve the feature representation.

Zheng, Shijie C., et al.[51] present tricycle, an R/Bioconductor package. By leveraging key features of the biology of the cell cycle, the mathematical properties of principal component analysis of periodic functions, and the use of transfer learning. They estimate a cell-cycle embedding using a fixed reference dataset and project new data into this reference embedding, an approach that overcomes key limitations of learning a dataset-dependent embedding. Tricycle then predicts a cell-specific position in the cell cycle based on the data projection.

Applying TL in microscopy imaging in order to enhance the imaging is becoming very effective and being used in several recent researches. To produce high-resolution slices of biopsy images from low-resolution ones, [52] suggest a combined architecture that combines a unique transfer learning technique and a deep super-resolution framework.

[53] introduces a versatile reconstruction method, ML-Structured illumination microscopy (SIM), which makes use of transfer learning to obtain a parameter-free model that allows a doubling of image resolution at speeds compatible with live-cell imaging.

[54] applied convolutional neural networks and transfer learning can be used to identify cancer tissue from real-time in vivo imaging with confocal laser microscopy In another research, Transfer learning was employed in the form

of fine-tuning.

[55] was able to effectively classify an image into four classes: normal, benign, invasive carcinoma, and in situ carcinoma. They produced results with an accuracy of 98.33 percent and sensitivity of 98.44 percent.

Importantly, the transferred knowledge does not always bring a positive impact on new tasks. If there is little in common between domains, knowledge transfer could be unsuccessful [17, 42, 56, 57].

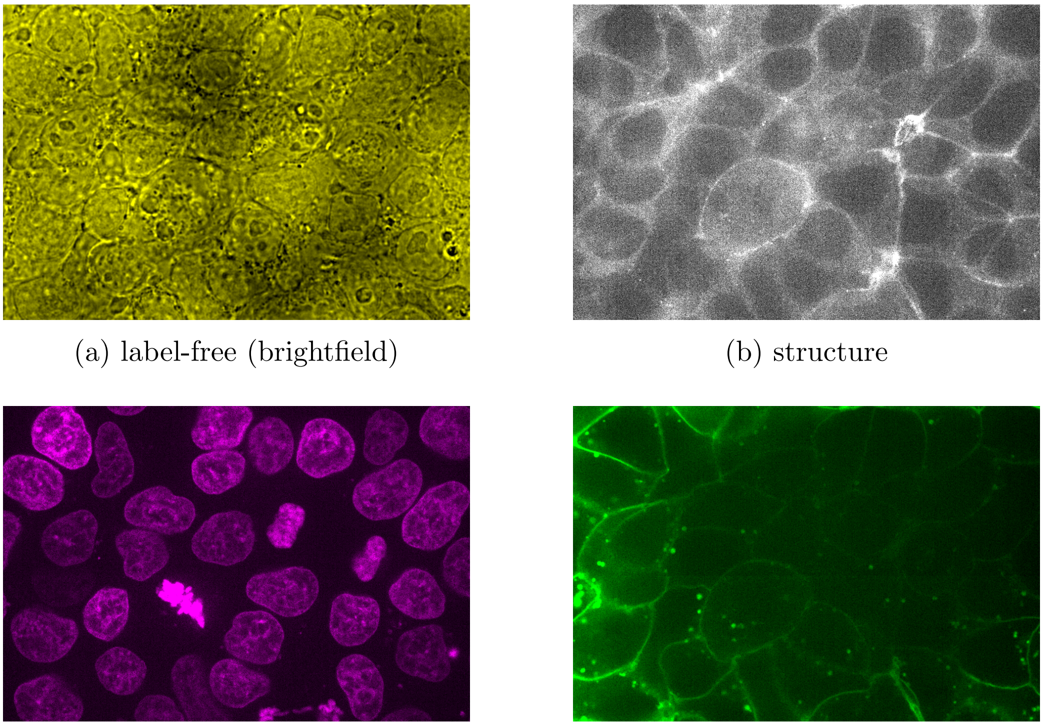
Chapter 3

Experimental Results

3.1 Pre-processing

Prior to diving into our primary outcomes, it's critical to comprehend the data and prepare it for further analysis. My research relied on a unique publicly available comprehensive 3D cell imaging datasets (Full description is available in Methods), from the Allen Institute of Cell Science [58]. The Allen Institute genetically edited over 40 high quality-certified fluorescently tagged human induced pluripotent stem cell lines (hiPSC) that target nineteen specific key cellular membrane-bound organelles and fundamental cellular structured components [59, 60] and imaged them in their native conditions and under several well characterized perturbations – a perfect dataset to systematically characterize inter-organelle interactions.

The Allen institute genetically edited different cell lines, each with a different key organelle fluorescently tagged. This dataset included images from nineteen different gene edited human stem cell lines, each organelle sub-dataset includes at least 350 3D images, each with a matching label-free (brightfield) and a labeled (fluorescent) channel as well as two additional channels with labeled organelles that are common to all cells, the cell membrane and nucleus (Fig.3.1).



(a) label-free (brightfield)

(b) structure

(c) additional channel of labeled organelle nucleus.

(d) additional channel of labeled organelle cell membrane.

Figure 3.1: Representative middle slices of an image in our database, in this case the structure imaged is endoplasmic reticulum. (a) In yellow is the label-free (brightfield) channel. (b) In gray labeled structure of endoplasmic reticulum. (c) In purple an additional channel of labeled organelle nucleus. (d) In green an additional channel of labeled organelle cell membrane.

Not all organelles were centered at the same focal plane (z-axis). Under the assumption that the most focused focal plane will result with a less blurred image we measured the standard deviation in pixel brightness across the different imaged focal planes. While the majority of organelles, including the nucleus, showed a normal distribution (Fig.3.2) of standard deviation over the z-axis, some organelles showed non-normal distributions (Fig. 3.3) because certain organelles are not located at the cell's center (e.g., tight junctions are located near the cell's periphery). It became clear that top and bottom image slices do not contain information regarding the organelles of interest and thus they were discarded, leaving 40 z-slices around the cell's center (determined as the center z-slice of the nucleus) spanning 11.6 μm.

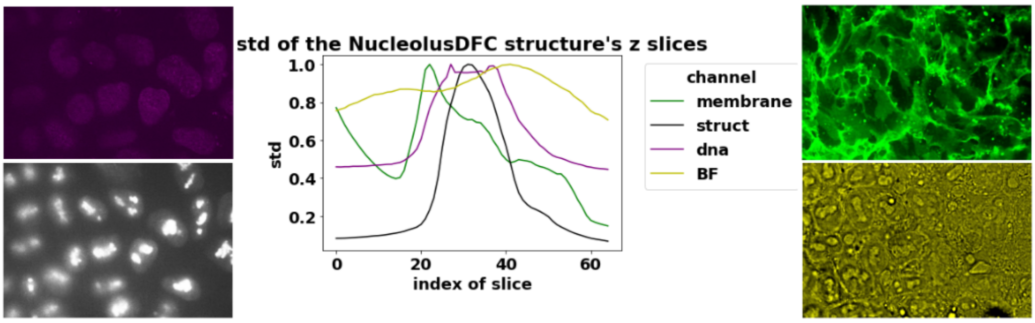
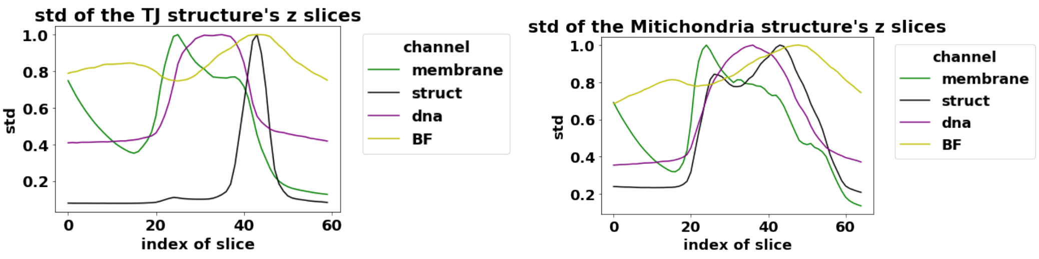


Figure 3.2: Standard deviation of pixel brightness per slice in z-axis for all channels in images of the Nucleus DFC organelle showing the mid slice of each channel correspondingly.



(a) Tight junctions

(b) Mitochondria

Figure 3.3: Standard deviation of pixel intensities (y-axis) per slice in the z-axis (x-axis in figure, "index of layer") of bright-field (BF), nucleus (dna), membrane, and an additional organelle (struct).

Our data suffered from day-today-variability or batch effects that can be caused by technical issues in the sample itself like difference in temperature, light exposure or slight variance in the growth medium. Other reasons might be, different imaging conditions or skills, as indicated by imaging days where in-silico labeling performance deteriorated compared to other days (Fig. 3.4) (chapter 4 Methods).

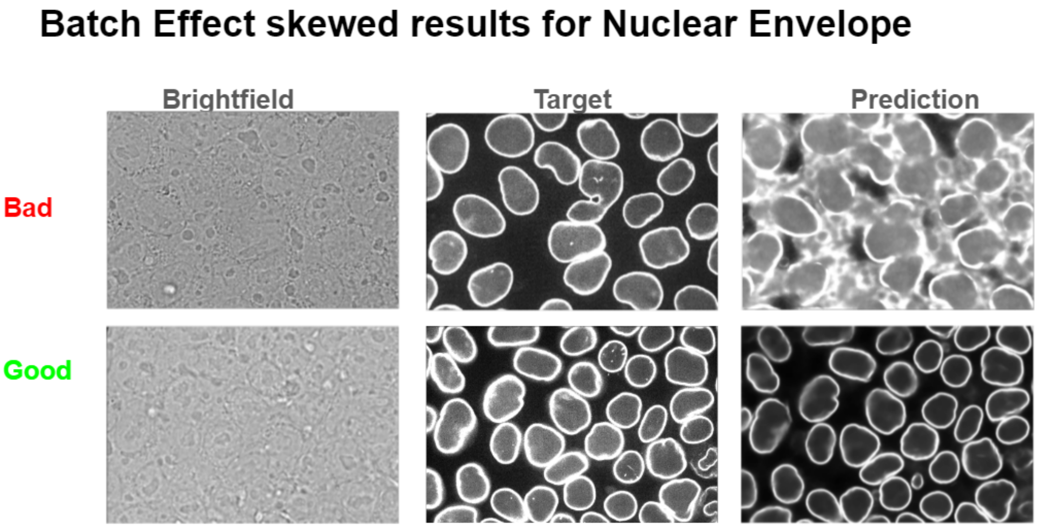


Figure 3.4: A comparison of mid slices of different images for the same structure, we can see that although the target image is looking very clear on both images the network performs differently, due to batch effects.

We analyzed datasets from different days as described at length in (chapter 4) and ended up with the dates listed below to be excluded from the full data-set (chapter 4 Methods, table 4.1).

:

Structure	Dates
Nuclear Envelope	08/05/2017, 09/05/2017, 12/05/2017
Endoplasmic-reticulum	28/07/2017
Actin-filaments	28/08/2017
Actomyosin-bundles	10/10/201,7 22/09/2017, 25/09/2017,26/09/2017
Tight-junctions	16/10/2017, 26/09/2017
Gap-junctions	06/03/2018 09/03/2018
Plasma-membrane	19/03/2018
Adherens-junctions	25/05/2018
Mitochondria	13/06/2017 , 09/06/2017
Desmosomes	19/07/2017 18/07/2017
Golgi	05/09/2017 , 25/08/2017, 29/08/2017
Centrosome	22/12/2017
Endosomes	12/10/2018

Table 3.1: Excluded days

Another criterion for exclusion was faulty images that were manually identified and discarded (Fig. 3.5).

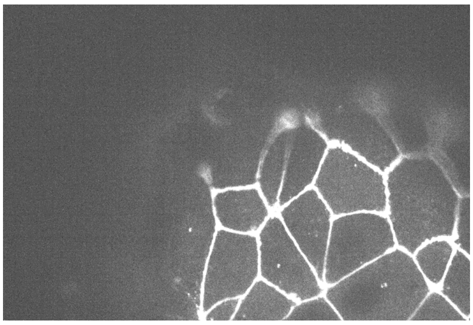
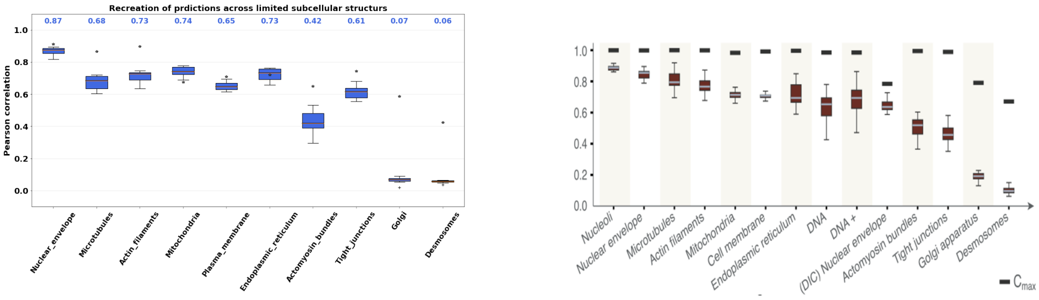


Figure 3.5: A corrupted image that was manually identified and excluded from our data set.

3.2 Reproducing Label Free and gaining a deeper insight into this method

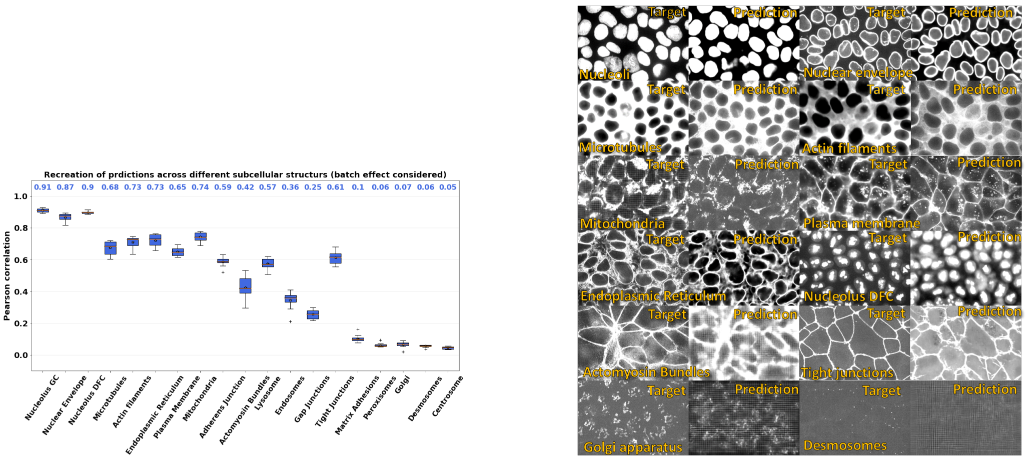
”For the things we have to learn before we can do them, we learn by doing them.” -Aristotle. It is crucial to understand the tools we are about to apply in this project. Therefore we reproduced Allen institute’s results by training models to predict organelles location from Bright Field (BF) images using the Allen institute fluorescently tagged images [1]. As presented in (Fig. 3.6) The models performance is very similar, slight variations may be seen due to variations in images used to train and test the models. Since the paper [1] was released more structures have been imaged and so we trained models for these structures (Fig. 3.7(a)). Examples of such predictions are presented in (Fig. 3.8(b)).



(a) Our reproduction of the Label-free method.

(b) The corresponding results that were published in [1].

Figure 3.6: Distributions of the image-wise Pearson correlation coefficient (r) between ground-truth (target) and predicted test images derived from the indicated subcellular structure models. Each target/predicted image pair in the test set is a point in the resultant r distribution; the 25th, 50th, and 75th percentile image pairs are spanned by the box for each indicated structure, with whiskers indicating the last data points within 1.5× the interquartile range of the lower and upper quartiles. The number of images in the test set was 10 for all distributions. (a)Our reproduction of the Label-free method. (b) The corresponding results that were published in [1].

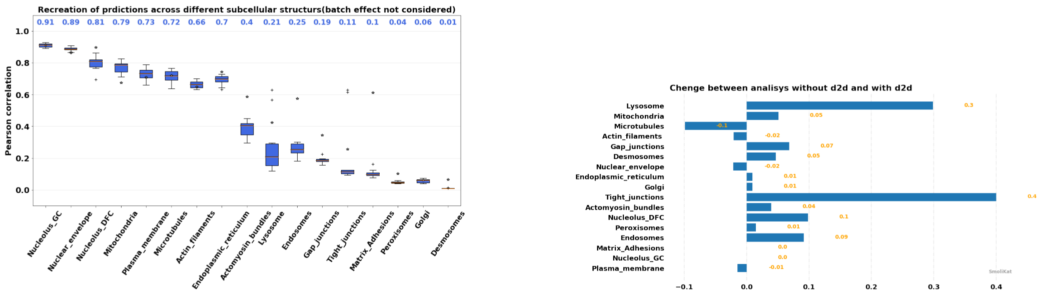


(a) Full asessment

(b) Representative labeled structure models and model predictions for 3D transmitted-light microscopy.

Figure 3.7: Extension of our reproduction of Label-free. (a) Data for the recreation was constructed with batch effects consideration, excluding images from problematic imaging days. (b) Representative labeled structure models and model predictions for 3D transmitted-light microscopy.

We trained models without considering batch effects to better comprehend the data as well as the constraints of the label free approach. The predictions of these models showed lower correlations in label free predictions. This analysis yielded a decline of 5.7 percent on average in accuracy (Fig. 3.8(a)). When comparing (Fig. 3.7(a)) and (Fig. 3.7(a)) the most notable impact was seen in organelles that are in the middle of the accuracy range, such as Tight Junction and Lysosome, while organelles at the extremes of the range, such as Desmosomes and Endosomes, on the low range, and Nucleus and Actin Filaments on the high range, showed little to no change.



(a) Partial results as explained in section C.

(b) The corresponding results to (a) that were published in [1].

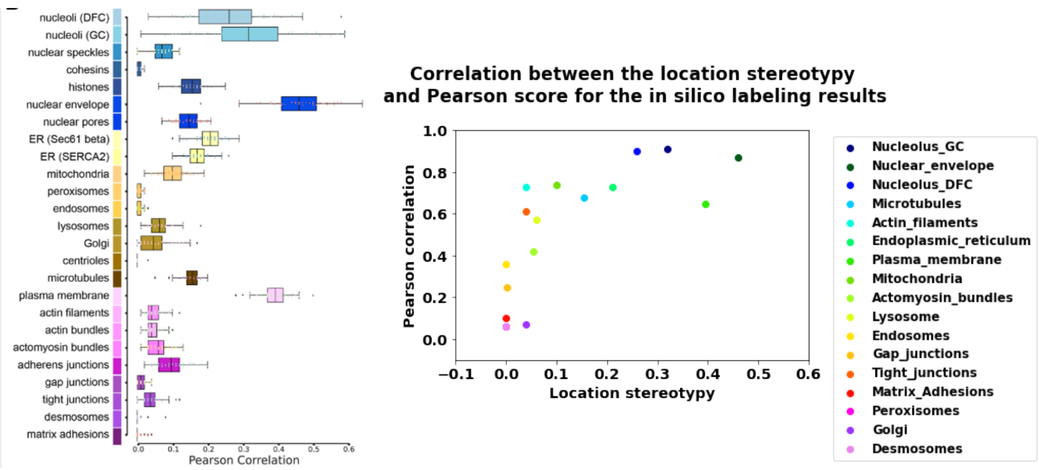
Figure 3.8: (a) Reproduction of Label Free without consideration of batch effects: the training dataset includes images from batch effect prone imaging days. (b) Representation of the variation in the models' performance with regard to the batch effect. The performance of the model that takes batch effects into consideration is significantly improved.

Not only the batch effect has an impact on this method but also it seems that there is a pattern where different groups of organelles are performing differently. Interestingly,the term of “stereotypy” -The extent to which a structure's individual location varied; was actually a side result of a larger research [28]. The aim was to develop generalizable, quantitative methods that permit direct comparison of the similarity of overall cell and nuclear shapes for 3D cell image data and achieve a simple and human-interpretable “shape space”. They developed A cell and nuclear shape-based coordinate system using a Principal Component Analysis (PCA)-based dimensional reduction approach. By Aligning all cells to their centroids. By morphing the parameterized intensity representation of each cell into the idealized cell and nuclear shape representing a nearby map point location in the shape space, they created a ‘morphed cell’.

To measure how variable the location of each individual structure is within the cell, required to calculate the 3D voxel-wise Pearson correlation between pairs of individual morphed cell images for each of the 25 cellular structures (Fig. 3.9(a)). Averaging those correlation values to generate a measure of the "location stereotypy" of each structure. Structures with a high stereotypy value have little cell-to-cell variation in their overall absolute positions for similarly shaped cells. Structures with a low stereotypy value may be found in distinct locations even for two cells whose shapes are very similar.

In (Fig. 3.9(b)) the correlation between organelle stereotypy and the performance of the model predicting this organelle is clear. The explanation is simple, if an organelle tends to appear in a fixed location it will be much easier to locate it then an organelle that may appear anywhere without any constraints. According to this reasoning, we expect that models that predict the more stereotypic organelles probably are able to do so by exploring and locating the immediate surroundings and neighbors of the structure, whereas for the less stereotypic structures, the model would probably have hard time finding such clues.

While batch effects are a general problem that affected the entire dataset, it is clear from the in-silico method that organelles that are more stereotypical, like plasma membrane and micro tubules, are less affected by this phenomenon and are maintaining high performance even when under the influence of batch effects, which is not the case for an organelle like a lysosome and tight junction.



(a) Box plots correspond to the values of the correlation for each of the 25 cellular structures. Figure adapted from [28]

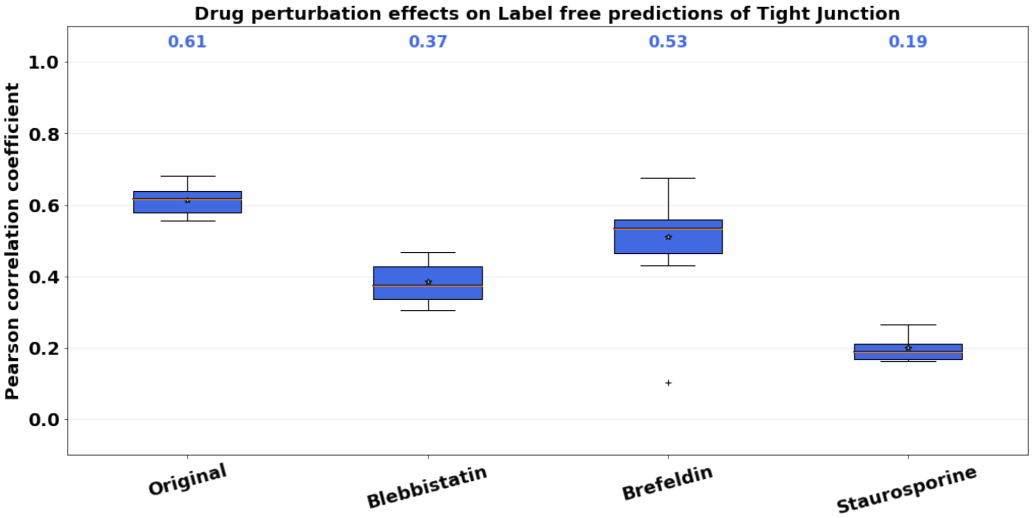
(b) Each organelles' in silico results are matched with the stereotypicality score it achieved. Showing high correlation between high performing models and stereotypicality of organelles location.

Figure 3.9: Correlation between the location stereotypy and Pearson score for the in silico labeling results.

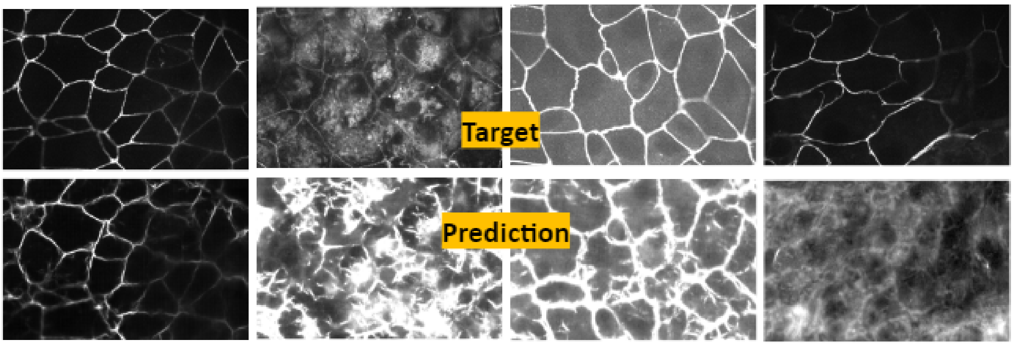
Further researching the models ability to exploit the structures' immediate environment led us to explore the impact perturbation have on the mapping between bright-field to the fluorescent label of that structure: We were interested in how the model's performance represented the effect that the drug has on the organelles' structural composition. To address this we have assessed the effects of various alterations on an organelle's structure and evaluated the effects of those changes on the prediction. Our hypothesis is that after the cells are perturbed the mapping between bright-field to fluorescent of that structure is also altered, making the models predictions inaccurate. The Allen Institute created a small data-set of various structures that were exposed to six different drugs using the Allen Cell Collection. Since Tight Junction was the only structure where we could visually see the impact of the drugs (as explained in Methods), we only tested three drug types that had visually distinct (Fig. 3.10(b)) structural effects on the structure. Understanding the drugs' mechanisms was crucial for understanding the results, The drugs are listed by the magnitude of the effect from minor to major:

- Brefeldin A- expected to see very little change if any.
- Blebbistatin- expected to see a decrease in correlation.
- Staurosporine- e expected to see a great decrease.

The results (Fig. 3.10) reflect that in-silico labeling performed worse after structure changing drug exposure, reinforcing our hypotheses the mapping from bright-field to the fluorescent target (TJ) has changed. This could be caused by direct change in TJ or because of changes in the surroundings and other organelles that the model is using to determine the TJ.



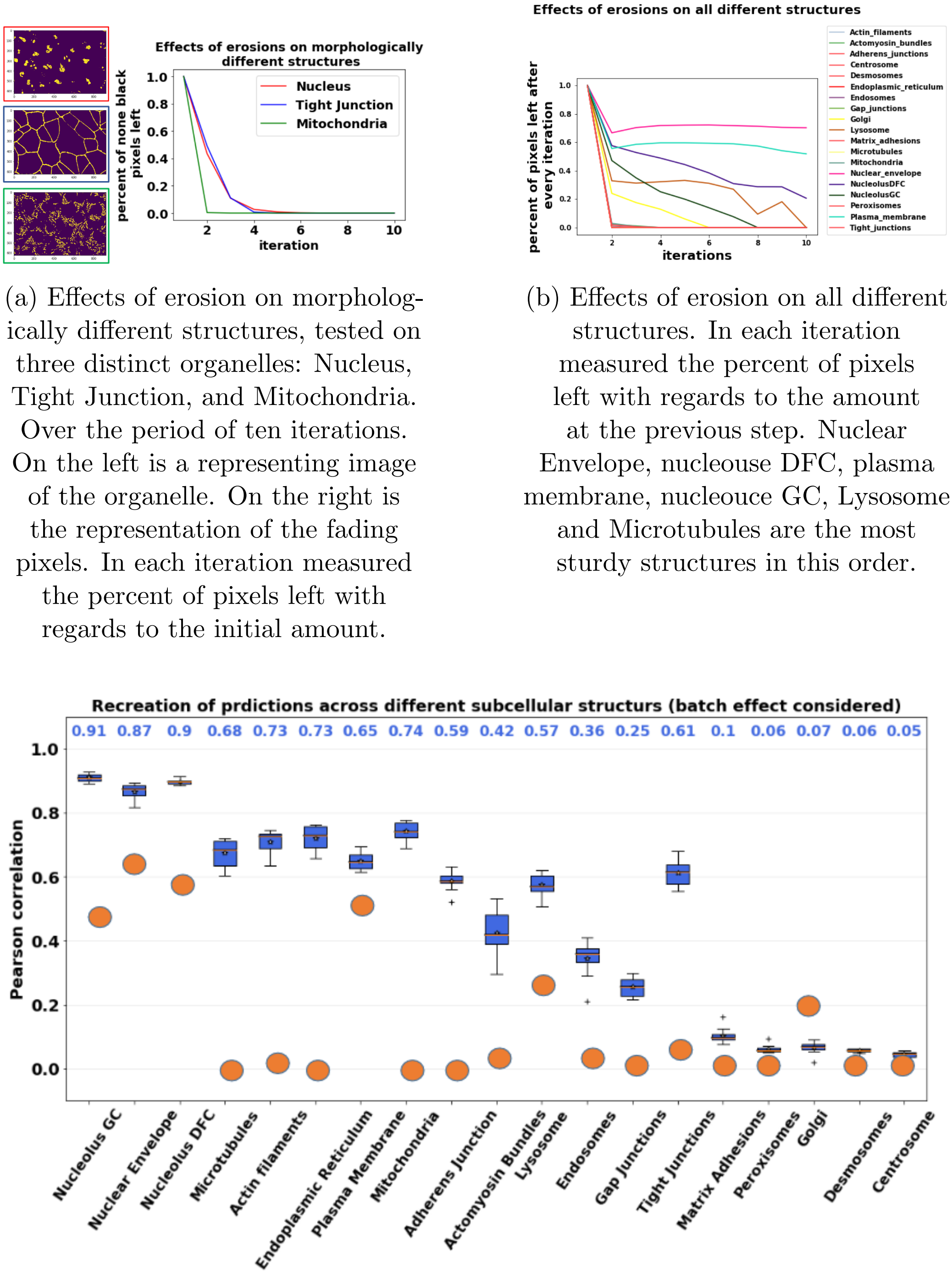
(a) Distributions of the image-wise Pearson correlation coefficient (r) between ground-truth (target) and predicted test images derived from the Tight junction subcellular structure model for data sets that were introduced to three different drugs compared with the control baseline (original); Blebbistatin, Berfeldin, and Staurosporine (Methods). Each target/predicted image pair in the test set is a point in the resultant r distribution; The number of images in the test set was 10, 14, 25, 5 correspondingly, and according to data availability. (All predictions were calculated using the original model only changing the test set).



(b) Examples of structure from control and perturbed samples. Top images are the ground truth, bottom are the predictions. From left to right: control, Blebbistatin, Brefeldin, and Staurosporine.

Figure 3.10: Evaluation of drug perturbation impact on predictions.

We had to challenge our growing intuition regarding the mapping between bright-field to the fluorescent label heaving to do great deal with spatial dependencies, and look into individual morphological properties of the different structures. Erosion is one of the fundamental operations in morphological image processing. We define *sturdy* structures structures that are more resistant to erosion are . We wanted to measure the correlation between the steadiness of a structure and its performance in in-silico labeling. We were not able to measure significant correlation (Fig. 3.11).



(a) Effects of erosion on morphologically different structures, tested on three distinct organelles: Nucleus, Tight Junction, and Mitochondria. Over the period of ten iterations. On the left is a representing image of the organelle. On the right is the representation of the fading pixels. In each iteration measured the percent of pixels left with regards to the initial amount.

(b) Effects of erosion on all different structures. In each iteration measured the percent of pixels left with regards to the amount at the previous step. Nuclear Envelope, nucleouse DFC, plasma membrane, nucleouce GC, Lysosome and Microtubules are the most sturdy structures in this order.

(c) Sturdiness, or resistance to erosion (orange dots) is plotted in context of the organelle's in-silico labeling performance. The sturdiness here is the results after one iteration of erosion.

Figure 3.11: Exploration of individual morphological properties of the different structures.

3.3 POC- Validating our approach on a well-known biological system composed of a small subset of structures

With the growing conviction that the mapping between bright-field to the fluorescent label of the structure has to do a great deal with spatial dependencies with its neighbors. We started thinking about how to measure it, and even pinpoint these neighbors. To make this POC more focused we chose a simple subset of three organelles that we know have a very significant role in cells' function and are spatially dependent. Based on the Central dogma of molecular biology, Nuclear envelope, Endoplasmic Reticulum (ER) and the Golgi apparatus, are the main players in cell protein production. The DNA contains the instructions for all proteins synthesized in the cell, contained in the nucleus bound by the nuclear envelope. Proteins created in the cell begin by being copied from the DNA, it is translated into a protein and folded in the ER [61], finally in the Golgi apparatus it is modified to its functional structure and packaged for transporting to its target destination [62].

In order to assert the best practices for our work, we executed a few supplementing analysis regarding our approach to Label-free. It was evident from the images that the third dimension contains some redundant slices that lack labeled information on the extremities. In the reprocessing, we experimented with only keeping the slices with the structure and reduced the image to fit into 40 z slices (Fig. 3.12). It's interesting to note that smaller structures did perform better in this format, while larger ones didn't or even showed decline. We hypothesize that the reason being the reduction of the reference area of the cell's perimeter for organelles that are large and dispersed across most of the cell's content. Our hypothesis is that this area is important for the model to predict the location of such structures. Over all the changes weren't of great significance thus we decided to continue working with whole z-axis range.

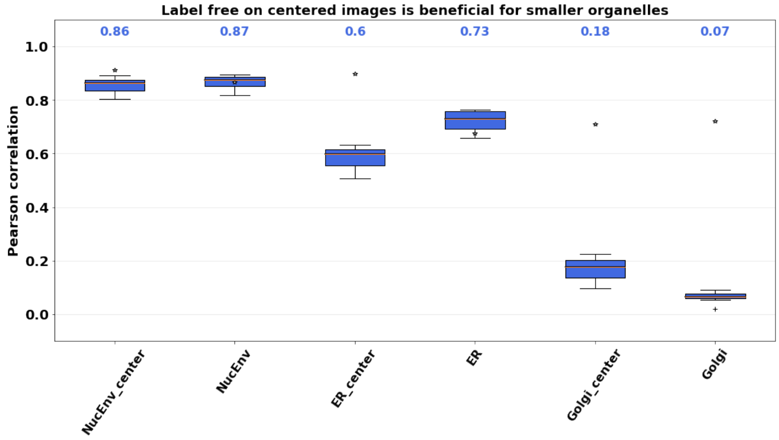


Figure 3.12: Assessment of image centering impact. Distributions of the image-wise Pearson correlation coefficient (r) between ground-truth (target) and predicted test images derived from the indicated sub-cellular structure models for the same data set in two variations; Whole image, 70 z-axis slices, and a cropped version with 40 z-axis slices containing the whole structure (indicated by _center). Each target/predicted image pair in the test set is a point in the resultant r distribution; The number of images in the test set was 10 for all distributions. Data for the recreation was constructed with batch effects consideration, excluding images from problematic imaging days. Comparing three distinct structures: Nuclear envelope (NucEnv), Endoplasmic reticulum (ER), and Golgi apparatus (Golgi).

Considering the training set size, we found that training models on 30 photos was sufficient and more training images did not improve our models (Fig. 3.13), in concurrence with [1].

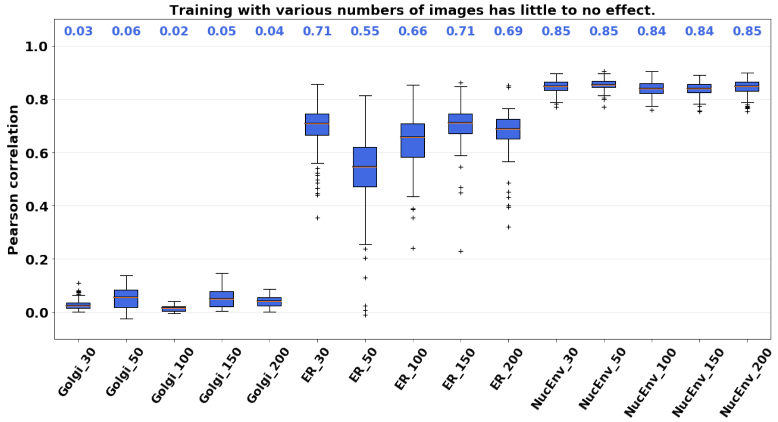
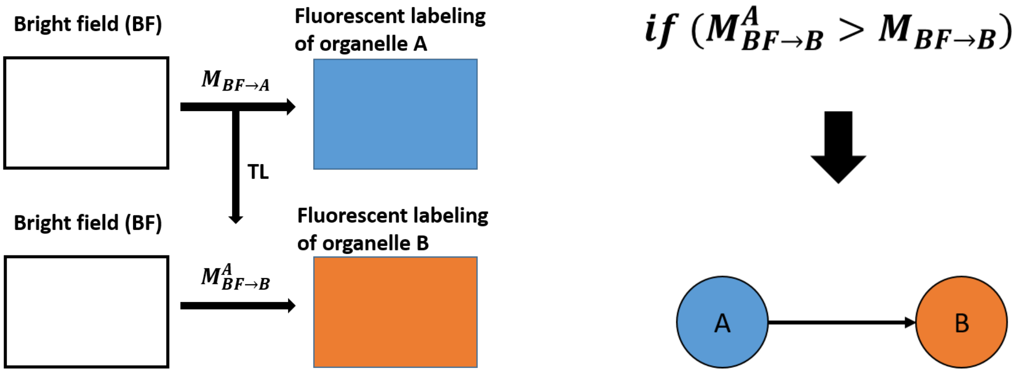


Figure 3.13: Evaluation of data volume's effects on training. Distributions of the image-wise Pearson correlation coefficient (r) between ground-truth (target) and predicted test images derived from the indicated sub-cellular structure models for the same data set in five variations; Baseline in 30 images, 50, 100, 150, and 200. Each target/predicted image pair in the test set is a point in the resultant r distribution; The number of images in the test set was 100 for all distributions. Comparing three distinct structures: Nuclear envelope (NucEnv), Endoplasmic reticulum (ER), and Golgi apparatus (Golgi).

Despite the fact that there are undoubtedly still many more questions about the methods we employ, we had to continue and reach the objective of our research. In order to bootstrap the in-silico method to our advantage the question we had to answer was: how much information that is learned by one model about an organelle contributes to the prediction of another? To answer this question we applied Transfer Learning; first, we trained a model to predict organelle A, then we trained a second model to predict organelle B initiating it with the weights of the first model (Fig. 3.14). If the second model is superior to a model trained directly to predict organelle B, then we can suggest that the spatial information of organelle A contributes to the prediction of organelle B, and that B is spatially-dependent on A (Methods).

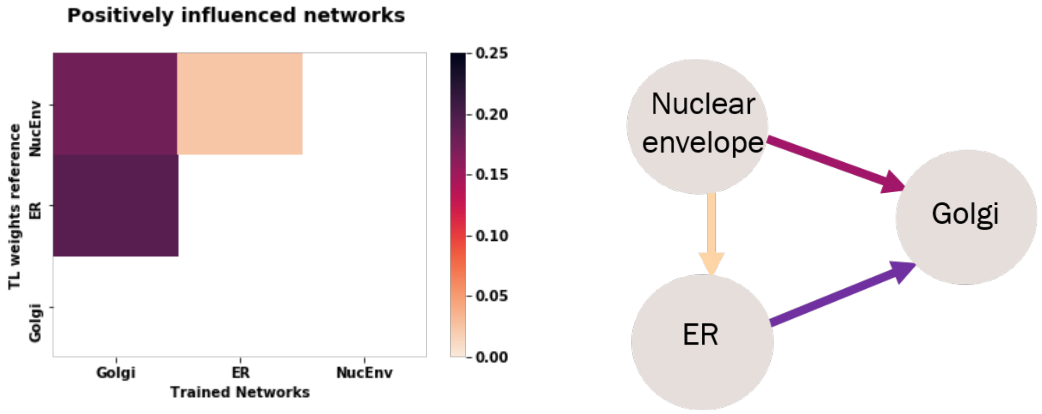

Bright field (BF)
Fluorescent labeling of organelle A
$M_{BF\to A}$
TL
Bright field (BF)
Fluorescent labeling of organelle B
$M^{A}_{BF\to B}$
$if\ (M^{A}_{BF\to B} > M_{BF\to B})$
A
B

(a) Top row represents the process of a label free model, trained predicting the fluorescent labeling of organelle A from a bright field unlabeled image, resulting in a model we annotate as follows: $\boldsymbol{M_{BF\to B}}$. The bottom raw represents the process in which the weights of $\boldsymbol{M_{BF\to B}}$ are loaded into a new model yet to be trained, which then is trained in a similar fashion to predict a different organelle B, in this instance the model has some previously learned insights which are applied in this new task, we annotate the resulting model as: $\boldsymbol{M^{A}_{BF\to B}}$.

(b) Analyzing the average person correlation of prediction of structure B with transfer learning from structure A and without, if the model with transfer learning showed significant improvement we conclude that the location of B is depending on the location of A, notice this dependency is not isomorphic.

Figure 3.14: Schematic of our approach for exploring and quantifying the spatial dependencies between organelles.

Using this architecture and applying it on the subset we chose, we trained models representing all the combinations of pairs of organelles, keeping in mind that the relations are not isomorphic. We expected to see that Nuclear Envelope would influence both Golgi and ER, ER would influence Golgi, and Golgi wont influence the other organelles based on the biological mechanisms mentioned above. The results confirmed our assumptions as shown in (Fig. 3.15). We measured the the percent of improvement in relation to the base model.



(a) Analysis of the results of our POC represented as a heat map, x-axis represents the structure on which we trained our models, y-axis is the structure of which weights were used to transfer. The map shows the percent of improvement each model achieved. The measurement of improvement: positive change in correlation (PCIC) is as fallows: $\boldsymbol{PCIC} = \frac{\boldsymbol{diff}}{1-\boldsymbol{M^{A}_{BF\to B}}}$ where - $\boldsymbol{diff} = \boldsymbol{M^{A}_{BF\to B}} - \boldsymbol{M^{A}_{BF\to B}}$

(b) A weighted directed graph, representing the spatial interactions. The arrows represent the direction of influence, for example the location of Golgi is influenced by Nuclear envelope. The color represents the magnitude of the impact.

Figure 3.15: The interplay between the nuclear envelope, endoplasmic reticulum (ER), and Golgi.

To validate our results we tested our networks with other images. We have chosen 100 images in a random manner. We predicted the location of the organelles with the basic models and compared the results with predictions of the same 100 images on our TL models(Fig. 3.16). We see that the same trends are consistent with the previous results. This time we also present negative influence that we observed (Explained in Methods).

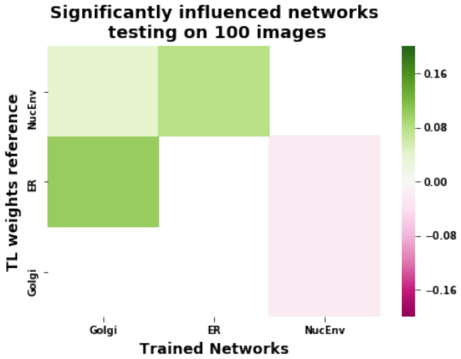
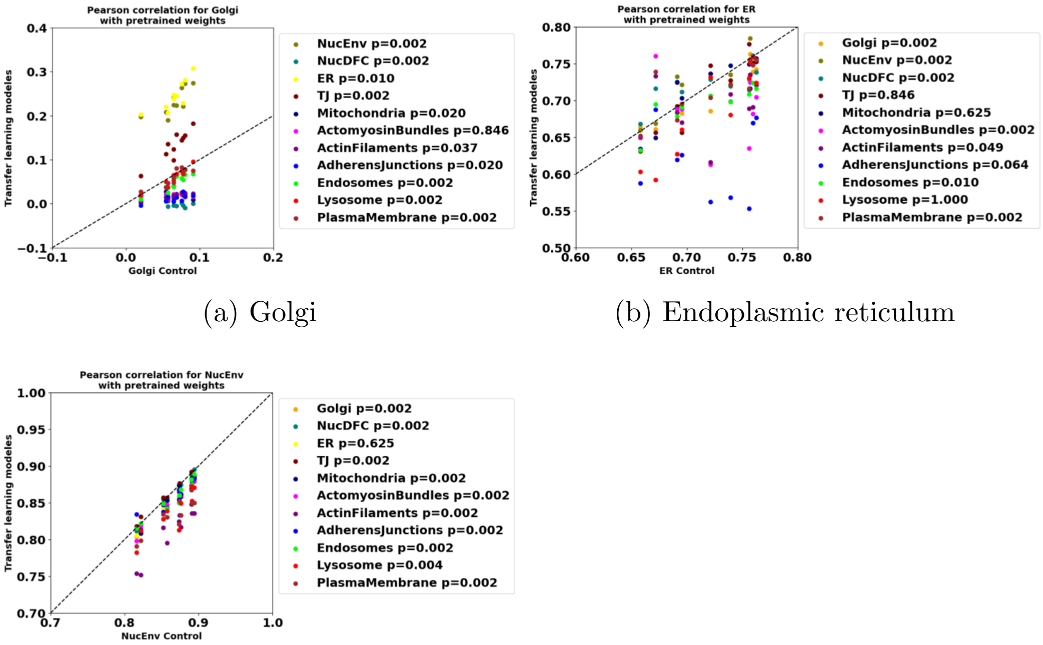


Figure 3.16: Analysis of the statistically significant results of our POC represented as a heat map was used to construct Fig. 24, x-axis represents the structure on which we trained our models, y-axis is the structure of which weights were used to transfer. The map shows the percent of improvement and negative effects each model achieved. This time each model was tested on a set of 100 images (negative TL is explained in Methods).

After validating our hypothesis with this small sub-set of structures, and before moving to a bigger scale of experiments it was important to examine the statistical significance of this changes. Using the Wilcoxon signed-rank test for each sample we evaluated the significance of the change by calculating the p value and considered significant only results that had p-score less than 0.05. At This point we wanted to incorporate more structures to our experiments, we extended our sub set to 12 structures, measuring the effects on the main three we did so far (Fig. 3.17). This analysis shows that the majority of the changes are significant. Another intriguing finding from this analysis was the observation that the prediction of nuclear envelopes is not improved by this approach; this is because the nuclear envelope is a very large and stereotypical structure with excellent results right away. Remembering the findings from Fig. 3.12, where we came to the conclusion that the information in the peripheral areas of the cell are necessary for this model, these results make sense. It may be the case that the additional information only serves to distract the model from the structure.



(a) Golgi

(b) Endoplasmic reticulum

(c) Nuclear envelope

Figure 3.17: Analysis of significance of our results. Each figure shows the comparison of the performance of the control model (x-axis) with models that were pre trained (y-axis) on organelles as indicated in the legend, color coded, the significance of the results is also indicated in the legend with the p value.

3.4 Main results

After the promising first results we were eager to see what other connections we are capable of finding. We increased the amount of organelles until we had a manageable yet versatile collection. We chose 14 organelles with distinct features and trained all the combinations of target and source TL models.

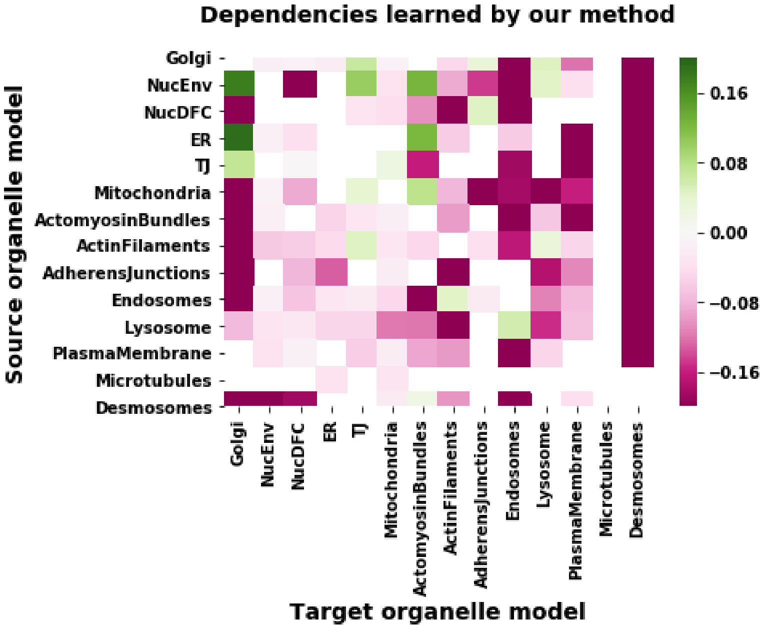


Figure 3.18: Analysis of the results represented as a heat map, x-axis represents the structure on which we trained our models, y-axis is the structure of which weights were used to transfer. The map shows the percent of improvement and decline each model achieved.

Looking at the results (Fig.3.18), we can see that there are mostly negative relations, which is a known phenomenon in TL, negative TL (as discussed in methods). It is usually explained as occurring when the divergence between the joint distributions of target and source data sets is very significant. Our hypothesis is that either this indicates organelles that are not spatially strongly correlated or the structural characteristics are so distinct that the model is not capable of making the transition. It is also possible that the source structure is not in correlation with the main areas the target structure is using and thus distracts it from his optimal solution. In (Fig. 3.19) is a directional and weighted graph of normalized relations between the organelles, it is a simplified and intuitive representation of organelle-organelle spatial dependencies. Such that nodes are representing the organelles and the vertices are weighted(represented with colour) and have directions showing who is the source or target. There were a few organelles that did not contribute significant positive connection, thus they are not participating in the graph. Now we have an explicit representation of the connections that were learned via our method. Looking into some biological possible rationalizations we found the following phenomena, explained here in a very simplified manner.

- Actin filaments → Tight Junction, is highlighted in the following papers, structural formation of actin is necessary for the functional formation of tight junction [63] [64] [65] [66] [67].

- Adherens Junction ← Golgi ↔ Tight Junction, Golgi-associated enzyme regulates the transport of transmembrane junction proteins through or from the Golgi, thereby controlling the integrity of endothelial cell–cell junctions [68] [67].

- ER → Golgi, "Newly synthesized proteins are transported from the endoplasmic reticulum via the Golgi to the trans-Golgi network.","Golgi stack is associated with a single endoplasmic reticulum (ER) exit site, forming a secretory unit." [66] [69]

- Nuc → Golgi, [69]

- Golgi→ lysosome, Fragmentation of golgi changes the location of lysosome [70].

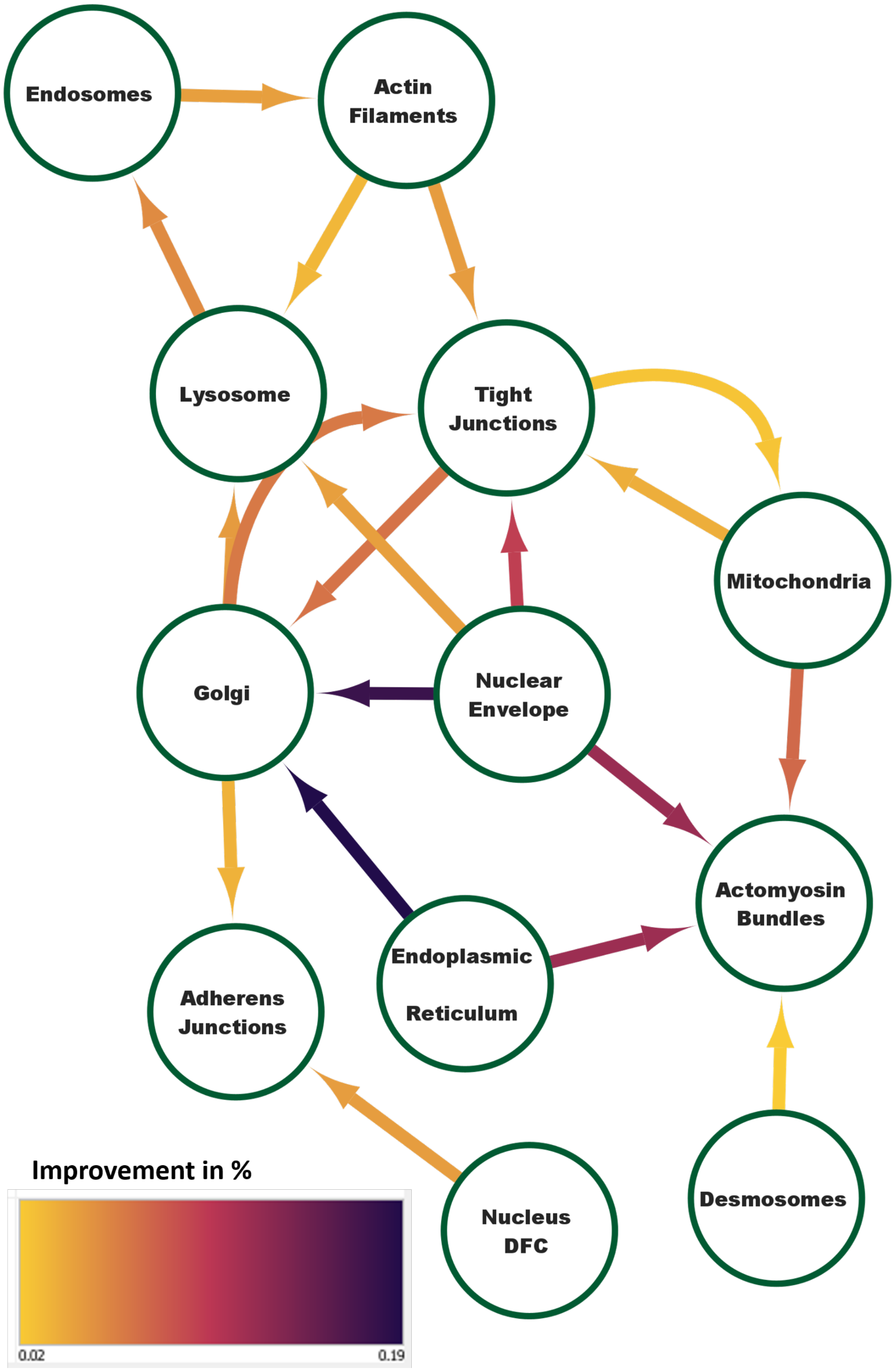
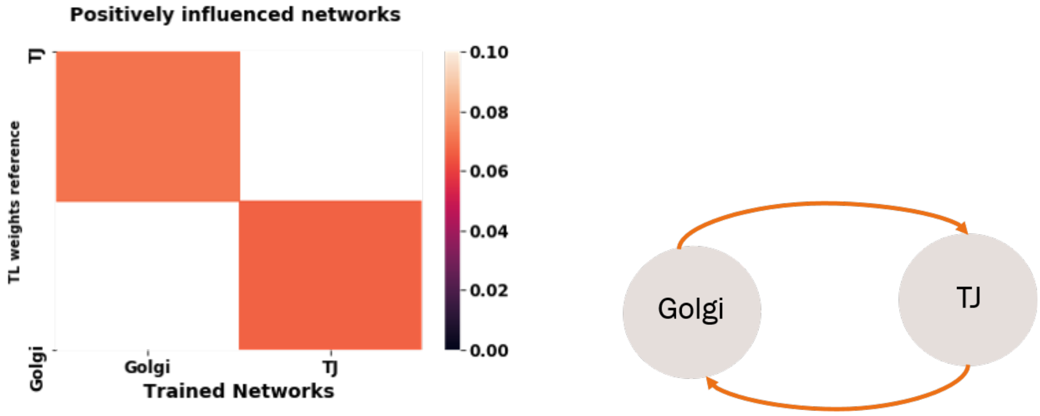


Figure 3.19: Organelle-Organelle interaction network. A weighted directed graph, representing the spatial interactions. The arrows represent the direction of influence, for example the location of Golgi is influenced by Nuclear envelope. The color represents the magnitude of the impact.

One of the intersting results we have come across was the Tight Junction and Golgi bidirectional spatial dependencies (Fig. 3.20), its a unique result because the relation between these two organelles is almost symmetric. We found evidence [67], [68], [66] that these relations are known and researched.

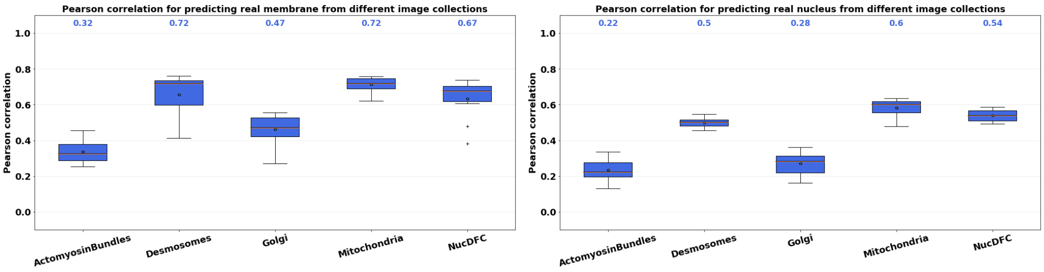


(a) Analysis of the results represented as a heat map, x-axis represents the structure on which we trained our models, y-axis is the structure of which weights were used to transfer. The map shows the percent of improvement each model achieved compared with a model without the use of transfer learning.

(b) Directed weighted graph representing the influence of TL on the performance of each structure Golgi on the left, and tight junction (TJ) on the right. The weight is represented with color gradients matching the map.

Figure 3.20: Tight junction and Golgi demonstrated bidirectional almost symmetric influence on one another.

Each image in our data set also contains two extra channels, representing the cell membrane and the nucleus, as was previously mentioned. Exploiting this underpinning new data collection provided a fantastic chance to advance our investigation. It was put to use in two separate ways. One extended our data set by using these additional channels as ”stand alone” data and applying them to train models of new organelles. Practically, we only required two models to be trained—one for each organelle—but comparing these models was interesting. We can observe from the results that the models were highly variable (Fig. 3.22), most likely because we neglected to examine the performance of these organelles when looking for batch effects.

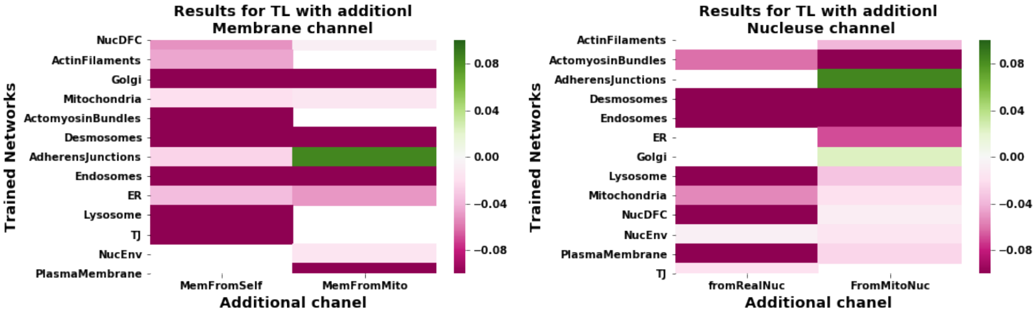


(a) Membrane

(b) Nucleus

Figure 3.21: Representative evaluation of the in-silico labeling for the additional channels each image contains. Distributions of the image-wise Pearson correlation coefficient (r) between ground-truth (target) and predicted test images derived from the indicated sub-cellular structure models from data sets of five organelles; Actomyosin bundles, Desmosomes, Golgi, Mitochondria and Nuclear DFC. Each target/predicted image pair in the test set is a point in the resultant r distribution.

Due to this, we selected the best model, which, interestingly, came from the mitochondria data set for both structures. And evaluated to see what impact these structures had on the rest of our data.



(a) Membrane

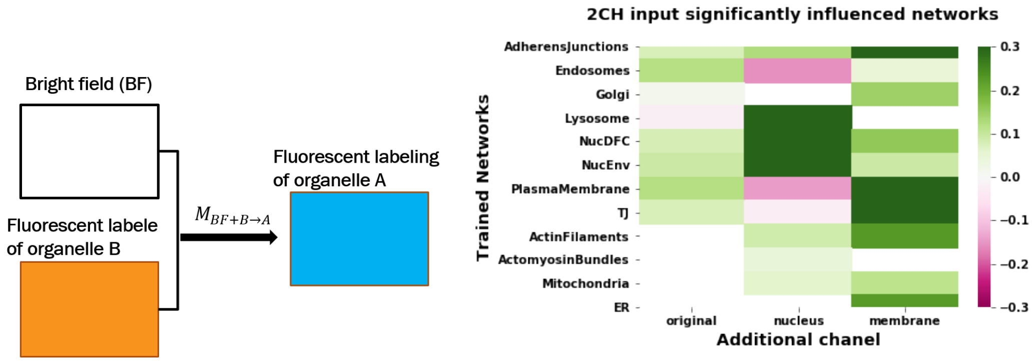
(b) Nucleus

Figure 3.22: Analysis of the results represented as a heat map, y-axis represents the structure on which we trained our models, x-axis is the structure of which weights were used to transfer(opposite from previous). The map shows the percent of improvement and decline each model achieved. in each map on the left are models that were pre-trained previously on the same BF images, on the right is the best performing model that was pre-trained on the mitochondria data set.

It was interesting to observe that the networks that were trained on the same BF images twice essentially, but with two different tasks resulted in a more significant decrease in performance. Also it was interesting to see that the Adherence junction benefited in both cases, while the Golgi only benefited from the Nucleus.

We see that additional channels are beneficial especially in similar organelles, for instance the nucleus improved results for the different nucleus structures while cell membrane improved the results of Plasma membrane.

Another method involved using additional real channels as an input for the model, resulting in two input channels, BF and additional organelle. Each time using a different channel.



(a) Schematic of the second approach. Left column represents the process of feeding a model with two input channels, bright field unlabeled image and fluorescent labeling of organelle B. The right column represents the output predicting the fluorescent labeling of organelle A. We annotate the resulting model as: $\boldsymbol{M_{BF+B\rightarrow A}}$.

(b) Analysis of the results represented as a heat map, y-axis represents the structure on which we trained our models, x-axis is the structure that was fed to the network as an additional channel. The map shows the percent of improvement and decline each model achieved. The left column was presented with two BF channels, the middle with the additional Nucleus and the right with the membrane.

Figure 3.23: Evaluation of introduction of two channels as an input for the model.

We see that additional channels are majorly beneficial especially in similar organelles, for instance the nucleus most significantly enhance results for the different nucleus structures while cell membrane impacts the results of Plasma membrane, Adherence junction and Tight junction the most. Its also interesting to see that even when we add another BF channel, meaning no new information was added this method managed to improve in-silico labelling.

When comparing the two methods, it becomes clear that they are very different from one another. The TL method works by utilizing spatial relations between organelles as evidenced by the improved performance on spatially-related organelles. On the other hand, we suspect that the additional channel method improves performance globally by providing the model with visual “hints” for each sample. That is, it provides the model with additional signal from the specific cells being predicted.

Chapter 4

Methods

4.1 Dataset description

Dataset description

Our data is Three-dimensional live cell microscopy. The 3D light microscopy data used to train and test the models consists of z-stacks of colonies of human embryonic kidney cells available at (see http:// www.allencell.org). Genome-edited (hiPSC) lines expressing a protein endogenously tagged with fluorescent protein (mEGFP) or red fluorescent protein (mTagRFP) that localizes to a particular subcellular structure. The EGFP-tagged proteins and their corresponding structures are: -tubulin (microtubules), -actin (actin filaments), desmoplakin (desmosomes), lamin B1 (nuclear envelope), fibrillarin (nucleoli), myosin IIB (actomyosin bundles), sec61B (endoplasmic reticulum), STGAL1 (Golgi apparatus), Tom20 (mitochondria), and ZO1 (tight junctions). The cell membrane was labeled by expression of mTagRFP.

All cell types were imaged for up to 2.5 h on a Zeiss spinning disk microscope with ZEN Blue 2.3 software and with a 100× /1.25-NA (numerical aperture) objective (Zeiss C-Apochromat × 100/1.25 W Corr), with up to four 16-bit data channels per image: transmitted light (either bright-field or DIC), cell membrane labeled with CellMask, DNA labeled with Hoechst, and EGFP-tagged cellular structure. Z-slice images were captured at a YX resolution of 624×924 px^2 with a pixel size of 0.108 μm px^{-1}, and 63×-objective z-slice images were captured at a YX resolution of $1{,}248{\times}1{,}848$ px^2 with a pixel scale of 0.086 μm px^{-1}. All z-stacks were composed of 50–75 z-slices with an inter-z-slice interval of 0.29 μm. Each Image contained 10–30 cells.

The data used to train and evaluate the models (Methods) based on 3D live-cell z-stacks, including train-test data splits. All multi-channel z-stacks were obtained from a database of images produced by the Allen Institute for Cell Science's microscopy pipeline (see http:// www.allencell.org).

Table 4.1: Data Dictionary: Listing for each of the 19 different structures available, the name of the protein that was used for in-silico labeling, the name of the structure, the amount of images available, and the imaging dates. The total number of available images is 11023.

Protein Names	**Structure**	**#Of Images**	**Imaging Dates**
Alpha-actinin-1, Beta-actin	Actin filaments	844	20180320, 20180323 ,20180326, 20180327, 20180330, 20180402, 20180403, 20180316, 20170919, 20170818, 20170828, 20170822

Continued on next page

Table 4.1 – *Continued from previous page*

Protein Names	**Structure**	**# Of Images**	**Imaging Dates**
Non-muscle myosin heavy chain IIB	Actomyosin bundles	426	20170925, 20171023, 20171010, 20170926, 20170922, 20171017, 20171020
Beta-catenin	Adherens junctions	590	20180515, 20180521, 20180514, 20180522, 20180525
Centrin-2	Centrosome	492	20171222, 20171219, 20180119, 20180122, 20171212, 20180123, 20171208, 20180112
Desmoplakin	Desmosomes	712	20170725, 20170718, 20170717, 20170726, 20170722, 20170808, 20170807, 20170724, 20170814, 20170728, 20170719, 20170721
Sec61 beta	Endoplasmic reticulum	347	20170811, 20170818, 20170807, 20170815, 20170728, 20170821, 20170804
Ras-related protein Rab-5A	Endosomes	374	20181001, 20181005, 20181009, 20181012
Connexin-43	Gap junctions	461	20180306, 20180309, 20180319, 20180320, 20180312, 20180313, 20180316
Sialyltransferase 1	Golgi	449	20170829, 20170918, 20170901, 20170912, 20170905, 20170825, 20170915

Continued on next page

Table 4.1 – *Continued from previous page*

Protein Names	**Structure**	**# Of Images**	**Imaging Dates**
LAMP-1	Lysosome	603	20171023, 20171106, 20171110, 20171113, 20171103, 20171030
Paxillin	Matrix adhesions	435	20180409, 20180416, 20180417
Alpha-tubulin	Microtubules	570	20170317, 20170331, 20170307, 20170310, 20170328, 20170306, 20170321, 20170227, 20170327, 20170301, 20170324, 20170320, 20170314, 20170303
Tom20	Mitochondria	1078	20170614, 20170626, 20170628, 20170612, 20170627, 20170623, 20170613, 20170712, 20170620, 20170609, 20170705, 20170619, 20170621, 20170616, 20170630
Lamin B1	Nuclear envelope	1017	20170519, 20170607, 20170530, 20170522, 20170523, 20170526, 20170606, 20170509, 20170515, 20170512, 20170531, 20170524, 20170508
Fibrillarin	Nucleolus (DFC)	608	20171027, 20171103, 20171031, 20171114, 20171117, 20171120, 20171110, 20171020
Nucleophosmin	Nucleolus (GC)	811	20180430, 20180501, 20180504, 20180507, 20180508, 20180511

Continued on next page

Table 4.1 – *Continued from previous page*

Protein Names	**Structure**	**# Of Images**	**Imaging Dates**
Peroxisomal membrane protein PMP34	Peroxisomes	372	20180928, 20181002, 20181008, 20181029, 20181030, 20181016, 20181116
CAAX domain of K-Ras	Plasma membrane	498	20180312, 20180313, 20180316, 20180319, 20180323, 20180326, 20180327
Tight junction ZO-1	Tight junctions	336	20171002, 20171016, 20171024, 20171009, 20170926
TOTAL		11023	

The Allen institute generated a small data-set in order to identify signature profiles of cellular organization for a range of well-characterized agonists and antagonists commonly used to perturb[1] specific cellular processes or pathways.

To do this, they used the cells from the Allen Cell Collection (7 different structures) introducing them to 6 different drugs. Since Tight Junction (TJ) was the only structure where we could visually see the impact of the drugs, we mainly tested three different drug types on it.

[1] "Perturbation is an alteration of the function of a biological system by external or internal means such as environmental stimuli, drug inhibition, and gene knockdown" [71].

- Blebbistatin- (S)-nitro blebbistatin (SNB), inhibits the non-muscle myosin II ATPase. (S)-nitro blebbistatin also produced a reorganization of the tight-junction protein ZO-1, with fragmentation of the cortical ring near the apical surface and redistribution throughout the entire volume of the cell.

- Brefeldin A- Inhibits protein transport from the endoplasmic reticulum to the golgi complex indirectly by preventing association of COP-I coat to the Golgi membrane [72]. It fragmented the Golgi apparatus, (which we know is connected to the TJ functionality) into multiple foci dispersed throughout the cell volume within 30 min of treatment. This fragmentation is similar to that normally observed during mitosis. None of the other structures studied showed major changes.

- Staurosporine- The main biological activity of staurosporine is the inhibition of protein kinases through the prevention of ATP binding to the kinase. This is achieved through the stronger affinity of staurosporine to the ATP-binding site on the kinase. Staurosporine is a prototypical ATP-competitive kinase inhibitor in that it binds to many kinases with high affinity, though with little selectivity [73]. Staurosporine induced re-localization of ZO-1 from a tightly localized band near the apical surface to a broader distribution throughout the cell surface.

To identify the center of each image we measured the standard deviation of pixel intensities across the z-planes. $s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$

High SD in a specific z-slice implies high variation in pixel intensities that can be interpreted as the presence of an organelle. The slice with the highest standard deviation in the nuclear channel was correlated with the center of the cell and this observation was used to localize the cell's center. Using the same approach we identified that all other organelles were located within a radius of 5.8 μm, thus we kept the 40 slices around the cell's center and excluded the rest (30% of data).

4.2 Day-to-day variability (batch effects)

Analysis of batch effects which is the issue of day to day variability was necessary. The data was collected over a period of many months. Different days of imaging cause high variation in the conditions, such as lighting, adjustment of the equipment and of course the quality of the fluorescent labels and the behavior of the cells on the plate as well as many others. To make sure the results are not affected by these factors we made an analysis of the different imaging days with the purpose of pinpointing problematic days that may have an impact on the performance of the models that is due to poor imaging data. We excluded days with measurable batch effects defined as follows, our dataset includes 4 days per organelle for each organelle and each day we partitioned the data to train set of 30 images and test set of 10 images in two ways, defining two corresponding models:

- M1 - Train set from three days, test set from the remaining day. This ensures that there is no data-contamination in terms of the daily batch effects.

- M2 - Train set from all days, the model enjoys information regarding the day at test.

We evaluate M1(I) versus M2(I) where I is an image from the day of the test. Days with non-parametric Wilcoxon rank-sign test p-value ¡ 0.05 (where M2 surpasses M1) were excluded because the information about the day was helpful during training, and it implies that this day was distinct from the others.

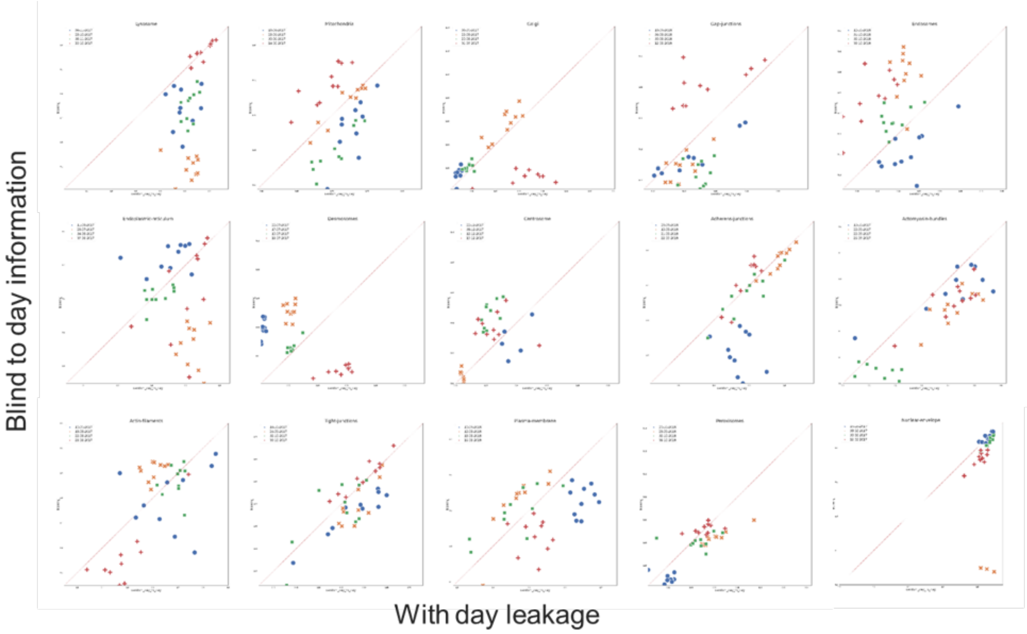


Figure 4.1: Assessment of batch effects. Each panel shows batch effects for a different organelle. Each color represents a different imaging day, and each data point represents an image. X-axis is the cross correlation coefficient when the training set includes images from the same imaging day (not the images at test, but with data leakage). Y-axis is the matched cross correlation coefficient when the training set was blind to the imaging day at test (no contamination). Batch effects were characterized by data points below the reference Y = X line. A subset of days suffering from batch effects, days that were affected by batch effects (e.g., orange markers at the most bottom-right panel) were excluded from further analysis.

4.3 Models

Model architecture description and training: Due to its proven effectiveness in image segmentation and tracking tasks, we used a Convolutional Neural Network (CNN) based on the U-Net architecture (Fig. 4.2). In general, because they are image-translation invariant, learn complex nonlinear relationships across multiple spatial areas, do not require the development of data-specific feature extraction pipelines, and are simple to implement and train, CNNs are especially effective for tasks involving images (such as classification, segmentation, and image-to-image regression).

CNNs have been shown to outperform other state-of-the-art models in basic image recognition [74] and have been used in biomedical imaging for a wide range of tasks including image classification, object segmentation [75] and estimation of image transformations [76]. This U-Net variant consists of layers that perform one of three convolution types, followed by a batch normalization and rectified linear unit (ReLU) operation. The convolutions are either 3-pixel convolutions with a stride of 1 pixel on zero-padded input (such that the input and output of that layer are the same spatial area), 2-pixel convolutions with a stride of 2 pixels (to halve the spatial area of the output), or 2-pixel transposed convolutions with a stride of 2 pixels (to double the spatial area of the output). There are no normalization or ReLU operations on the last layer of the network. The number of output channels per layer is shown in (Fig. 16). Our 3D models use 3D convolutions.

Owing to memory constraints associated with graphics processing unit computing, we trained the model on batches of 3D patches ($64 \times 64 \times 32\ px^3$, YXZ), which were randomly subsampled uniformly both across all training images and spatially within an image. The training procedure took place in a typical forward–backward fashion, with model parameters updated via stochastic gradient descent (backpropagation) to minimize the MSE between output and target images. The models were trained using the Adam optimizer [77] with a learning rate of 0.001 and with beta values of 0.5 and 0.999 for 50,000 mini- batch iterations. We used a batch size of 16 for the models due to hardware restrictions, the memory on some of our GPUs was a little low. Running on one of the following machines: titan gtx, rtx2080, gtx1080, rtx3090, rtx3090, each model completed training in approximately 16 to 26 h.

For prediction tasks, we minimally crop the input image such that its size in any dimension is a multiple of 16, to accommodate the multi-scale aspect of the CNN architecture. Prediction Three-dimensional light microscopy model results analysis and validation. For our 3D light microscopy applications, model accuracy was quantified by the Pearson correlation coefficient between the pixel intensities of the model's output, y, and independent ground- truth test images, x : $r = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2\sum(y-\overline{y})^2}}$.

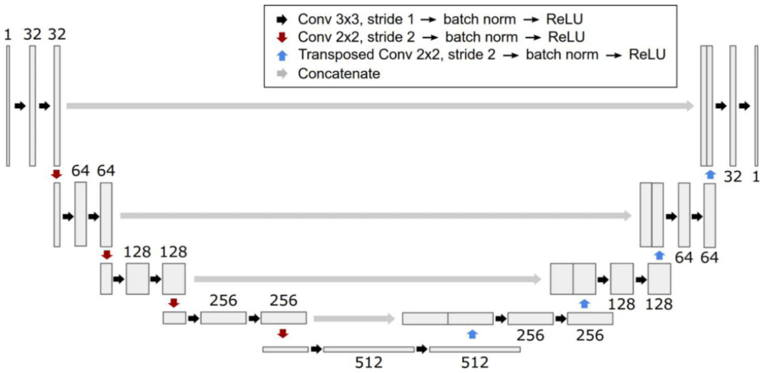

Conv 3x3, stride 1 → batch norm → ReLU
Conv 2x2, stride 2 → batch norm → ReLU
Transposed Conv 2x2, stride 2 → batch norm → ReLU
Concatenate
1 32 32
64 64
128 128
256 256
512
512
256 256
128 128
64 64
32 1

Figure 4.2: Schema of the model's composition. There are no batch normalization or ReLU layers on the last layer of the network, and the number of output channels per layer is shown above the box of each layer. Figure adapted from Lecture Notes in Computer Science 234–241 (2015).

4.4 Transfer learning

Transfer learning (TL) is a method used to transfer knowledge acquired from one task to resolve another[16]. For instance, before learning to fly a real airplane, pilots use a variety of simulations, such as video games, to gain prior skills, knowledge, and experience.

In the context of machine learning, By transferring the knowledge found in various but related source domains, TL seeks to increase the performance of target learners on target domains [16], [39], [40], [41]. This is very useful when data in the target domain is scarce and it is not possible to train a good model with it, so TL gives the model a head start. Sometimes TL is also used to direct the model in a specific direction when the target task is not easily deducible from the target data.

The idea of transfer learning was first discussed in the context of education in the last century, it was first considered in a computational context in 1995, and went through conceptual refinement in 2005 to the manner that we comprehend it today [44].

Inductive transfer learning, transductive transfer learning, and unsupervised transfer learning are the three subsettings most frequently used to classify transfer learning, depending on the circumstances between the source and target domains and tasks. [44].

Regardless of whether the source and target domains are the same or not, the target task in the context of inductive transfer learning differs from the source task. To create an objective prediction model for usage in the target domain in this situation, some labeled data from the target domain are necessary. Additionally, based on various circumstances involving labeled and unlabeled data in the source domain, we can further categorize the inductive transfer learning setting into two cases:

- a. There is a sizable amount of labeled data in the source domain. In this case, the multitask learning setting and inductive transfer learning are analogous. Multitask learning attempts to learn the source and target tasks simultaneously, whereas inductive transfer learning simply aspires to achieve high performance in the target task by transferring knowledge from the source task.
- b. No labeled data in the source domain are available. In this case, the inductive transfer learning setting is similar to the self-taught learning setting, which was first proposed by Raina et al.[78]. In the self-

taught learning setting, the label spaces between the source and target domains may be different, which implies the side information of the source domain cannot be used directly. Thus, it's similar to the inductive transfer learning setting where the labeled data in the source domain are unavailable.

In the transductive transfer learning setting, the source and target tasks are the same, while the source and target domains are different. In this situation, no labeled data in the target domain are available while a lot of labeled data in the source domain are available. In addition, according to different situations between the source and target domains, we can further categorize the transductive transfer learning setting into two cases.

- a. The feature spaces between the source and target domains are different.
- b. The feature spaces between domains are the same, but the marginal probability distributions of the input data are different. It is related to domain adaptation for knowledge transfer in text classification, and sample selection bias.

Finally, in the unsupervised transfer learning setting, similar to the inductive transfer learning setting, the target task is different from but related to the source task. However, the unsupervised transfer learning focuses on solving unsupervised learning tasks in the target domain, such as clustering, dimensionality reduction, and density estimation [79], [80]. In this case, there is no labeled data available in both source and target domains in training.

Approaches to transfer learning in the above three different settings can be summarized into four cases based on "What to transfer".

- The first context can be referred to as instance-based transfer learning (or instance transfer) approach [45], [46], which assumes that certain parts of the data in the source domain can be reused for learning in the target domain by reweighting. Instance reweighting and importance sampling are two major techniques in this context.

- A second case can be referred to as feature-representation-transfer approach [78], [81]. The intuitive idea behind this case is to learn a "good" feature representation for the target domain. In this case, the knowledge used to transfer across domains is encoded into the learned feature representation. With the new feature representation, the performance of the target task is expected to improve significantly.

- A third case can be referred to as parameter-transfer approach [82], [83], which assumes that the source tasks and the target tasks share some parameters or prior distributions of the hyperparameters of the models. The transferred knowledge is encoded into the shared parameters or priors. Thus, by discovering the shared parameters or priors, knowledge can be transferred across tasks.

- Finally, the last case can be referred to as the relational knowledge-transfer problem [84], which deals with transfer learning for relational domains. The basic assumption behind this context is that some relationship among the data in the source and target domains is similar.

Thus, the knowledge to be transferred is the relationship among the data.

Transferred knowledge does not always bring a positive impact on new tasks. If there is little in common between domains, knowledge transfer could be unsuccessful [56]. For example, learning to ride a bicycle cannot help us learn to play the piano faster. Besides, the similarities between domains do not always facilitate learning, because sometimes the similarities may be misleading. For example, although Spanish and French have a close relationship with each other and both belong to the Romance group of languages, people who learn Spanish may experience difficulties in learning French, such as using the wrong vocabulary or conjugation. This occurs because previous successful experience in Spanish can interfere with learning the word formation, usage, pronunciation, conjugation, and so on, in French. In the field of psychology, the phenomenon that previous experience has a negative effect on learning new tasks is called negative transfer [17]. Similarly, in the transfer learning area, if the target learner is negatively affected by the transferred knowledge, the phenomenon is also termed as negative transfer [44], [57]. Whether negative transfer will occur may depend on several factors, such as the relevance between the source and the target domains and the learner's capacity of finding the transferable and beneficial part of the knowledge across domains. Despite being obvious, this explanation obscures several important aspects of negative transfer, including the next three points.:

- 1. Negative transfer should be defined with regard to the algorithm used. It is important to consider what the negative impact is compared with. For example, it will be misleading to only compare with the best possible algorithm only using the target data, because the increase in risk (the expected value of specific task loss) may not come from using the source-domain data, but the difference in algorithms.

- 2. Divergence between the joint distributions is the main cause of negative transfer. As negative transfer is algorithm specific, it is expected to doubt the existence of a transfer learning algorithm that can always improve the expected risk compared to its target-domain only baseline. It turned out this depends on the divergence between distribution in the source and the target domain [85]. Then, an ideal transfer would figure out and take advantage of the similar part, leading to improved performance. However, if an algorithm fails to discard the divergent part and instead rely on it, one can expect negative transfer to happen.

- 3. Another important factor of negative transfer is the size of the labeled target data, which can have a mixed effect. On one hand, for the same algorithm and distribution divergence, negative transfer condition depends on how well the algorithm can do using target data alone. Where there is no labeled target data, only using unlabeled target data would result in a weak random model and thus negative transfer conditions are unlikely to be satisfied. When labeled target data is available, a better target-only baseline can be obtained using semi-supervised learning methods and so negative transfer is relatively more likely to occur. At the other end of the spectrum, if there is an abundance of labeled target data, then transferring from an even

> slightly different source domain could hurt the generalization. Thus, this shows that negative transfer is relative.

In the recent years researchers started incorporating this tool in their research in the fields of microscopy [86], biology, genetics etc. to benefit tasks such as protein folding, and labeling cancer cells (add references). In this research we present an innovative use of TL to learn about organelle - organelle spatial dependencies. Our objective was to measure the interactions between different organelles. We believe that the information about the structure, organization, and patterns, is stored in the networks we have trained in advance. Thus, to create a measurable analysis, we wanted to pass this knowledge from one network to another and study the impact it creates, our approach is under the instance-based transfer learning case. The method we came up with was applying Transfer Learning on our initial networks. We load the weights generated by network A (that predicts organelle A) as an initialization for a new network that we are going to train to predict organelle B. The results of the new network $\boldsymbol{M^{A}_{BF\to B}}$ are compared to the results of the original network $\boldsymbol{M_{BF\to B}}$ that predicts organelle B in order to determine if organelle B is spatially dependent on organelle A (Fig. 3.14). Using the Wilcoxon signed-rank test, we make sure that we only compare the findings that are statistically significant. We measure the positive change in correlation (significance of our prediction). The PCIC (positive change in correlation) as follows: $\boldsymbol{PCIC} = \frac{\boldsymbol{diff}}{\boldsymbol{1 - M^{A}_{BF\to B}}}$ where - $\boldsymbol{diff} = \boldsymbol{M^{A}_{BF\to B}} - \boldsymbol{M^{A}_{BF\to B}}$ We also measured the negative change (NCIC) in a similar fashion: $\boldsymbol{NCIC} = \frac{\boldsymbol{diff}}{\boldsymbol{M^{A}_{BF\to B}}}$.

I would like to express two important disclaimers:

- The performance of label free models can be enhanced with the help of TL as shown in [34][87]. It is critical to recognize that this work does not center on accomplishing it. By analyzing how various networks interact with one another, we use TL to learn about the spatial dependencies between organelles.

- Intentionally, we utilized a really rudimentary TL method. Although we might have used a variety of methodologies and procedures to improve the performance for each structure depending on its particular characteristics, our objective was to provide results that were simple to compare.

Chapter 5

Discussion and Conclusions

To provide a comprehensive methodology for predicting organelle-organelle interactions, as well as measuring the magnitude of the impact structures have on each others' location. We have investigated a unique dataset of 3D microscopy imaging, investigated the limitations and weak points of the in-silico labeling method as well as methods to enhance it, and applied an innovative approach for using transfer learning. The main outcome of this research is a prototype of a data-driven organelle interactome atlas with computationally predicted organelle-organelle interactions, represented as a directional weighted graph. In this graph one can see the impact each organelle has on its neighbors and the magnitude of this impact.

Our method is a powerful tool for giving Biologists an intuition about the inner workings of the cell. It can be used to reveal new intricate connections between organelles in the cell, revealing unknown components in systems and pathways in the cells' function. By revealing these new insights that were not available with currently existing tools researchers will be able to develop new specific and impactful medical treatments such as drugs, genetic modifications and even diet alterations that can save patients lives. Following this logic, and also based on our results with drug perturbations (Fig. 3.10) our method can actualize as a tool for measuring the impact of different treatments and procedures on spatial dependencies in multi organelle systems.

The fresh viewpoint on TL that this research offers is another important outcome; in this case, we used it to discover spatial relationships rather than to enhance learning. To access the features that the model has learned, in other words. It is important to notice that our results represent the features that our method was able to learn given the data that was available to us and under the time constraints of a masters degree . Here I would like to list some limitations, constraints and possible pitfalls our research is prone to suffer from: variance in cell types, as well as in the imaging resolution, leakiness of in-silico predictions, sensitivity of our method.

Areas of challenge: First I would like to address the issue of variance in cell types, there are many different types of cells in different organs of our body. These differences are important for our body's healthy operation, in addition there is also the difference of these cells among different ethnicities and cultures. These differences will oftentimes be expressed in different shapes and compositions of their organelles, but more importantly in order to execute different functionalities the interactions between different organelles will differ, obviously leading to different outcomes in our method. Addressing this will require extensive work of acquiring datasets of imaging of cells with different characteristics and analyzing them. Staying on the topic of variance of our data, it is also important to notice that different imaging data has different imaging resources, variety in quality, resolution and other technical conditions. Even in our own data set we had batch effects which we addressed (in chapter. 4.2) by preprocessing and eliminating problematic days. It is interesting to note that some organelle interactions occur at multiple spatial scales [88]. This phenomena might be captured if data with variable range scale will be available, then the atlas would have to have different layers to reflect this.

Variance in the resolution of structures is another issue, some imaging resolutions are able to capture bigger structures like the nucleus, but Golgi and Desmosomes lack important structural resolution. To solve this issue much higher resolution images would be required, the down side of this solution will be the significant increase in computation time and resources required. We managed this problem by measuring the impacts relatively in a manner that takes into account the performance of the in-silico models for each structure (Fig. 3.22(a)). Taking a step back it is also important to notice that the method of in-silico that we rely on is also not fault proof, thus validations on data from multiple sources is required in order to validate the results. To complete this atlas data for more organelles is required, which is not biologically yet possible to annotate ground truth images of all organelles in a complete manner.

Negative transfer learning (NTL) is another topic that requires further investigation in this context, as the main reasons for this phenomena are explained it is important to further investigate the cause in our case, we have a few suspicions that we didn't have time to validate: First is very straight forward, models that have really poor result to start with, will only confuse the model we are trying to train, thus it will cause NTL. Structures might have negative relations with each other, canceling each other might be reflected as NTL. Imaging differences between the source and target datasets like high variance in the resolution, quality, etc. might also result in NTL.

Together, the computational strategies addressed in this study demonstrate how image-based information can be measured, incorporated, and explored in the context of data-driven cell biology. Future research must solve the mentioned crucial issues in order for this potential to be fully fulfilled. In the long term, we envision a comprehensive cell atlas, which can be mined for patterns and relationships across a wide range of experiments and modalities.

Bibliography

[1] Ounkomol, Chawin, Seshamani, Sharmishtaa, Maleckar, Mary M, Collman, Forrest, and Johnson, Gregory R. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nature methods*, 15(11):917–920, 2018.

[2] Friedman, Jonathan R, Lackner, Laura L, West, Matthew, DiBenedetto, Jared R, Nunnari, Jodi, and Voeltz, Gia K. Er tubules mark sites of mitochondrial division. *Science*, 334(6054):358–362, 2011.

[3] Murley, Andrew and Nunnari, Jodi. The emerging network of mitochondria-organelle contacts. *Molecular cell*, 61(5):648–653, 2016.

[4] Wu, Haoxi, Carvalho, Pedro, and Voeltz, Gia K. Here, there, and everywhere: The importance of er membrane contact sites. *Science*, 361 (6401):eaan5835, 2018.

[5] Valm, Alex M, Cohen, Sarah, Legant, Wesley R, Melunis, Justin, Hershberg, Uri, Wait, Eric, Cohen, Andrew R, Davidson, Michael W, Betzig, Eric, and Lippincott-Schwartz, Jennifer. Applying systems-level spec-

tral imaging and analysis to reveal the organelle interactome. *Nature*, 546(7656):162–167, 2017.

[6] Saheki, Yasunori and De Camilli, Pietro. Endoplasmic reticulum–plasma membrane contact sites. *Annual review of biochemistry*, 86: 659–684, 2017.

[7] Shai, Nadav, Yifrach, Eden, van Roermund, Carlo WT, Cohen, Nir, Bibi, Chen, IJlst, Lodewijk, Cavellini, Laetitia, Meurisse, Julie, Schuster, Ramona, Zada, Lior, et al. Systematic mapping of contact sites reveals tethers and a function for the peroxisome-mitochondria contact. *Nature communications*, 9(1):1–13, 2018.

[8] Cohen, Sarah, Valm, Alex M, and Lippincott-Schwartz, Jennifer. Interacting organelles. *Current opinion in cell biology*, 53:84–91, 2018.

[9] Scorrano, Luca, De Matteis, Maria Antonietta, Emr, Scott, Giordano, Francesca, Hajnóczky, György, Kornmann, Benoît, Lackner, Laura L, Levine, Tim P, Pellegrini, Luca, Reinisch, Karin, et al. Coming together to define membrane contact sites. *Nature communications*, 10(1):1–11, 2019.

[10] Kroemer, Guido and Jäättelä, Marja. Lysosomes and autophagy in cell death control. *Nature Reviews Cancer*, 5(11):886–897, 2005.

[11] Chang, Diane TW and Reynolds, Ian J. Mitochondrial trafficking and morphology in healthy and injured neurons. *Progress in neurobiology*, 80(5):241–268, 2006.

[12] Carlton, Jeremy G, Jones, Hannah, and Eggert, Ulrike S. Membrane and organelle dynamics during cell division. *Nature Reviews Molecular Cell Biology*, 21(3):151–166, 2020.

[13] Garini, Yuval, Young, Ian T, and McNamara, George. Spectral imaging: principles and applications. *Cytometry Part A: The Journal of the International Society for Analytical Cytology*, 69(8):735–747, 2006.

[14] Christiansen, Eric M, Yang, Samuel J, Ando, D Michael, Javaherian, Ashkan, Skibinski, Gaia, Lipnick, Scott, Mount, Elliot, O'neil, Alison, Shah, Kevan, Lee, Alicia K, et al. In silico labeling: predicting fluorescent labels in unlabeled images. *Cell*, 173(3):792–803, 2018.

[15] Ronneberger, Olaf, Fischer, Philipp, and Brox, Thomas. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[16] Torrey, Lisa and Shavlik, Jude. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[17] Perkins, David N, Salomon, Gavriel, et al. Transfer of learning. *International encyclopedia of education*, 2:6452–6457, 1992.

[18] Sevakula, Rahul K, Singh, Vikas, Verma, Nishchal K, Kumar, Chandan, and Cui, Yan. Transfer learning for molecular cancer classification using deep neural networks. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(6):2089–2100, 2018.

[19] Wang, Sheng, Li, Zhen, Yu, Yizhou, and Xu, Jinbo. Folding membrane proteins by deep transfer learning. *Cell systems*, 5(3):202–211, 2017.

[20] Matassoni, Marco, Gretter, Roberto, Falavigna, Daniele, and Giuliani, Diego. Non-native children speech recognition through transfer learning. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6229–6233. IEEE, 2018.

[21] Sun, Han, Chen, Yu, Aved, Alexander, and Blasch, Erik. Collaborative multi-object tracking as an edge service using transfer learning. In *2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 1112–1119. IEEE, 2020.

[22] Phillips, Melissa J and Voeltz, Gia K. Structure and function of er membrane contact sites with other organelles. *Nature reviews Molecular cell biology*, 17(2):69–82, 2016.

[23] Rowland, Ashley A, Chitwood, Patrick J, Phillips, Melissa J, and Voeltz, Gia K. Er contact sites define the position and timing of endosome fission. *Cell*, 159(5):1027–1041, 2014.

[24] Neher, Richard and Neher, Erwin. Optimizing imaging parameters for the separation of multiple labels in a fluorescence image. *Journal of microscopy*, 213(1):46–62, 2004.

[25] Heinrich, Larissa, Bennett, Davis, Ackerman, David, Park, Woohyun, Bogovic, John, Eckstein, Nils, Petruncio, Alyson, Clements, Jody, Xu,

C Shan, Funke, Jan, et al. Automatic whole cell organelle segmentation in volumetric electron microscopy. *bioRxiv*, 2020.

[26] Cai, Yin, Hossain, M Julius, Hériché, Jean-Karim, Politi, Antonio Z, Walther, Nike, Koch, Birgit, Wachsmuth, Malte, Nijmeijer, Bianca, Kueblbeck, Moritz, Martinic-Kavur, Marina, et al. Experimental and computational framework for a dynamic protein atlas of human cell division. *Nature*, 561(7723):411–415, 2018.

[27] Cho, Nathan H, Cheveralls, Keith C, Brunner, Andreas-David, Kim, Kibeom, Michaelis, André C, Raghavan, Preethi, Kobayashi, Hirofumi, Savy, Laura, Li, Jason Y, Canaj, Hera, et al. Opencell: Endogenous tagging for the cartography of human cellular organization. *Science*, 375(6585):eabi6983, 2022.

[28] Viana, Matheus P, Chen, Jianxu, Knijnenburg, Theo A, Vasan, Ritvik, Yan, Calysta, Arakaki, Joy E, Bailey, Matte, Berry, Ben, Borensztejn, Antoine, Brown, Jackson M, et al. Robust integrated intracellular organization of the human ips cell: where, how much, and how variable. *biorxiv*, pages 2020–12, 2021.

[29] Xu, C Shan, Pang, Song, Shtengel, Gleb, Müller, Andreas, Ritter, Alex T, Hoffman, Huxley K, Takemura, Shin-ya, Lu, Zhiyuan, Pasolli, H Amalia, Iyer, Nirmala, et al. An open-access volume electron microscopy atlas of whole cells and tissues. *Nature*, 599(7883):147–151, 2021.

[30] Breiman, Leo. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[31] Cherkauer, Kevin J. Human expert-level performance on a scientific image analysis task by a system using combined artificial neural networks. In *Working notes of the AAAI workshop on integrating multiple learned models*, volume 21. Citeseer, 1996.

[32] Horton, Paul and Nakai, Kenta. Better prediction of protein cellular localization sites with the it k nearest neighbors classifier. In *Ismb*, volume 5, pages 147–152, 1997.

[33] Ortiz de Solorzano, C, Malladi, R, Lelievre, SA, and Lockett, SJ. Segmentation of nuclei and cells using membrane related protein markers. *journal of Microscopy*, 201(3):404–415, 2001.

[34] Al-Kofahi, Yousef, Zaltsman, Alla, Graves, Robert, Marshall, Will, and Rusu, Mirabela. A deep learning-based algorithm for 2-d cell segmentation in microscopy images. *BMC bioinformatics*, 19(1):1–11, 2018.

[35] Sullivan, Devin P and Lundberg, Emma. Seeing more: A future of augmented microscopy. *Cell*, 173(3):546–548, 2018.

[36] Ruan, Xiongtao Murphy, Robert F. Evaluation of methods for generative modeling of cell and nuclear shape. *Bioinformatics*, 2018.

[37] Rivenson, Yair, Liu, Tairan, Wei, Zhensong, Zhang, Yibo, de Haan, Kevin, and Ozcan, Aydogan. Phasestain: the digital staining of label-

free quantitative phase microscopy images using deep learning. *Light: Science & Applications*, 8(1):1–11, 2019.

[38] Guo, Syuan-Ming, Yeh, Li-Hao, Folkesson, Jenny, Ivanov, Ivan E, Krishnan, Anitha P, Keefe, Matthew G, Hashemi, Ezzat, Shin, David, Chhun, Bryant B, Cho, Nathan H, et al. Revealing architectural order with quantitative label-free imaging and deep learning. *Elife*, 9:e55502, 2020.

[39] Zhuang, Fuzhen, Qi, Zhiyuan, Duan, Keyu, Xi, Dongbo, Zhu, Yongchun, Zhu, Hengshu, Xiong, Hui, and He, Qing. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

[40] Tian, Qing, Ma, Chuang, Cao, Meng, and Chen, Songcan. Unsupervised domain adaptation through transferring both the source-knowledge and target-relatedness simultaneously. *arXiv preprint arXiv:2003.08051*, 2020.

[41] Ren, Chuan-Xian, Ge, Pengfei, Yang, Peiyi, and Yan, Shuicheng. Learning target-domain-specific classifier for partial domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):1989–2001, 2020.

[42] Weiss, Karl, Khoshgoftaar, Taghi M, and Wang, DingDing. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

[43] Ribani, Ricardo and Marengoni, Mauricio. A survey of transfer learning for convolutional neural networks. In *2019 32nd SIBGRAPI conference*

on graphics, patterns and images tutorials (SIBGRAPI-T), pages 47–57. IEEE, 2019.

[44] Pan, Sinno Jialin and Yang, Qiang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[45] Dai, Wenyuan, Xue, Gui-Rong, Yang, Qiang, and Yu, Yong. Transferring naive bayes classifiers for text classification. In *AAAI*, volume 7, pages 540–545, 2007.

[46] Turhan, Burak. On the dataset shift problem in software engineering prediction models. *Empirical Software Engineering*, 17(1):62–74, 2012.

[47] Morid, Mohammad Amin, Borjali, Alireza, and Del Fiol, Guilherme. A scoping review of transfer learning research on medical image analysis using imagenet. *Computers in biology and medicine*, 128:104115, 2021.

[48] Wang, Tongxin, Johnson, Travis S, Shao, Wei, Lu, Zixiao, Helm, Bryan R, Zhang, Jie, and Huang, Kun. Bermuda: a novel deep transfer learning method for single-cell rna sequencing batch correction reveals hidden high-resolution cellular subtypes. *Genome biology*, 20(1):1–15, 2019.

[49] Widmer, Christian and Rätsch, Gunnar. Multitask learning in computational biology. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 207–216. JMLR Workshop and Conference Proceedings, 2012.

[50] Stumpf, Patrick S, Du, Xin, Imanishi, Haruka, Kunisaki, Yuya, Semba, Yuichiro, Noble, Timothy, Smith, Rosanna CG, Rose-Zerili, Matthew, West, Jonathan J, Oreffo, Richard OC, et al. Transfer learning efficiently maps bone marrow cell types from mouse to human using single-cell rna sequencing. *Communications biology*, 3(1):1–11, 2020.

[51] Zheng, Shijie C, Stein-O'Brien, Genevieve, Augustin, Jonathan J, Slosberg, Jared, Carosso, Giovanni A, Winer, Briana, Shin, Gloria, Bjornsson, Hans T, Goff, Loyal A, and Hansen, Kasper D. Universal prediction of cell-cycle position using transfer learning. *Genome biology*, 23(1):1–27, 2022.

[52] Chen, Jintai, Ying, Haochao, Liu, Xuechen, Gu, Jingjing, Feng, Ruiwei, Chen, Tingting, Gao, Honghao, and Wu, Jian. A transfer learning based super-resolution microscopy for biopsy slice images: the joint methods perspective. *IEEE/ACM transactions on computational biology and bioinformatics*, 18(1):103–113, 2020.

[53] Christensen, Charles N, Ward, Edward N, Lu, Meng, Lio, Pietro, and Kaminski, Clemens F. Ml-sim: universal reconstruction of structured illumination microscopy images using transfer learning. *Biomedical optics express*, 12(5):2720–2733, 2021.

[54] Gessert, Nils, Bengs, Marcel, Wittig, Lukas, Drömann, Daniel, Keck, Tobias, Schlaefer, Alexander, and Ellebrecht, David B. Deep transfer learning methods for colon cancer classification in confocal laser mi-

croscopy images. *International journal of computer assisted radiology and surgery*, 14(11):1837–1845, 2019.

[55] Munien, Chanaleä and Viriri, Serestina. Classification of hematoxylin and eosin-stained breast cancer histology microscopy images using transfer learning with efficientnets. *Computational Intelligence and Neuroscience*, 2021, 2021.

[56] Zhang, Wen, Deng, Lingfei, Zhang, Lei, and Wu, Dongrui. A survey on negative transfer. *arXiv preprint arXiv:2009.00909*, 2020.

[57] Wang, Zirui, Dai, Zihang, Póczos, Barnabás, and Carbonell, Jaime. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11293–11302, 2019.

[58] Horwitz, Rick. Interdisciplinary team science in cell biology. *Trends in cell biology*, 26(11):796–798, 2016.

[59] Roberts, Brock, Haupt, Amanda, Tucker, Andrew, Grancharova, Tanya, Arakaki, Joy, Fuqua, Margaret A, Nelson, Angelique, Hookway, Caroline, Ludmann, Susan A, Mueller, Irina A, et al. Systematic gene tagging using crispr/cas9 in human stem cells to illuminate cell organization. *Molecular biology of the cell*, 28(21):2854–2874, 2017.

[60] Drubin, David G and Hyman, Anthony A. Stem cells: the new “model organism”. *Molecular biology of the cell*, 28(11):1409–1411, 2017.

[61] Schwarz, Dianne S and Blower, Michael D. The endoplasmic reticulum: structure, function and response to cellular signaling. *Cellular and molecular life sciences*, 73(1):79–94, 2016.

[62] Cooper, Geoffrey M, Hausman, Robert E, and Hausman, Robert E. *The cell: a molecular approach*, volume 4. ASM press Washington, DC, 2007.

[63] Rajasekaran, Sigrid A, Beyenbach, Klaus W, and Rajasekaran, Ayyappan K. Interactions of tight junctions with membrane channels and transporters. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1778(3):757–769, 2008.

[64] Rajasekaran, Ayyappan K and Rajasekaran, Sigrid A. Role of na-k-atpase in the assembly of tight junctions. *American Journal of Physiology-Renal Physiology*, 285(3):F388–F396, 2003.

[65] Cereijido, Marcelino, Contreras, Rubén G, Shoshani, Liora, Flores-Benitez, David, and Larre, Isabel. Tight junction and polarity interaction in the transporting epithelial phenotype. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1778(3):770–793, 2008.

[66] Köhler, Katja and Zahraoui, Ahmed. Tight junction: a co-ordinator of cell signalling and membrane trafficking. *Biology of the Cell*, 97(8): 659–665, 2005.

[67] Ugalde-Silva, Paul, Gonzalez-Lugo, Octavio, and Navarro-Garcia, Fernando. Tight junction disruption induced by type 3 secretion system effectors injected by enteropathogenic and enterohemorrhagic escherichia coli. *Frontiers in cellular and infection microbiology*, page 87, 2016.

[68] Regan-Klapisz, Elsa, Krouwer, Vincent, Langelaar-Makkinje, Miriam, Nallan, Laxman, Gelb, Michael, Gerritsen, Hans, Verkleij, Arie J, and Post, Jan Andries. Golgi-associated cpla2α regulates endothelial cell–cell junction integrity by controlling the trafficking of transmembrane junction proteins. *Molecular biology of the cell*, 20(19):4225–4234, 2009.

[69] Egea, Gustavo, Serra-Peinado, Carla, Gavilan, Maria P, and Rios, Rosa Maria. Cytoskeleton and golgi-apparatus interactions: a two-way road of function and structure. *Cell Health and Cytoskeleton*, 7:37, 2015.

[70] Harada, A, Takei, Y, Kanai, Y, Tanaka, Y, Nonaka, S, and Hirokawa, N. Golgi vesiculation and lysosome dispersion in cells lacking cytoplasmic dynein. *The Journal of cell biology*, 141(1):51–59, 1998.

[71] Dubitzky, Werner, Wolkenhauer, Olaf, Cho, Kwang-Hyun, and Yokota, Hiroki. *Encyclopedia of systems biology*, volume 402. Springer New York, 2013.

[72] Helms, J Bernd and Rothman, James E. Inhibition by brefeldin a of a golgi membrane enzyme that catalyses exchange of guanine nucleotide bound to arf. *Nature*, 360(6402):352–354, 1992.

[73] Karaman, Mazen W, Herrgard, Sanna, Treiber, Daniel K, Gallant, Paul, Atteridge, Corey E, Campbell, Brian T, Chan, Katrina W, Ciceri, Pietro, Davis, Mindy I, Edeen, Philip T, et al. A quantitative analysis of kinase inhibitor selectivity. *Nature biotechnology*, 26(1):127–132, 2008.

[74] Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[75] Zhou, Kevin, Greenspan, Hayit, and Shen, Dinggang. *Deep learning for medical image analysis*. Academic Press, 2017.

[76] Litjens, Geert, Kooi, Thijs, Bejnordi, Babak Ehteshami, Setio, Arnaud Arindra Adiyoso, Ciompi, Francesco, Ghafoorian, Mohsen, Van Der Laak, Jeroen Awm, Van Ginneken, Bram, and Sánchez, Clara I. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[77] Kingma, Diederik P and Ba, Jimmy. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[78] Raina, Rajat, Battle, Alexis, Lee, Honglak, Packer, Benjamin, and Ng, Andrew Y. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, pages 759–766, 2007.

[79] Dai, Wenyuan, Yang, Qiang, Xue, Gui-Rong, and Yu, Yong. Self-taught clustering. In *Proceedings of the 25th international conference on Machine learning*, pages 200–207, 2008.

[80] Wang, Zheng, Song, Yangqiu, and Zhang, Changshui. Transferred dimensionality reduction. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 550–565. Springer, 2008.

[81] Dai, Wenyuan, Xue, Gui-Rong, Yang, Qiang, and Yu, Yong. Co-clustering based classification for out-of-domain documents. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 210–219, 2007.

[82] Schwaighofer, Anton, Tresp, Volker, and Yu, Kai. Learning gaussian process kernels via hierarchical bayes. *Advances in neural information processing systems*, 17, 2004.

[83] Gao, Jing, Fan, Wei, Jiang, Jing, and Han, Jiawei. Knowledge transfer via multiple model local structure mapping. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 283–291, 2008.

[84] Mihalkova, Lilyana, Huynh, Tuyen, and Mooney, Raymond J. Mapping and revising markov logic networks for transfer learning. In *Aaai*, volume 7, pages 608–614, 2007.

[85] Gong, Mingming, Zhang, Kun, Liu, Tongliang, Tao, Dacheng, Glymour, Clark, and Schölkopf, Bernhard. Domain adaptation with conditional transferable components. In *International conference on machine learning*, pages 2839–2848. PMLR, 2016.

[86] Jin, Yuncheng, Chen, Jiajia, Wu, Chenxue, Chen, Zhihong, Zhang, XIngyu, Shen, Hui-liang, Gong, Wei, and Si, Ke. Wavefront reconstruction based on deep transfer learning for microscopy. *Optics Express*, 28 (14):20738–20747, 2020.

[87] Caicedo, Juan C, Cooper, Sam, Heigwer, Florian, Warchal, Scott, Qiu, Peng, Molnar, Csaba, Vasilevich, Aliaksei S, Barry, Joseph D, Bansal, Harmanjit Singh, Kraus, Oren, et al. Data-analysis strategies for image-based cell profiling. *Nature methods*, 14(9):849–863, 2017.

[88] Glancy, Brian, Kim, Yuho, Katti, Prasanna, and Willingham, T Bradley. The functional impact of mitochondrial structure across subcellular scales. *Frontiers in physiology*, 11:541040, 2020.

תוכן עניינים

ישום תיוג באמצעות למידה עמוקה והעברת למידה על מנת לנתח קשרי אברון־אברון במרחב

סמוליאנסקי יקטרינה

עבודת גמר לתואר מוסמך למדעי הטבע

אוניברסיטת בן־גוריון בנגב

2022

תקציר

כיצד אברונים פועלים יחד ליצירת הרכב המאפשר את תפקוד התא היא שאלה בסיסית בביולוגיה של התא. ההתקדמות בתחום מעוכבת על ידי היעדר כלים שיטתיים לנתח אינטראקציות מרחביות בזמן בין־אורגנלות, בעיקר בשל מגבלות טכניות בתיוג בו־זמני של מספר אברונים בתוך אותו תא חי. אנו משלבים טכניקות חישוביות חדשניות ומערך נתונים מקיף זמין לציבור של הדמיית תאים תלת מימדית כדי לאמת קשרים ידועים ולהעלות השערות חדשות לגבי תלות מרחבית בין־אברונים.

לאחרונה הוכח כי טכניקות מודרניות של לימוד מכונה מצליחות לחלץ מיקומי האברונים בתוך התא מתוך תמונות ללא תווית, על ידי טכניקה המכונה "תיוג אינסיליקו". העברת למידה היא טכניקת למידת מכונה שבה מודלים חישוביים שהוכשרו למשימה אחת משמשים כנקודת התחלה לאימון מודל

חדש למשימה אחרת אך קשורה. האינטואיציה להצלחת למידה בהעברה נובעת מההבנה שאימון של מודלים תוך שימוש בנקודמת התחלה טובה יותר יכולים לספק יתרונות כאשר נתוני האימון מוגבלים בנפחם. אנו משתמשים בתכונה זו כאמצעי למדידת תלות מרחבית בין אברונים שונים. אם מודל תיוג "אין סיליקו" שאומן ללוקליזציה של אברון משופר על ידי העברת למידה ממודל שאומן ללוקליזציה של אברון אחר, אנחנו אומרים שהמידע המקודד במיפוי לאברון האחרון שימושי לחיזוי של האברון הקודם. השוואה כזו בין שני מודלים מגדירה קישור א־סימטרי המקודד תלות מרחבית בין האברונים האחרונים לאברונים הראשונים. יישמנו מתודולוגיה זו לבניית רשת תלות מרחבית בין־אברון על ידי שילוב היחסים הזוגיים בין 13 אברונים בתאי גזע פלוריפוטנטיים שצולמו בתלת־ממד, המסומנים באופן אנדוגני, הזמינים על ידי מכון אלן למדעי התא.

התוצאות הראשוניות שלנו מאמתות מספר קשרים ידועים בין אברונים ומציעות תחזיות חדשות שיש לאמתן בניסוי. לדוגמה, גם מעטפת הגרעין וגם רשתית תוך־פלזמית תורמים לחיזוי של מנגנון הגולג'י, והמעטפת הגרעינית תורמת לחיזוי הרשתית התוך פלזמית. הגולג'י ו"המצתים ההדוקים" מקושרים דו־כיוונית. קרום הפלזמה תורם לניבוי של "הצמתים הצמודים" ואף אברון לא משפר את חיזוי הדסמוזומים.

אולטימיבית, הפרויקט שלנו יספק מתודולוגיה מקיפה ומאומתת לחיזוי אינטראקציות אברונים־אברונים, תוך עקיפת מחסומים טכניים ארוכי שנים במיקרוסקופיה. הקמת משאב עשיר של רשתות אינטראקציה חזויה של אברון־אברון יכולה לספק קפיצת מדרגה גדולה לעבר "הגביע הקדוש" של ביולוגיה של התא ־ הבנה כוללת של תאים כמערכות מורכבות משולבות.

אוניברסיטת בן־גוריון בנגב

הפקולטה למדעי הטבע

המחלקה למדעי המחשב

ישום תיוג באמצעות למידה עמוקה והעברת למידה על מנת לנתח קשרי אברון־אברון במרחב

חיבור זה מהווה חלק מהדרישות לקבלת התואר מוסמך למדעי
הטבע (M.Sc)

סמוליאנסקי יקטרינה

בהנחיית **אסף זריצקי ואנדרי שרף**

אוגוסט 2022