Cell Nuclear Morphology Analysis Using 3D Shape Modeling, Machine Learning and Visual Analytics

by

Alexandr Kalinin

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2018

Doctoral Committee:

Professor Brian D. Athey, Co-Chair
Professor Ivo D. Dinov, Co-Chair
Associate Professor Jason J. Corso
Research Professor Gerald Higgins
Professor Kayvan Najarian

Alexandr Kalinin

akalinin@umich.edu

ORCID ID: 0000-0003-4563-3226

To my dear parents, Natalya and Andrey, my brother Anton, and my niece Masha.

ACKNOWLEDGEMENTS

This dissertation is a result of a journey that took longer than just my doctoral studies. Along the way, I have relied on and have been helped by many people, whose combined explicit and implicit impact on this work exceeds my own by a great degree.

Foremost, I would like express my deep gratitude to my advisors, Ivo Dinov and Brian Athey. Since we have worked together at UCLA, Ivo's patience and careful guidance was central for my professional and personal development as a researcher. Brian's energy and leadership helped me to learn seeing the bigger picture and to be more confident in taking ownership of my work. But most of all, I thank them for believing in me as a PhD applicant and welcoming me onto this journey that I have not been originally planing on taking, but which turned out more rewarding than I could have ever imagined.

I would also like to thank my dissertation committee members, Gerry Higgins, Kayvan Najarian, and Jason Corso for their invaluable opinions and contributions that in many ways have shaped my multiple projects into the final form of this dissertation, and helped me to stay on track and to press on with completing my doctoral program on time.

Many colleagues, faculty, and staff at the Department of Computational Medicine and Bioinformatics, Athey lab, and Statistics Online Computational Resource (SOCR) at the University of Michigan have provided important contributions including ideas, improvement suggestions, and other assistance in the development of this work. I am also thankful to the online data science community ODS.ai for providing valuable

resources and motivation to be involved in important side projects, and specifically to Vladimir Iglovikov, Alexey Shvets, and Alexander Rakhlin for welcoming me to work and publish together with them. The Fulbright Program and the Stanford US-Russia Forum (SURF) have provided me with unique opportunities for meeting many amazing leaders who have inspired and motivated me to work beyond my course of study.

Among my other teachers, I would like to distinguish Vladimir Ivanovich Ovsov and Anatolii Nikolaevich Puliaev, who went an extra mile to instill love of math and sciences in me.

Among the many friends that have helped me in maintaining my sanity while in Ann Arbor, I am especially grateful to, in order of appearance, Shashank Jariwala, Andreia Gonçalves, Ricardo D'Oliveira Albanus, Sushma Chaluvadi, Evgeny Kagan, Siyu Liu, and Christopher Castro for being there for me in all ups and downs of my journey and sharing theirs back with me. What I have learned from and with them over many beers is not in any way less than what I got out of my formal training. I also thank Maxim Shvetsov, Pavel Ovchinnikov, Nikolay Golovko, Dmitry Ponomarev, Viktoria Kvashnina, Alexander Kuznetsov, and all my other friends in Russia for cheering for me from overseas over the years and for always celebrating my brief visits back home.

Most of all, I thank my dear parents, Natalya and Andrey, my brother Anton, and my niece Masha, for their unconditional love, understanding, and support.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Quantitative analysis of morphological changes in a cell nucleus is important for the understanding of nuclear architecture and its relationship with cell differentiation, development, proliferation, and disease. Changes in the nuclear form are associated with reorganization of chromatin architecture related to altered functional properties such as gene regulation and expression. Understanding these processes through quantitative analysis of morphological changes is important not only for investigating nuclear organization, but also has clinical implications, for example, in detection and treatment of pathological conditions such as cancer.

While efforts have been made to characterize nuclear shapes in two or pseudo-three dimensions, several studies have demonstrated that three dimensional (3D) representations provide better nuclear shape description, in part due to the high variability of nuclear morphologies. 3D shape descriptors that permit robust morphological analysis and facilitate human interpretation are still under active investigation. A few methods have been proposed to classify nuclear morphologies in 3D, however, there is a lack of publicly available 3D data for the evaluation and comparison of such algorithms. There is a compelling need for robust 3D nuclear morphometric techniques to carry out population-wide analyses.

In this work, we address a number of these existing limitations.

First, we present a largest publicly available, to-date, 3D microscopy imaging dataset for cell nuclear morphology analysis and classification. We provide a detailed description of the image analysis protocol, from segmentation to baseline evaluation of a number of popular classification algorithms using 2D and 3D voxel-based mor-

phometric measures. We proposed a specific cross-validation scheme that accounts for possible batch effects in data.

Second, we propose a new technique that combines mathematical modeling, machine learning, and interpretation of morphometric characteristics of cell nuclei and nucleoli in 3D. Employing robust and smooth surface reconstruction methods to accurately approximate 3D object boundary enables the establishment of homologies between different biological shapes. Then, we compute geometric morphological measures characterizing the form of cell nuclei and nucleoli. We combine these methods into a highly parallel computational pipeline workflow for automated morphological analysis of thousands of nuclei and nucleoli in 3D. We also describe the use of visual analytics and deep learning techniques for the analysis of nuclear morphology data.

Third, we evaluate proposed methods for 3D surface morphometric analysis of our data. We improved the performance of morphological classification between epithelial vs mesenchymal human prostate cancer cells compared to the previously reported results due to the more accurate shape representation and the use of combined nuclear and nucleolar morphometry. We confirmed previously reported relevant morphological characteristics, and also reported new features that can provide insight in the underlying biological mechanisms of pathology of prostate cancer. We also assessed nuclear morphology changes associated with chromatin remodeling in drug-induced cellular reprogramming. We computed temporal trajectories reflecting morphological differences in astroglial cell sub-populations administered with 2 different treatments vs controls. We described specific changes in nuclear morphology that are characteristic of chromatin re-organization under each treatment, which previously has been only tentatively hypothesized in literature. Our approach demonstrated high classification performance on each of 3 different cell lines and reported the most salient morphometric characteristics.

We conclude with the discussion of the potential impact of method development

in nuclear morphology analysis on clinical decision-making and fundamental investigation of 3D nuclear architecture. We consider some open problems and future trends in this field.

CHAPTER I

Introduction

The cell nucleus is an essential structure that contains the genome and maintains its three-dimensional structural organization (*Wilson*, 1925; *White*, 1977; *Jevtić et al.*, 2014; *Stephens et al.*, 2018b). Nuclear morphology is a study of size and shape of a cell nucleus that are regulated by complex biological mechanisms related to cell differentiation, development, proliferation, and disease (*Jevtić et al.*, 2014; *Uhler and Shivashankar*, 2018). More specifically, morphology of a cell nucleus is determined by both the cytoskeletal links and the degree of chromatin condensation within the nucleus (*Uhler and Shivashankar*, 2018). DNA is folded into a chromatin fiber by histone and nonhistone proteins and the associated chemical modifications on the histone proteins (*Allis and Jenuwein*, 2016). The chromatin fiber, depending on such modifications as histone acetylation or methylation, dictates the higher-order compaction, thereby implying both genetic and nongenetic functions of the genome (*Bustin and Misteli*, 2016). This higher-order structure comprises megabase-pair topologically associated domains (TADs) leading to a highly organized chromatin structure (*Higgins et al.*, 2015; *Gonzalez-Sandoval and Gasser*, 2016). Changes in nuclear morphology are reflective of reorganization of chromatin architecture and are related to altered functional properties such as gene regulation and expression (*Jevtić et al.*, 2014; *Uhler and Shivashankar*, 2018). Conversely, studies in mechanobiol-

ogy show that external geometric constraints and mechanical forces that deform the cell nucleus affect chromatin dynamics and gene and pathway activation (*Uhler and Shivashankar*, 2017, 2018). Recent studies showed revealed separate roles for chromatin and lamins in determining the nuclear mechanical properties and morphology (*Stephens et al.*, 2018a,b). Thus, nuclear morphological quantification becomes of major relevance as the studies of the reorganization of the chromatin and DNA architecture in the spatial and temporal framework, known as the 4D nucleome, emerge (*Chen et al.*, 2015; *Cremer et al.*, 2015; *Higgins et al.*, 2015; *Zheng et al.*, 2018). Cellular structures of interest in the context of the 4D nucleome include not only the nucleus itself, but also the nucleolus and nucleolar-associating domains, chromosome territories, TADs, lamina-associating domains, and loop domains in transcription factories (*Higgins et al.*, 2015, 2017).

At the same time, quantitative analyses of nuclear and nucleolar morphological changes also have clinical implications, for example, in detection and treatment of pathological conditions such as cancer (*Montanaro et al.*, 2008; *Veltri and Christudass*, 2014; *Zink et al.*, 2004). Abnormal nuclear morphology has been used as one of the gold standards for cancer diagnoses for nearly a century in tests such as the Pap smear (*Zink et al.*, 2004; *Papanicolaou and Traut*, 1941; *Stephens et al.*, 2018b; *Uhler and Shivashankar*, 2018). However, since morphologies in cell populations are highly heterogeneous, morphometric assays performed by pathologists are highly subjective and rely on human interpretation (*Uhler and Shivashankar*, 2018). To address these limitations, automated analysis of cell morphology is employed to improve the accuracy and efficiency for the pathology detection. For example, the Cell-CT® platform relies on automated 3D morphometry and machine learning algorithms to assess the morphology of epithelial cells in sputum samples for early stage lung cancer detection (*Wilbur et al.*, 2015; *Meyer et al.*, 2015; *Pantanowitz et al.*, 2018). The relevance of nuclear morphology to both understanding fundamental principles of cellular organi-

zation and improving disease detection and treatment presents a compelling need for accurate and robust ways to analyze cell nuclear morphology.

While many efforts have been made to develop cell and nuclear morphological characteristics in 2D or pseudo-3D (*Huang et al.*, 2014b; *Pincus and Theriot*, 2007), several studies have suggested that 3D measures provide better results for nuclear morphometry description and discrimination (*Choi and Choi*, 2007; *Meyer et al.*, 2009). Although a number of signal processing and computer vision algorithms have been proposed to analyze cell and nuclear morphological phenotypes using 3D representations (*Dufour et al.*, 2015), there is a lack of publicly available 3D cell imaging datasets that could serve for the evaluation of various tools and methods. This limitation becomes of great importance in the modern reality of big data microscopy, when the ability to evaluate different approaches on publicly available data is needed for better dissemination of the current state of the art methods for bioimage analysis (*Caicedo et al.*, 2017; *Meijering et al.*, 2016).

The way nuclear morphologies can be quantified depends on their representation extracted from image data (*Pincus and Theriot*, 2007). Many 3D morphometric measures are applied as is to 3D geometric objects represented by volumetric data. However, such morphological representations can be noisy, and they may lose fine geometric details or even break the objects topological structure. 3D shape descriptors that permit robust morphological analysis and facilitate human interpretation are still under active investigation (*Dufour et al.*, 2015). Additionally, the dimensionality and volume of acquired data, various image acquisition conditions, and great variability of cell shapes in a population present challenges for 3D shape analysis methods that should be scalable, robust to noise, and specific enough across cell populations at the same time. Thus, there is a need for robust 3D nuclear morphometric techniques to carry out population-wide analysis (*Pegoraro and Misteli*, 2016).

The remainder of this work examines and addresses a number of current limitations

relevant to the 3D nuclear morphological analysis.

The second chapter addresses the lack of available 3D imaging data for nuclear morphology analysis and describes a new, biggest publicly available dataset of 3D microscopic images for cell nuclear morphology analysis and classification. We provide detailed description of image analysis protocol, from segmentation to a baseline evaluation of a number of popular classification algorithms using 2D and 3D voxel-based morphometric measures. Contents of the first chapter were partially published in *Kalinin et al.* (2018d,a).

In the third chapter we focus on the development of the robust and accurate 3D nuclear morphometry methods. We employ 3D surface modeling, morphometric feature extraction, and machine learning to construct a high-throughput computational workflow for automated 3D morphology classification. We also discuss the use of visual analytics and deep learning for the analysis of nuclear morphology data. Contents of the second chapter were partially published in *Kalinin et al.* (2017, 2018c,b); *Ching et al.* (2018).

In the fourth chapter we evaluate proposed methods for 3D surface morphometry on our data. Specifically, we demonstrate the efficiency and accuracy of out approach for morphological discrimination of combined nuclear and nucleolar morphologies of prostate cancer cells in epithelial and mesenchymal conditions, outperforming previously proposed solutions. We also introduce a new experiment on evaluating nuclear morphology changed over time during drug-induced cellular reprogramming of astroglial cells. We were able to quantitatively show the differences in over time morphometric measures three-way: between two treatments and the controls. High efficiency and analytical performance of proposed methods are demonstrated, along with the ability to derive new biological insight from the obtained results. Contents of the fourth chapter were partially published in *Kalinin et al.* (2018c); *Ching et al.* (2018).

We conclude with the discussion of the potential impact of method development in nuclear morphology analysis on clinical decision-making and fundamental investigation of 3D nuclear architecture. We discuss future perspectives in both basic science and transnational applications as well as open science considerations. Finally, we consider some open problems and future trends in this field. Some conclusions were partially published in *Kalinin et al.* (2018c,b); *Ching et al.* (2018).

CHAPTER II

3D Cell Nuclear Morphology Imaging Dataset

2.1 Introduction

Although a number of signal processing and computer vision algorithms have been proposed to analyze cell and nuclear morphological phenotypes using 3D representations (*Dufour et al.*, 2015), there is a lack of publicly available 3D cell imaging datasets that could serve for the evaluation of various tools and methods. This limitation becomes of great importance in the modern reality of big data microscopy, when the ability to evaluate different approaches on publicly available data is needed for better dissemination of the current state of the art methods for bioimage analysis (*Meijering et al.*, 2016; *Caicedo et al.*, 2017; *Ellenberg et al.*, 2018).

To begin to address the lack of data for 3D cell nuclear morphological analysis and enable objective evaluation of the methods for nuclear morphometric classification, we created a 3D cell nuclear morphology dataset (*Kalinin et al.*, 2018d). The dataset includes 3D confocal fluorescence microscopy volumetric images of cell nuclei and nucleoli of two different cell collections: primary human fibroblast cells and human prostate cancer cells (PC3). In turn, each collection contains images of cells in two different phenotypic states that have previously been shown to exhibit quantifiable changes in nuclear or nucleolar morphology. This allows for the evaluation of quntitative methods in morphometry on 2 sub-sets of data as binary classification

problems.

We also provide a baseline classification performance evaluation of simple voxel-based morphometric analysis methods. First, we use 3D automatic segmentation methods to extract individual nuclear and nucleolar binary masks from the original z-stack images. We then extract common 2D and 3D voxel-based measures of binary mask morphology and combine them into per-nucleus feature vectors. These feature vectors then used to evaluate a number of machine learning algorithms to provide morphology classification performance baselines. To account for batch effects, while enabling calculations of interval estimates for the Area under the Precision-Recall curve (AUPR) and the Area Under the Receiver Operating Characteristic (ROC) curve (AUC) performance metrics, we propose a specific cross-validation scheme.

Finally, we evaluate how much of the difference between cell phenotypic conditions can be explained by morphology as opposite to the pixel intensity information. We perform classification of both binary and intensity 2D projections of 3D microscopic images of fibroblast cells. More specifically, we compare direct classification of pixel data from either raw intensity images or binary masks, which contain only object morphology information, but not texture.

2.2 Sample preparation

The presented dataset is composed of two different cell collections (*Kalinin et al.*, 2018d). Each collection includes 3D volumetric images of cells in two phenotypic states that have been shown to exhibit different nuclear and/or nucleolar morphology.

The first collection includes images of primary human fibroblast cells (newborn male) that were purchased from ATCC (BJ Fibroblasts CRL-2522 normal). In order to introduce morphology changes, a part of this collection was subjected to a G0/G1 Serum Starvation Protocol (*Langan and Chou*, 2011). This protocol is used for cell cycle synchronization and has previously been shown to cause morphology changes in

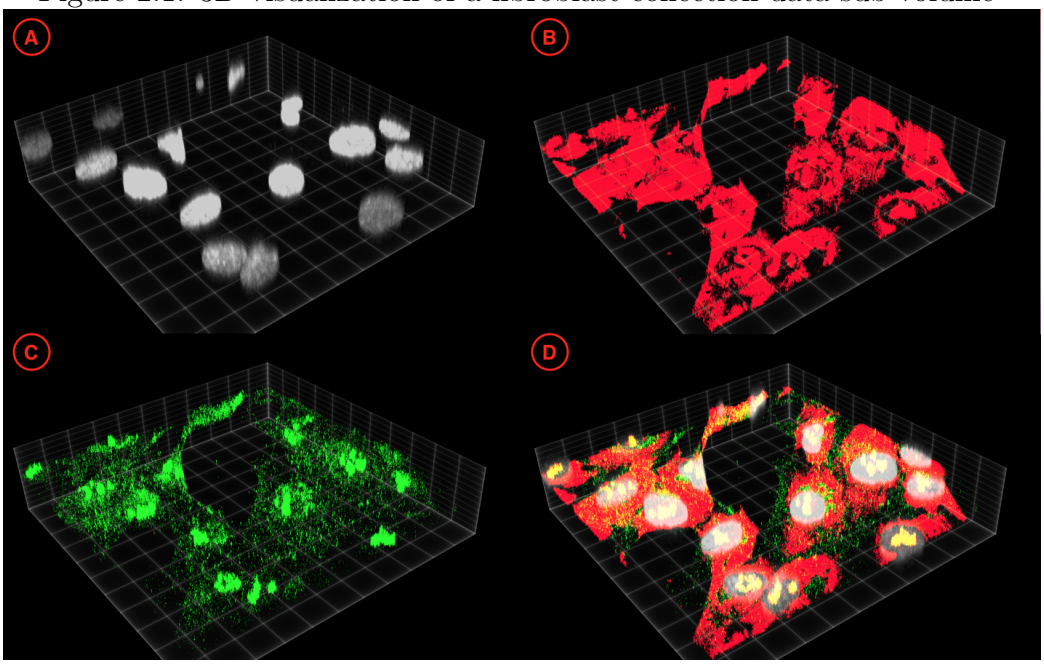
human fibroblasts, affecting nuclear size and shape (*Seaman et al.*, 2015). Full details of the fibroblast cell sample preparation protocol are given in A.1. Cell cycle profiles were confirmed for synchronized serum-starved and proliferating fibroblasts with flow cytometry, see A.4. As a result, the first collection contains 3D volumetric images of cells in the following phenotypic classes: (1) proliferating fibroblasts (PROLIF), and (2) cell cycle synchronized by the serum-starvation protocol (SS). These classes serve as two categories in a binary morphology classification setting.

The second collection contains images of human prostate cancer cells (PC3). Through the course of progression to metastasis, malignant cancer cells undergo a series of reversible transitions between intermediate phenotypic states bounded by pure epithelium and pure mesenchyme (*Veltri and Christudass*, 2014). These transitions in prostate cancer are associated with quantifiable changes in both nuclear and nucleolar structure (*Montanaro et al.*, 2008; *Verdone et al.*, 2015). Microscope slides of prostate cancer cell line PC3 were cultured in: (1) epithelial (EPI), and (2) mesenchymal transition (EMT) phenotypic states, as described in (*Verdone et al.*, 2015). Full details of the PC3 cell sample preparation protocol are given in A.2. Thus, this setting can also be treated as a binary classification task.

2.3 Image acquisition

Cells in both collections are labeled with 3 different fluorophores: DAPI (4',6-diamidino-2-phenylindole), a common stain for the nuclei, fibrillarin antibody (anti-fibrillarin) and ethidium bromide (EtBr), both used for nucleoli staining. Although anti-fibrillarin is a commonly used nucleolar label, we find it to be too specific, which makes the extraction of a shape mask problematic. It has been shown that EtBr can be used for staining dense chromatin, nucleoli, and ribosomes (*Biggiogera and Biggiogera*, 1989). We find that it provides better overall representation of nucleolar shape. Anti-fibrillarin is combined with EtBr by co-localization to confirm correct

Figure 2.1: 3D visualization of a fibroblast collection data sub-volume



Notes. Figure panels show: (A) DAPI channel; (B) EtBr channel; (C) anti-fibrillarin channel; (D) a composite image. Images are thresholded by 25% the for the clarity of visual appearance and visualized using ClearVolume (*Royer et al.*, 2015).

detection of nucleoli locations as described below. 3D imaging used a Zeiss LSM 710 laser scanning confocal microscope with a 63x PLAN/Apochromat 1.4NA DIC objective. Full details of sample staining and imaging are given in A.3 and A.5.

For multichannel data in the vendor-defined format, the channels are separated and saved as individual volumes labeled as c0, c1, c2, representing the DAPI, anti-fibrillarin, and EtBr channels, respectively, see figure 2.1. Each channel-specific volume is then re-sliced into a $1,024 \times 1,024 \times Z$ lattice ($Z = \{30, 50\}$), where regional sub-volumes facilitate the alignment with the native tile size of the microscope. All sub-volumes are saved as multi-image 3D TIFF volumes. For every sub-volume, accompanying vendor meta-data are extracted from the original data.

As a result, the fibroblasts collection includes the total of 178 sub-volumes (64 PROLIF and 112 SS), see table 2.1. The PC3 collection includes the total of 101 sub-volumes (50 EPI and 51 EMT), see table 2.2.

Class	Sub-volumes	GBs
PROLIF	64	10.6
SS	112	19.2
TOTAL	178	29.8

Table 2.1: The size of the fibroblast cell collection. Sub-volumes column shows the number of $1024 \times 1024 \times Z$ sub-volumes per channel.

Class	Sub-volumes	GBs
EPI	50	15.7
EMT	51	21.3
TOTAL	101	37.0

Table 2.2: The size of the PC3 cell collection. Sub-volumes column shows the number of $1024 \times 1024 \times Z$ sub-volumes per channel.

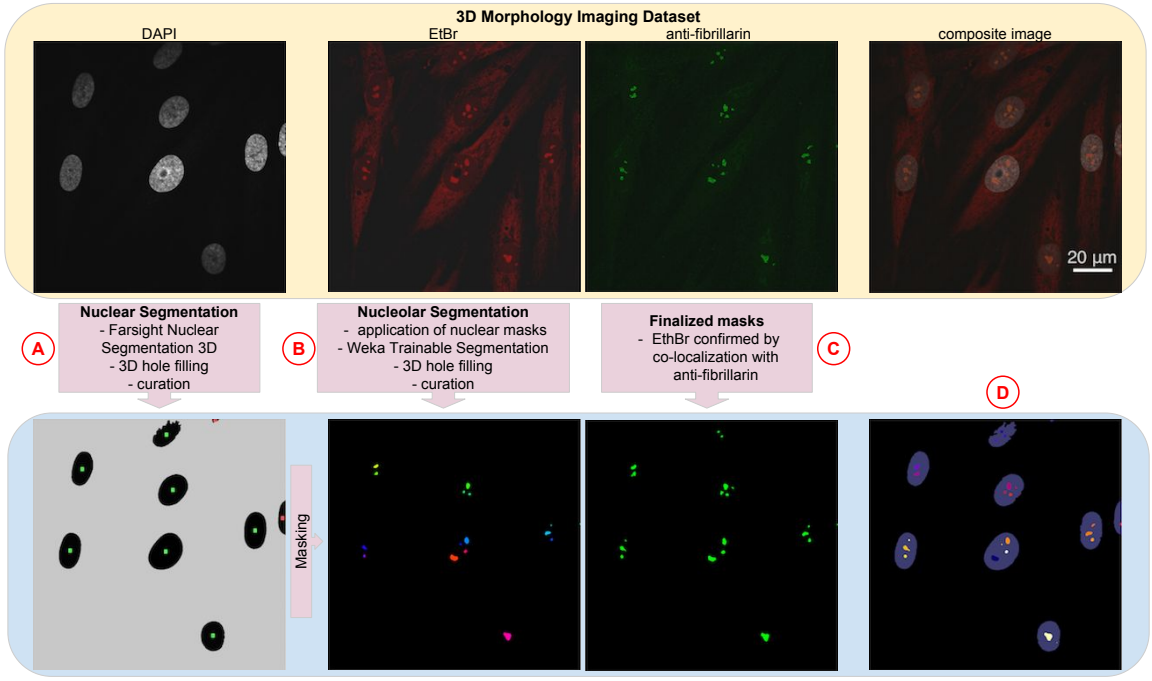
2.4 Segmentation

To establish baseline morphometry classification results, we first segment nuclei and nucleoli from the original data sub-volumes as described in (*Kalinin et al.*, 2018d). Then, we extract multiple voxel-based morphometric characteristics from 3D binary masks and their 2D projections (2D masks). We use these features to evaluate the performance of a number of widely used classification algorithms. We also assess possible batch effects in data by comparing two different cross-validation techniques.

2.4.1 Nuclear segmentation

Model-based cell segmentation approaches are the most common in bioimage analysis and typically perform well for fluorescence microscopy images of cultured cells (*Caicedo et al.*, 2017). Moreover, they allow to avoid a very labor-intensive process of manual pixel-level expert annotation of large 3D volumetric imaging data. After testing a number of implementations of 3D thresholding-based and watershed-like methods in commonly used bioimage analysis packages, we perform the automatic 3D segmentation of nuclei using Nuclear Segmentation algorithm from the Farsight

Figure 2.2: A schematic view of the dataset segmentation protocol



Notes. Figure panels show exemplar 2D slices of fibroblast data: (A) steps for the DAPI segmentation process that produces nuclear masks after hole-filling (color-coded by quality control filter); (B) steps for EtBr segmentation that outputs nucleolar masks (colored by connected component labeling); (C) co-localization nucleolar segmented masks with the segmented anti-fibrillarin channel; (D) the composite image of segmented data.

toolkit (*Al-Kofahi et al.*, 2010). This tool was created specifically to segment DAPI-stained nuclei in 2D or 3D, it does not require a labeled training set, has a convenient command line interface, and demonstrated stable results on these data. The algorithm implements multiple steps which include a graph-cut algorithm to binarize the sub-volumes, a multi-scale Laplacian of Gaussian filter to convert the nuclei to blob masks, fast clustering to delineate the nuclei, and nuclear contour refinement using graph-cuts with alpha-expansions (*Al-Kofahi et al.*, 2010).

After segmentation of the DAPI channel sub-volumes, figure 2.2, data were converted to 16-bit 3D TIFF files, each segmented nucleus was represented as a binary mask, and given a unique index value. Post-segmentation processing of nuclear masks included 3D hole filling and a filtering step that removed the objects if they span the edge of a tile, are connected to other objects, or their compactness or voxel count val-

ues were outside of the empirically estimated interval. This quality control protocol allowed to remove most of the artifacts, as confirmed by visual inspection. Details of the curation and post-processing protocol are described in A.7.

2.4.2 Nucleolar segmentation

Since nucleolar labels are not very specific and produce strong background, see figure 2.1, segmentation of nucleoli using model-based approaches did not demonstrate acceptable results. Therefore, segmentation of objects within the nucleus was performed using the Trainable Weka Segmentation (*Arganda-Carreras et al.*, 2017), a machine learning tool for microscopy pixel classification bundled with Fiji (*Schindelin et al.*, 2012), a commonly used bioimage analysis framework. The Trainable Weka Segmentation plugin is the most popular segmentation tool in the ImageJ technological landscape (*Schindelin et al.*, 2015), and it is convenient to use for labeling biological structures in 3D images, since it does not require the exact mask contour tracing. Instead, it allows the extraction of a number of features from scarcely labeled pixel groups from both classes, which then are used train a classification algorithm from the WEKA Data Mining software package (*Hall et al.*, 2009). Intra-nuclear segmentation was independently performed on EtBr and anti-fibrillarin stained nucleoli. Nuclear masks were used to isolate sub-nuclear segmentations in the EtBr and anti-fibrillarin channels to objects within a nucleus. An individual Random Forest classification model (*Liaw and Wiener*, 2002) was created for each channel by using a random selection of 10% of the sub-volumes within that channel for training. Trained models were then applied to all sub-volumes and nucleolar masks were created from the resulting probability maps and labeled as connected components, figure 2.2. Finally, both EtBr and anti-fibrillarin segmented volumes were used as input to a co-localization algorithm to validate the segmented EtBr-stained nucleoli based on the presence of anti-fibrillarin, figure 2.2.

The quality control protocol for nucleolar masks was similar to that for the nuclear masks. Since uneven staining can cause occasional segmentation artifacts, filtering step also measured spherical compactness of identified objects (*Montero and Bribiesca*, 2009) and removed the masks if their compactness were outside of the empirically estimated interval.

2.5 Baseline analysis and classification

2.5.1 Voxel-based morphometric feature extraction

We extracted 2D and 3D voxel-based morphometric features of both nuclear and nucleolar binary masks, shown in figure 2.2, using image processing library, scikit-image (*van der Walt et al.*, 2014).

The 2D feature set included: area of the object, area of the 2D bounding box, diameter of a circle with the same area as the object, ratio of the object area to the bounding box area, convex hull area, eccentricity, two biggest eigenvalues of the inertia tensor of the region, major and minor axis of an ellipse fitted to the region, the angle between the X-axis and the major axis of the fitted ellipse, perimeter of an object which approximates the contour of the region, the ratio of the region area to the convex hull area.

The set of 3D morphometry features included: object volume, volume of the 3D bounding box, diameter of a sphere with the same volume as the object, and ratio of the object volume to the bounding box volume.

In order to aggregate the nucleolar features per nucleus we computed median, minimum, maximum, and standard deviation for each morphometry measure across the nucleoli within one nucleus. Correspondingly, nuclei that did not have any internally positioned nucleoli were excluded from the further analysis. The number of detected nucleoli per nucleus was included as an individual feature. Thus, the total

number of features per nucleus was $5 \times N + 1$, where N is the number of either 2D or 3D morphometric measures.

Feature preprocessing included feature standardization by subtracting the mean and scaling to unit variance of the training set. In this study, we assigned the label of the whole image to every single cell extracted from it.

2.5.2 Cross-validation

We compared various supervised classification algorithms from scikit-learn, a popular Python machine learning toolkit (*Pedregosa et al.*, 2011), including Gaussian Naive Bayes (NB), Linear Discriminant Analysis (LDA), k nearest neighbors classifier (kNN), support vector machines with linear (SVM) and Gaussian/Radial Basis Functions (RBF) kernels, Random Forest (RF), Extremely Randomized Trees (ET), and Gradient Boosting (GBM). All classifiers used default hyper-parameters.

We evaluated the possible batch effect that could occur during the sample preparation and image acquisition (*Caicedo et al.*, 2017). In order to do so, we compare the traditional k-fold cross-validation (CV) scheme with alternatives that can account for such batch effects. A traditional approach to address this issue is to use Leave-One-Group-Out CV (LOGO). LOGO is a cross-validation scheme which holds out the samples according to a provided list of groups. In case of image-level labeled nuclei classification, group information encodes the image from which the specific nucleus was extracted. However, the disadvantage of LOGO lies in inability to compute per-split metrics such as the Area Under the Receiver Operating Characteristic (ROC) curve (AUC), the Area under the Precision-Recall curve (AUPR), and F1 score, as they require the presence of samples from both classes in the testing set.

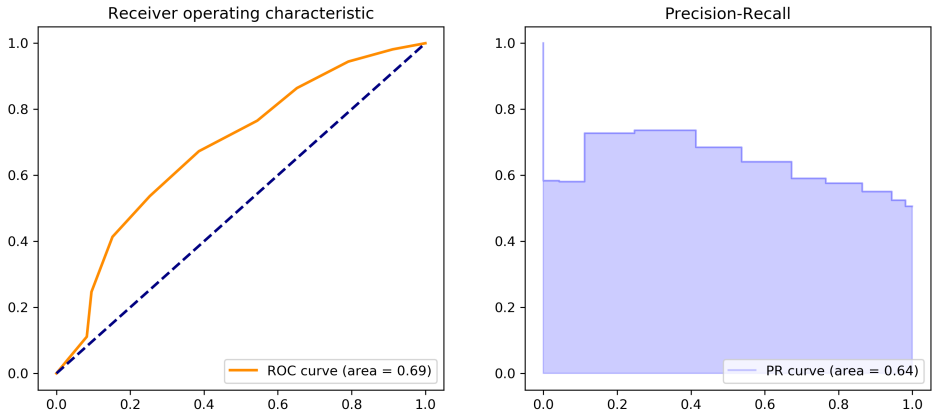
A receiver operating characteristic, or simply ROC curve, is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied (*Davis and Goadrich*, 2006), see figure 2.3. It is created by plotting the

fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold settings. TPR is also known as sensitivity, and FPR is one minus the specificity or true negative rate.

The precision-recall curve shows the trade-off between precision and recall for different threshold, (*Davis and Goadrich*, 2006). A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate, see figure 2.3. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall). A system with high recall but low precision returns many results, but most of its predicted labels are incorrect when compared to the training labels. A system with high precision but low recall is just the opposite, returning very few results, but most of its predicted labels are correct when compared to the training labels. An ideal system with high precision and high recall will return many results, with all results labeled correctly. We do not use any interpolation to compute AUPR, since a linear interpolation of points on the precision-recall curve provides an overly-optimistic measure of classifier performance (*Davis and Goadrich*, 2006; *Flach and Kull*, 2015).

As both AUC and AUPR rely on the presence of both positives and negatives in the test set, they cannot be calculated per-split in LOGO scheme, all nuclei from a single image (group) will have the same label. One option is to go through all splits in LOGO and predict labels on the whole dataset before calculating global values of AUC and AUPR, but that method only gives a point estimate of a metric, but not the interval. Instead, we suggested Leave-2-Opposite-Groups-Out (L2OGO) scheme, see figure 2.4. L2OGO ensures that: (1) all masks derived from one image fall either in the training or testing set, and (2) testing set always contains masks from 2 images of different classes. Each training set is thus constituted by all the samples except the

Figure 2.3: ROC and PR curves



Notes. Examples of ROC and PR curves for voxel-based morphological classification of fibroblast nuclei in 2D. 33% of data was used for testing and the rest was used to train a Random Forest classification model with default parameters from the scikit-learn software package (*Pedregosa et al.*, 2011).

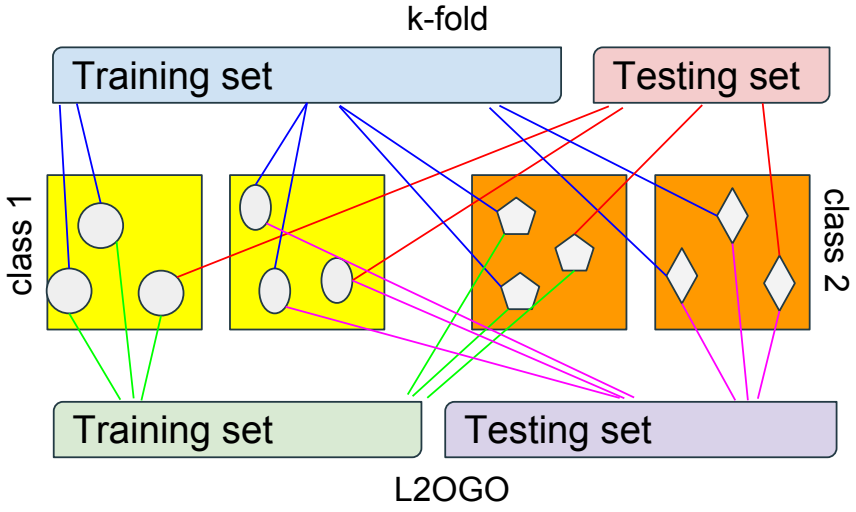
ones related to a specific group. L2OGO enables per-split evaluation of performance metrics such as AUPR and AUC. Since original volumes are of different size and contain different number of nuclei, we joined smaller volumes into bigger groups to reduce class imbalance in testing sets and the variance of the performance metric estimates. Given the fact that L2OGO may introduce or augment class imbalance, besides AUC we also compute AUPR and F1 score to compare algorithms, as they have been shown to be more suitable in such settings (*Saito and Rehmsmeier*, 2015).

2.5.3 Fibroblast voxel-based morphometric analysis

After the curation process and the exclusion of nuclei without detected nucleoli, the full collection of segmented fibroblasts consists of total 965 nuclear (498 SS and 470 PROLIF) and 2,181 nucleolar (1,151 SS and 1,030 PROLIF) binary masks. 2D and 3D morphometric measures of nuclear and nucleolar masks are merged into per-nucleus feature vectors as described above.

Next, we evaluate the performance of algorithms for fibroblast morphometric clas-

Figure 2.4: Comparison of cross-validation strategies



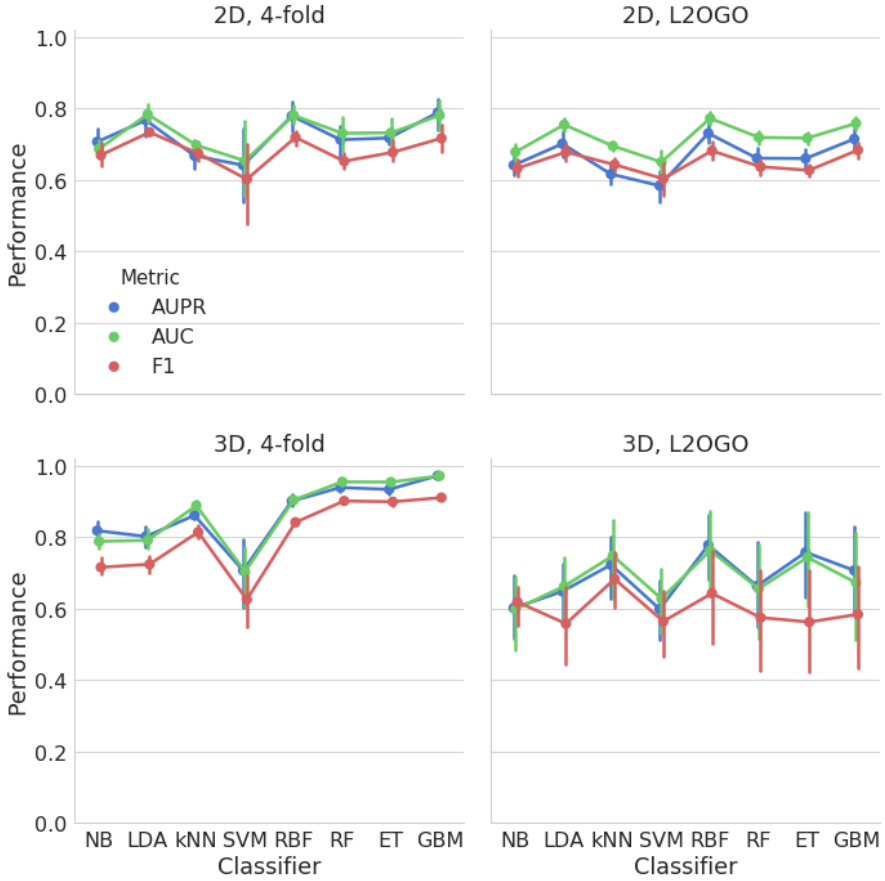
Notes. Schematic comparison of k-fold cross-validation procedure with Leave 2 Opposite Groups Out cross-validation in the presence of batch effects. Two image classes are: round objects in yellow and faceted objects in orange. Batch effects are shown as slight object shape differences across images of the same class.

sification on 2 different CV schemes: 20 splits in L2OGO and a 7 times repeated 4-fold CV in order to obtain more stable estimates. Results in figure 2.5 do not show any apparent batch effects in the 2D classification setting. However, 3D performance estimates for all classifiers using L2OGO are more pessimistic compared to 4-fold CV, which indicates the possibility of batch effects and overly optimistic classification results in 4-fold CV. As expected, L2OGO led to an increased variance of metrics, especially in the F1 score, which can be explained by classifiers' sensitivity to different class imbalances in each iteration of this scheme. Within L2OGO, a number of algorithms showed higher performance on 3D morphometry compared to 2D features. The best overall result is achieved by the Gaussian SVM (RBF) classifier in 3D with the median $AUC = 0.814 \pm 0.245$, $AUPR = 0.724 \pm 0.206$, and $F1 = 0.709 \pm 0.185$.

2.5.3.1 2D maximum intensity projection classification

Fluorescent labels are not always specific to the object of interest and often produce noisy background, see figure 2.1. In order to assess changes in the nuclear

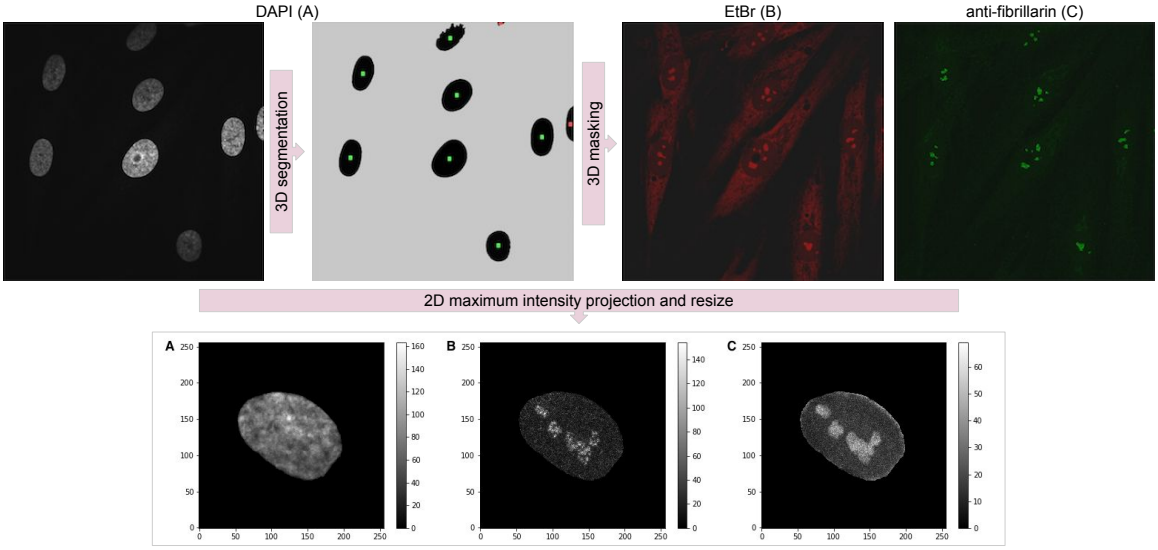
Figure 2.5: Voxel-based fibroblast morphometry classification



Notes. The comparison of cross-validation strategies and commonly used algorithms to evaluate the classification performance and possible batch effects using combined morphometric features of 2D and 3D fibroblast nuclear and nucleolar binary masks.

architecture, we first apply nuclear masks provided with the dataset to all 3 channels of original microscopy data. Due to the anisotropy in original data, we then re-scale volumes in Z dimension by a factor extracted from the corresponding meta-data. Since each of $1,024 \times 1,024 \times Z$ sub-volumes typically contains between 1 and 5 nuclei, we crop re-scaled volumes into smaller $256 \times 256 \times 57$ sub-volumes, centered at the centroid of the corresponding nuclear mask and zero-pad them, when necessary. Finally, we produce 2D representation of sub-volumes by a maximum intensity projection along the Z dimension, as shown in figure 2.6. As a result, we create a set of 999 256×256 images per channel.

Figure 2.6: Fibroblast 2D maximum intensity projections



Notes. An exemplar visualization of 256 × 256 2D maximum intensity projections of a masked, re-scaled, and cropped fibroblast sub-volumes in: (A) DAPI channel, c0; (B) anti-fibrillarin channel, c1; (C) EtBr channel, c2.

We compare classification algorithms from scikit-learn (*Pedregosa et al.*, 2011) with default hyper-parameters. Every image is flattened into a 1D feature vector. Feature preprocessing includes subtracting the mean and scaling to unit variance of the training set. In order to further assess batch effects in the intensity images and binary masks, we compare 4-fold cross-validation with L2OGO.

First, we evaluate the performance of algorithms for fibroblast nuclear classification using only 2D morphological information, i.e. binary masks. We compute AUC per channel using 2 different CV schemes: 20 splits in L2OGO and a 10 times repeated 4-fold CV. Results in table 2.3 do not show any apparent batch effects in the 2D classification setting in any of the channels, as performance levels L2OGO are only slightly lower compared to 4-fold CV. As expected, classifiers are not able to pick up complex morphological relationships from flattened binary vectors, even when 3 channels are combined. Results are dominated by the voxel-based morphometry features extracted from binary masks. The best overall result with L2OGO is achieved by the

Table 2.3: Classification AUC (*mean* ± *std*) on binary masks for 2 cross-validation schemes (CV: 4-fold and L2OGO) and a number of algorithms (Clf: NB, LDA, kNN, SVM, RBF, RF, ET, GBM) per image channel (c0, c1, c2, and all 3 channels combined). VBM: voxel-based morphometry.

		Binary images				VBM
CV	Clf	c0	c1	c2	c0c1c2	c0c2
4-fold	kNN	0.604 ± 0.056	0.615 ± 0.059	0.610 ± 0.059	0.620 ± 0.062	0.706 ± 0.030
	SVM	0.597 ± 0.067	0.583 ± 0.075	0.582 ± 0.054	0.650 ± 0.071	0.646 ± 0.090
	RBF	0.647 ± 0.053	0.648 ± 0.078	$\mathbf{0.644 \pm 0.056}$	0.678 ± 0.062	$\mathbf{0.785 \pm 0.027}$
	RF	0.635 ± 0.061	0.642 ± 0.061	0.635 ± 0.054	0.661 ± 0.064	0.735 ± 0.029
	ET	0.581 ± 0.059	0.577 ± 0.077	0.573 ± 0.058	0.651 ± 0.054	0.732 ± 0.034
	GBM	$\mathbf{0.659 \pm 0.049}$	$\mathbf{0.658 \pm 0.062}$	0.641 ± 0.061	$\mathbf{0.721 \pm 0.054}$	0.783 ± 0.032
L2OGO	kNN	0.582 ± 0.047	0.582 ± 0.047	0.582 ± 0.047	0.582 ± 0.047	0.695 ± 0.030
	SVM	0.557 ± 0.057	0.568 ± 0.050	0.555 ± 0.046	0.564 ± 0.053	0.650 ± 0.077
	RBF	0.616 ± 0.060	0.616 ± 0.060	0.616 ± 0.060	0.616 ± 0.060	$\mathbf{0.772 \pm 0.041}$
	RF	0.635 ± 0.038	0.636 ± 0.037	0.626 ± 0.050	0.627 ± 0.041	0.719 ± 0.039
	ET	0.637 ± 0.039	0.637 ± 0.039	$\mathbf{0.641 \pm 0.044}$	0.636 ± 0.042	0.717 ± 0.034
	GBM	$\mathbf{0.642 \pm 0.037}$	$\mathbf{0.644 \pm 0.038}$	0.639 ± 0.037	$\mathbf{0.643 \pm 0.041}$	0.758 ± 0.041

Gaussian SVM (RBF) classifier in with $AUC = 0.772 \pm 0.041$, $AUPR = 0.731 \pm 0.063$, and $F1 = 0.682 \pm 0.060$.

Next, we evaluate the performance using only 2D pixel intensity information. Results in table 2.4 indicate possible batch effects. The performance on the nuclear c0 channel does not benefit from the presence of additional information compared to only 2D masks. But nucleolar-stained channels c1 and c2 demonstrate 20% gain in performance even using more conservative L2OGO CV. However, L2OGO here leads to a large variance of the performance metric. On average, the EtBr channel (c2) seems to provide a sightly better representation of nucleolar structure compared to the anti-fibrillarin (c1). Almost all classifiers in both channels show results superior of those obtained with morphometric features, see table 2.3. Combining all 3 channels gives the best result, demonstrating the complement nature of stains. The best overall result is achieved by the the Gaussian SVM (RBF) classifier with $AUC = 0.990 \pm 0.029$, $AUPR = 0.980 \pm 0.040$, and $F1 = 0.877 \pm 0.177$.

Although DAPI structure classification did not benefit from using the intensity information, our results indicate usefulness of intensities of nucleolar labels: anti-

Table 2.4: Classification AUC (*mean* ± *std*) on raw intensity images for 2 cross-validation schemes (CV: 4-fold and L2OGO) and a number of algorithms (Clf: NB, LDA, kNN, SVM, RBF, RF, ET, GBM) per image channel (c0, c1, c2, and all 3 channels combined).

		Raw intensity images			
CV	Clf	c0	c1	c2	c0c1c2
4-fold	kNN	0.581 ± 0.059	0.771 ± 0.048	0.862 ± 0.041	0.865 ± 0.039
	SVM	0.610 ± 0.077	0.726 ± 0.080	0.829 ± 0.059	0.896 ± 0.043
	RBF	0.647 ± 0.058	0.814 ± 0.052	0.892 ± 0.0326	0.938 ± 0.026
	RF	0.630 ± 0.040	0.868 ± 0.039	0.890 ± 0.035	0.948 ± 0.022
	ET	0.606 ± 0.054	0.864 ± 0.045	0.875 ± 0.035	0.961 ± 0.021
	GBM	$\mathbf{0.673 \pm 0.046}$	$\mathbf{0.919 \pm 0.031}$	$\mathbf{0.912 \pm 0.026}$	$\mathbf{0.974 \pm 0.011}$
L2OGO	kNN	0.552 ± 0.030	0.755 ± 0.207	0.826 ± 0.170	0.933 ± 0.044
	SVM	0.579 ± 0.053	0.671 ± 0.166	0.794 ± 0.183	0.964 ± 0.068
	RBF	0.579 ± 0.053	0.766 ± 0.261	0.844 ± 0.204	$\mathbf{0.990 \pm 0.021}$
	RF	0.579 ± 0.053	0.823 ± 0.202	0.841 ± 0.188	0.966 ± 0.057
	ET	0.613 ± 0.047	0.816 ± 0.209	0.839 ± 0.181	0.975 ± 0.034
	GBM	$\mathbf{0.637 \pm 0.045}$	$\mathbf{0.844 \pm 0.235}$	$\mathbf{0.857 \pm 0.204}$	0.990 ± 0.029

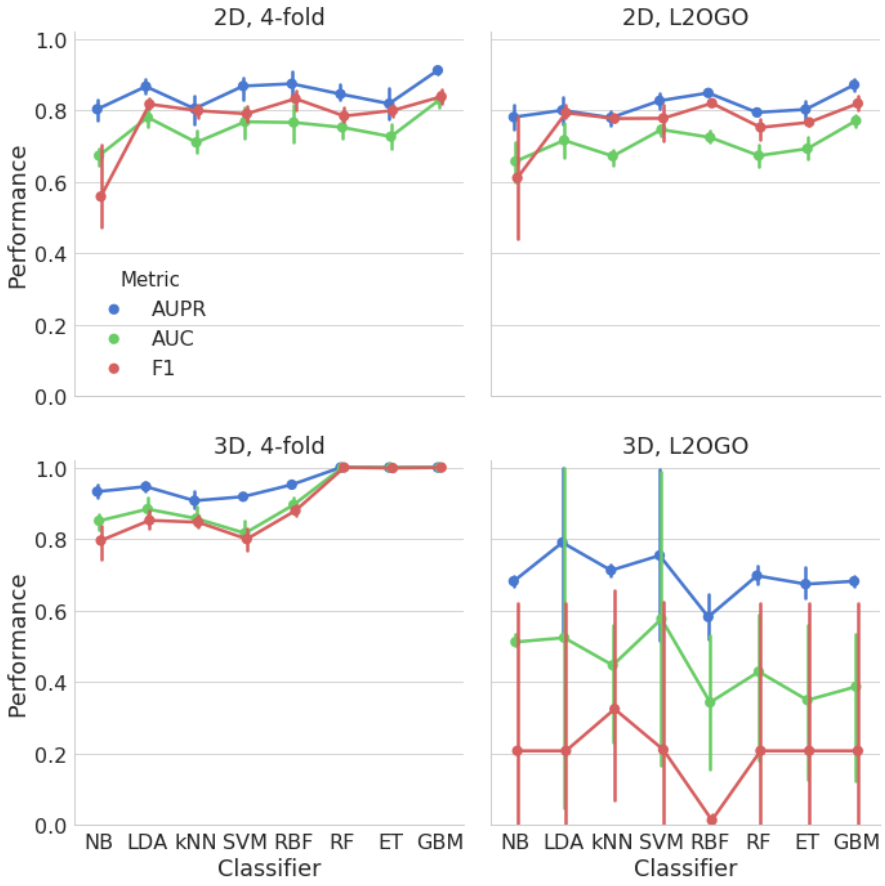
fibrillarin and EtBr. Nuclear morphometry extracted from binary masks seems to reflect most of the relevant changes. Increased potential for batch effects is only observed in classification of nucleolar structures in channels c1 and c2. Interestingly, combining 3 channels together seems to alleviate this issue and lead to near-perfect performance in L2OGO scheme.

The evaluation as presented here has a number of drawbacks and requires further investigation. First, we only use flattened vectors of pixels, while there exist multiple methods for texture feature extraction, which may speed up the calculation. Alternatively, deep learning-based methods can be used for automatic feature learning (*Ching et al.*, 2018; *Kalinin et al.*, 2018b). Second, we only evaluate performance on 2D maximum intensity projections of 3D images. Bigger study could further address similar issues in the original 3D space. Finally, we assume each nucleus in the same image to be representative of the phenotypic label that is provided for the whole image. This can be addressed by using methods that are robust to label noise (*Kalinin and Lisitsin*, 2011).

2.5.4 PC3 voxel-based morphometric analysis

After the exclusion of nuclei without detected nucleoli, the segmented PC3 collection consists of 458 nuclear (310 EPI and 148 EMT) and 1,101 nucleolar (649 EPI and 452 EMT) binary masks.

Figure 2.7: Voxel-based PC3 morphometry classification



Notes. The comparison of cross-validation strategies and commonly used algorithms to evaluate the classification performance and possible batch effects using combined morphometric features of 2D and 3D fibroblast nuclear and nucleolar binary masks.

After merging smaller EMT groups, L2OGO scheme produced 4 pairs of groups as training and testing sets. Given smaller number of volumes and apparent class imbalance, we compared L2OGO to 4-fold CV repeated 2 times to match the total number of splits. Similar to the previous experiment, 2D morphometry classification performance was quite similar for both CV schemes, see figure 2.7. However, in 3D,

the performance of algorithms degraded as measured by L2OGO CV, such that no methods performed better than in 2D. This can indicate possible batch effects, given the perfect performance estimates for 3 classifiers on 2D features. However, it is hard to judge given the large performance metrics' variation in 3D. In this case, the best classification by single classifier was the result of applying the Gradient Boosting classifier (GB) with the median $AUC = 0.774 \pm 0.017$, $AUPR = 0.875 \pm 0.019$, $F1 = 0.818 \pm 0.018$.

Results of classification on both collections suggest that the combination of the voxel-based morphometry and common algorithms with default parameters can provide a good baseline performance. Best performance is typically achieved by the application either of Gaussian SVM or tree-based ensemble algorithms, such as Gradient Boosting. This can be explained by the ability of such models to capture complex non-linear relationships in data (*Gao et al.*, 2018; *Tang et al.*, 2018), for example, between nuclear volume and surface area, without manually introducing them as individual features. We controlled for over-fitting using cross-validation, which showed that standard deviations of performance metrics were not increasing substantially when using non-linear models. Using 3D masks can improve the performance as it did in fibroblast classification. However, it suggests that having the three-dimensional information sometimes can lead to more apparent batch effects and, thus, may require more complex validation schemes.

2.6 Concluding remarks

A lack of publicly available 3D cell imaging datasets limits the evaluation of various 3D cell and nuclear morphology analysis solutions. To address this limitation, we present a new dataset that consists of two collections of 3D volumetric microscopic images. Each collection includes images of cells in two phenotypic states and, thus, poses a binary classification problem that can be used for the assessment of cell

nuclear and nucleolar morphometry analysis methods. We share these data publicly to promote results reproducibility, facilitate open-scientific development, and enable collaborative validation. To the best of our knowledge, this 3D imaging dataset is one of the largest publicly available datasets of its type.

In order to establish baseline evaluation of simple voxel-based morphometric analysis methods, we provide an example of 3D image processing workflow: from segmentation, to feature extraction, to morphometric analysis. First, we use both model-based and machine learning segmentation methods to extract individual nuclear and nucleolar binary masks in 3D. Then, we extract commonly used 2D and 3D voxel-based measures of binary mask morphology and combine them into per-nucleus feature vectors. We compare a number of commonly used machine learning classification algorithms on both collections of data using voxel-based morphometric measures. To account for batch effects, while enabling calculations of AUC and AUPR performance metrics, we also propose a specific cross-validation scheme (L2OGO). Our results indicate potential usefulness of 3D cell imaging data for morphology analysis. However, they also indicate the possibility of stronger batch effects compared to the 2D setting, which may be related to the different in imaging resolution in Z dimension.

As a possible limitation of this work, the microscope settings did not meet the Nyquist sample rates and may have created distortions in the digitized images (*Cole et al.*, 2011). However, sampling was consistent across experiments. Larger variability of the performance estimates in 3D using the suggested CV scheme (L2OGO) may be reduced by better class balancing or loss weighting during the each iteration of the cross-validation process. Although produced nuclear and nucleolar binary masks are visually inspected, they are produced by segmentation algorithms rather than hand-labeled by an expert. We provide an example of 3D image processing workflow, which, in general, does not have to always include segmentation (*Caicedo et al.*, 2017). The size of the produced 3D morphological dataset should be big enough

to use segmentation-free deep learning-based morphology analysis approaches (*Ching et al.*, 2018; *Kalinin et al.*, 2018b). Recent examples in medical image analysis have already demonstrated successful applications of such models in the small data regime (*Rakhlin et al.*, 2018; *Iglovikov et al.*, 2018; *Shvets et al.*, 2018b,a) and on devices with limited resources (*Solovyev et al.*, 2018). Finally, we assume each cell in the same image to be representative of the same phenotypic label that is provided on the level of the whole image. However, this assumption does not always hold. One 3D volumetric image can contain cells of multiple phenotypes. This can be addressed by using methods for weakly-supervised classification that are robust to label noise.

Imaging protocols, original and segmented data, and the source code are made publicly available on the project web-page: `http://www.socr.umich.edu/projects/3d-cell-morphometry/data.html`. Additionally, extracted morphometric features are made available for interactive exploration and analysis online via our visual analytics platform SOCRAT (*Kalinin et al.*, 2017).

CHAPTER III

3D Cell Nuclear Surface Morphometry

3.1 Introduction

The first part of this chapter describes 3D morphometry metrics for nuclear and nucleolar shape description and classification. First, surfaces of 3D masks extracted from the microscopy data are reconstructed using Laplace-Beltrami eigen-projection and topology-preserving boundary deformation (*Shi et al.*, 2010). Then, we computed intrinsic and extrinsic geometric metrics, which are used as derived signature vectors (shape biomarkers) to characterize the complexity of the 3D shapes and discriminate between observed clinical and phenotypic traits. These metrics include volume, surface area, mean curvature, curvedness, shape index, and fractal dimension (*Koenderink and Van Doorn*, 1992; *Thompson et al.*, 1996; *Meyer et al.*, 2003). Although these methods were previously used in recent neuroimaging studies (*Dinov et al.*, 2009; *Fani et al.*, 2013; *Moon et al.*, 2015), this was the first attempt to apply robust smooth LB-based surface reconstruction with intrinsic and extrinsic morphometric measure extraction to 3D cell nuclear and nucleolar shape modeling and morphometry (*Kalinin et al.*, 2018c). Suggested modeling and analysis methods are not restricted to nuclear and nucleolar shapes and can be used for the shape quantification of other cellular compartments, depending on their topology.

3.1.1 3D shape representation and morphometric measures

The way cell nuclear shapes can be measured depends on their representation extracted from image data (*Pincus and Theriot*, 2007). Many 3D morphometric measures are applied as is to 3D geometric objects represented by volumetric data (*Kalinin et al.*, 2018d). However, voxel-based shape representations are noisy, and they may lose fine geometric details or even break the objects topological structure. Moreover, these representations are not intrinsic, and vary when changing pose or deforming the object. A recent review of approaches to 3D cell shape description by *Dufour et al.* (2015) separated them into three categories in increasing order of complexity: landmark-based, graph-based, and moment-based. This last category includes approaches that are widely used in cellular morphology and allow the user to obtain a global representation that combines low-order moments describing the coarse conformation with high-order moments retaining information at higher frequency. Typically, before applying these methods, a binary mask or outline of the shape (surface) is first extracted from image data, which is done by most segmentation methods. These masks are assumed to have a sphere-like topology and can be projected onto an appropriate basis. Two popular approaches of this type are spherical harmonics (SPHARM) (*Brechbühler et al.*, 1995) and spherical wavelets (*Antoine and Vandergheynst*, 1999). Both methods first map the surface of interest onto the sphere using appropriate spherical parameterization techniques, and then project it onto a reference function basis living on the sphere.

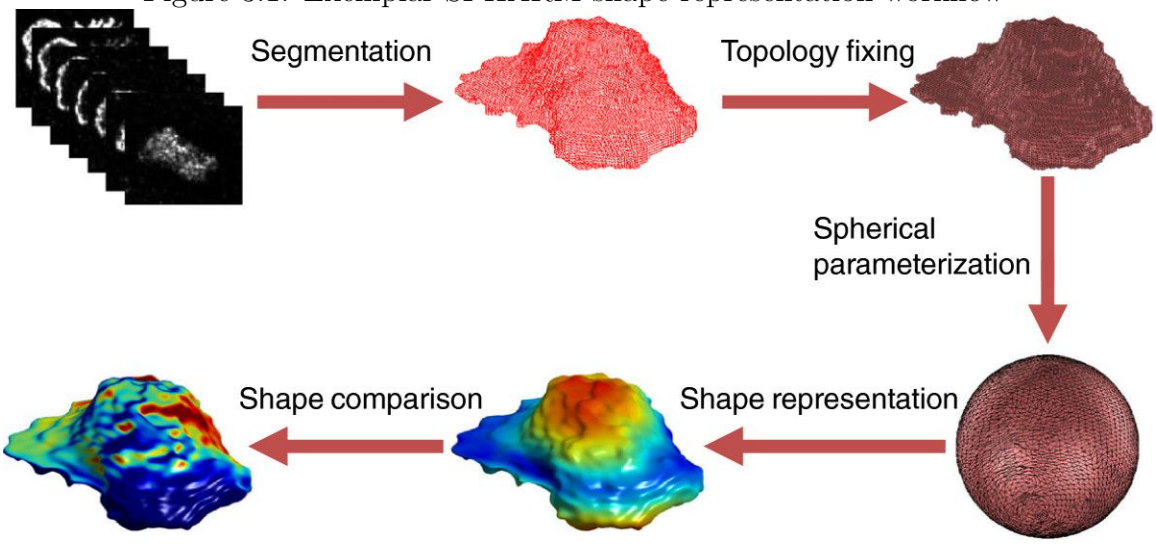
SPHARM is arguably one of the most widely applied cell morphology modeling approaches (*Khairy et al.*, 2008; *Singh et al.*, 2011; *Ducroz et al.*, 2012; *Du et al.*, 2013). In SPHARM, the spherical signal is projected onto a basis of Legendre polynomials, extending the classical Fourier analysis to signals on the two-sphere (*Ducroz et al.*, 2012). In the same way that vectors can be described through projections onto each axis (using scalar products), expansion coefficients (scalar product between func-

tions) can be used to describe functions. On the unit sphere, an orthonormal basis for the Hilbert space of square-integrable function is given by the spherical harmonics: $Y_l^m(\theta,\varphi) = k_{l,m}, P_l^m(\cos\theta)e^{im\varphi}$, where l and m are respectively the degree and order of the harmonic, $k_{l,m}$ is the expansion coefficient and P_l^m is the associated Legendre polynomial (*Ducroz et al.*, 2012). Using this basis, any spherical scalar function $f(\theta,\varphi)$ can be expanded into $f(\theta,\varphi) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l} \hat{f}(l,m)Y_l^m(\theta,\varphi)$, where $\hat{f}(l,m))$ is the (l,m) harmonic coefficient, given by: $\hat{f}(l,m) = k_{l,m}\int_0^{\pi}\int_0^{2\pi} e^{-im\varphi}f(\theta,\varphi)P_l^m(\cos\theta)\sin\theta d\varphi\, d\theta$. The coefficients $\hat{f}(l,m)$ are unique and can thus describe any arbitrary shape. The spectral decomposition of the input signal is then straightforward: lower degrees (i.e. l) correspond to low frequencies and hence describe the global shape of the object, while higher degrees describe the details of the surface. Higher dimensional (non-scalar) spherical functions can also be expanded using spherical harmonics, by expanding each component of the function independently. The spherical harmonic transform can be performed on a surface defined in the Cartesian space (x,y,z) and parametrized into a spherical signal defined in the polar system (θ,φ) as $v(\theta,\varphi) = (x(\theta,\varphi)y(\theta,\varphi)z(\theta,\varphi))^T$ (*Ducroz et al.*, 2012). As (θ,φ) runs over the sphere, $v(\theta,\varphi)$ runs over the object surface. By applying the SPHARM transform to each component of $v(\theta,\varphi)$ independently, coefficients with three components can be obtained (*Ducroz et al.*, 2012). SPHARM coefficients describe general conformation of the shape of interest at different spatial scales, are rotation invariant, and can be directly used as features for further analysis (*Shen and Makedon*, 2006; *Ducroz et al.*, 2012).

SPHARM coefficients can be computed by first performing surface reconstruction and spherical parametrization using the CALD algorithm (*Shen and Makedon*, 2006). Then, the object surface is expanded of into a complete set of spherical harmonic basis functions. Finally,the SHREC method (*Ducroz et al.*, 2012) is used to minimize the mean square distance between corresponding surface parts.

However, SPHARM methods are most appropriate when low order approxima-

Figure 3.1: Exemplar SPHARM shape representation workflow



Notes. Based on the topology fixed binary volume and the spherical parameterization result, spherical harmonics are employed to describe an object shape. After that, measurements of local and global dynamic cell shape changes can be conducted. Adapted from *Du et al.* (2013) under CC-BY 2.0 license.

tion is satisfactory and become less effective in preserving surface details, as artificial oscillations start to appear when higher order basis functions are incorporated (*Shi et al.*, 2010). The spherical parameterization introduces metric distortion which compounds the reconstruction error (*Seo and Chung*, 2011). More robust smooth surface reconstruction can be obtained from a 3D binary mask via Laplace-Beltrami (LB) eigen-projection (*Shi et al.*, 2010). On a unit sphere, the LB eigen-functions correspond to spherical harmonics, so overall they can be viewed as a generalization of the SPHARM to the complex geometry manifold with local adaptation of the basis to the dataset at hand (*Lévy*, 2006; *Seo and Chung*, 2011). Consider a closed compact manifold $\mathcal{M} \subset \mathbb{R}^3$. Let $L^2(\mathcal{M})$ be the space of square integrable functions on $\mathcal{M}$ with the inner product $\langle f, g \rangle_{\mathcal{M}} = \int_{\mathcal{M}} f(\mathbf{p}) g(\mathbf{p}) d\mu(\mathbf{p})$, where μ is the Lebesgue measure such that $\mu(\mathcal{M})$ is the total area of $\mathcal{M}$ (*Seo and Chung*, 2011). The orthonormal basis in $L^2(\mathcal{M})$ is given by the eigenfunctions of $\Delta_{\mathcal{M}} \psi_j = -\lambda \psi_j$, where $\Delta_{\mathcal{M}}$ is the LB-operator in $\mathcal{M}$ (*Lévy*, 2006; *Seo and Chung*, 2011). The eigen-functions $\psi_0, \psi_1, \psi_2, ...$ can be sorted by the corresponding eigenvalues, $0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq ....$ Then, for surface

coordinates $p = (p^1, p^2, p^3)'$, each coordinate function $p^i(\mathbf{p}) \in L^2(\mathcal{M})$ can be represented as a linear combination of the LB eigen-functions: $p^i(\mathbf{p}) = \sum_{j=0}^{K-1} \beta_j^i \psi_j(\mathbf{p})$, where $\beta_j^i = \langle p^i, \psi_j \rangle_{\mathcal{M}}$ are Fourier coefficients, and K is the number of basis functions (*Seo and Chung*, 2011). Laplace-Beltrami (LB) eigen-projection does not require the spherical parameterization and thus avoids the corresponding metric distortion during reconstruction. It has been also suggested that the representation using the LB eigen-functions has far less between-subject reconstruction error variability and converges faster to the ground truth with fewer basis functions than SPHARM (*Seo and Chung*, 2011). The proposed method has been demonstrated to produce smoother and more detailed surfaces compared to both the SPHARM and the topology preserving level sets (*Han et al.*, 2003). Extracted surfaces are smooth, accurately represent the shape of an object, and can be further used for morphometric analysis.

In order to extract shape geometric characteristics, boundary surfaces of binary masks are typically reconstructed from voxel data and discretized as meshes. At the next step, various useful morphometric descriptors can be computed based on this representation. Useful extrinsic and intrinsic geometric descriptors aim to distinguish between global and local shape features. Intrinsic measures capture shape properties that are invariant under transformations (e.g., affine: rotation, translation and scaling). Various shape morphometry measures, like surface area and Gaussian curvature, represent invariant metrics of complexity, which are stable under special transformations of the surface (e.g., bending) that do not affect the inner geometry of the boundary of the 3D volume (*Batchelor et al.*, 2002). Alternatively, shape metrics, e.g., mean L_2-norm and the extrinsic curvature index, are sensitive to affine transformation and other shape morphology in the ambient space. Shape index and curvedness are morphometric descriptors that can capture local shape features, independently or in relation to the size of an object (*Koenderink and Van Doorn*, 1992). Combination of the object surface reconstruction with the extraction of such shape

measures demonstrated high performance in recent neuroimaging studies for discriminatory morphometric analysis of complex 3D shapes of cortical and subcortical brain areas (*Dinov et al.*, 2009; *Fani et al.*, 2013; *Moon et al.*, 2015).

3.1.2 High-throughput processing workflow protocol

When it comes to a choice of tools for 3D cell nuclear morphometrics, reproducibility and implementation availability are among major concerns in the field of bioimage analysis (*Dufour et al.*, 2015). To date, many of the widely available software tools for cell shape morphometry were either developed for the analysis of 2D (*Pincus and Theriot*, 2007; *Ramo et al.*, 2009; *Held et al.*, 2010; *Pau et al.*, 2010; *Kamentsky et al.*, 2011; *Chiang et al.*, 2015), or pseudo-3D images (*Peng and Murphy*, 2011). Other tools only implement slice-by-slice or voxel-based morphometry (*Schindelin et al.*, 2012; *de Chaumont et al.*, 2012; *Kankaanpaa et al.*, 2012; *Ollion et al.*, 2013), providing a coarse approximation of the global cell shape that is sensitive to increasing amounts of noise and usually fails to characterize morphological variations occurring at different spatial scales. Other common limitations of many 3D cell morphology solutions include a lack of high-throughput processing capabilities or restrictions to the specific programming language or platform that dictate principles of tool implementation (*Eliceiri et al.*, 2012; *Peng et al.*, 2014; *Li et al.*, 2016). Implementations of methods in a bioimage analysis landscape are highly diverse. They range across programming languages, software libraries, and file formats, which increases module interoperability issues and makes code reuse extremely difficult. Re-implementing underlying methods is often very challenging, time-consuming, and error prone (*Ince et al.*, 2012). Some of the existing bioimage analysis frameworks, including ImageJ (*Schneider et al.*, 2012), rely on a plugin architecture, which allows their extension via third-party contributions (*Schindelin et al.*, 2012; *de Chaumont et al.*, 2012; *Ollion et al.*, 2013). High-throughput capabilities of some of these software packages

are limited to processing of multiple objects simultaneously within its graphical user interface (GUI), for example, Tango (*Ollion et al.*, 2013). More advanced packages, such as CellProfiler 2.0 (*Kamentsky et al.*, 2011), BioimageXD (*Kankaanpaa et al.*, 2012), and Icy (*de Chaumont et al.*, 2012), provide a basic graphical interface to assemble elementary tasks into reusable pipelines that make it possible to execute in GUI and batch modes. However, these solutions are still limited to specific scripting languages and libraries supported by the main software package. They also dont provide a straightforward way to take advantage of the growing number of parallel hardware configurations, such as clusters, clouds, and high-performance computing, which limits the scalability of these solutions.

An alternative to plugin-based solutions, software platforms with modular design allow integration of already existing solutions into workflows without re-implementing them in a specific language, and provide methods for optimizing module interaction, re-usage, and extension. An example of an extensive and feature rich solution for building and executing complex workflows is the LONI Pipeline (*Dinov et al.*, 2009, 2010). This client-server platform enables users to efficiently describe atomic modules and end-to-end protocols in a graphical canvas using a large library of powerful computational tools. The Pipeline back-end server has extensive support for parallel execution on a grid cluster, including automated data converting, formatting and transfer, optimal job submission and management, pausing execution, and combining local and remote software and data sources. Most importantly, it makes it very easy to create new custom modules from any software that supports a command line interface (CLI). The Pipeline allows users to take advantage of a highly diverse set of tools and connect them together as steps of a computational protocol that is then executed in a high-throughput, parallel fashion. Validated individual modules and end-to-end workflows may be saved, reused in other workflows, easily modified and repurposed. Additionally, the LONI Pipeline saves information about executed steps

(such as software origin, version, and architecture) providing provenance information (*Dinov et al.*, 2010, 2011).

We develop a reproducible pipeline workflow implementing the entire process that can be customized and expanded for deep exploration of associations between 3D nuclear and nucleolar shape phenotypes in health and disease. High-throughput imaging (HTI) can include automatization of liquid handling, microscopy-based image acquisition, image processing, and statistical data analysis (*Pegoraro and Misteli*, 2016). Our work focuses on the last two aspects of this definition. We implemented a streamlined multi-step protocol using a diverse set of tools to achieve optimal performance compared to alternatives at each step of analysis. These tools are represented as individual modules seamlessly connected in the LONI Pipeline workflow. This workflow meets modern standards for high-throughput imaging processing and analysis and is mostly automated with a focus on validity and reproducibility. Our implementation is massively parallel, customizable, and provides fully automated execution and data provenance out-of-the-box. At the final step of the workflow, we employed machine learning methods to investigate the associations between cell phenotypes and treatment conditions using cell shape morphometric measures as features.

3.1.3 Visual analytics for morphometric data analysis

Data visualization and analytics are crucial components of any study of complex biomedical data (*Dinov*, 2016). The goal of visual analytics (VA) is to support analytical reasoning and decision making with a combination of highly interactive visualizations and data analysis techniques. This includes data management, computational transformation, hypothesis testing, and knowledge discovery (*Keim et al.*). Visual analytics workflow encompasses an iterative process in which data analysts interactively interrogate their datasets. Visual analytics workflow encompasses an iterative process in which an analyst interrogates the dataset in hand in the form of

interactive dialogue motivated by an analytical question and supported by visualizations and data analysis components.

The rapid advances in web-based information, communication and computation technologies support the explosive growth of interactive services and tools implementing novel solutions for exploration and visualization of large, complex, incongruent, multisource and incomplete data. Web-based visualization solutions dramatically reduce deployment issues by running directly in the desktop or mobile web browser, yielding a high degree of accessibility and avoiding complex installation, version update and incompatibility, and other problems characteristic of standalone software (*Steed et al.*, 2014). Together with increased graphical and interactive capabilities native to modern web browsers (HTML5/JavaScript), this enabled enhancement of visualizations by user-focused data-driven real-time interactions.

Furthermore, integrating statistical and machine learning methods with visualizations can greatly amplify visual data analysis approaches (*Tukey*, 1977). Similarly, improved computational capabilities of web browsers enabled implementations of mathematical, statistical, machine learning, and computing JavaScript libraries (*Khan et al.*, 2014). Combining these resources with existing interactive visualization frameworks would open a path to the development of more effective and powerful VA web-based systems without re-implementing standard components from scratch. However, there are a number of challenges associated with this endeavor such as significant incongruences in design, development, and deployment. Earlier review of information visualization system architectures pointed out the difficulty of identifying common design patterns within existing visualization tools, and consequently the high cost for users to learn and evaluate unfamiliar systems (*Heer and Agrawala*, 2006). Moreover, building general purpose visual analytics web systems is even more challenging than creating visualization tools, since VA application design requires their uniform integration with data management and analysis solutions into a com-

plex large-scale web application. Existing VA applications implement this approach combining web-based interactive visualization libraries with specific analytical functionality, however, such solutions remain scattered and very problem-specific. Thus, practices for development of sophisticated large-scale VA web applications are not well established (*Booth et al.*, 2014).

Our motivation to address these challenges comes in part from over a decade of designing, building, and maintaining the Statistics Online Computational Resource (SOCR) (*Dinov*, 2006). SOCR includes a large collection of web applications for in-browser data processing, analysis, and visualization. SOCR implemented a feature-rich educational web toolkit for in-browser interactive data visualization, modeling, and statistical analysis (*Dinov*, 2006). Visualization components were based on open-source Java charting library JFreeChart (*Object Refinery Limited*, 2017), and included SOCR Charts (*Dinov*, 2006), Hyperbolic Wheel (*Lam and Dinov*, 2012), and Motion Charts (*Al-Aziz et al.*, 2010) that allowed computation of data summary statistics and provided a number data representation types, including raw data display, data mapping, and over 30 various highly interactive data plots, charts and diagrams, including 3D, spatial, cartographic, and GIS data visualizations. SOCR Modeler (*Dinov*, 2006) implemented interactive visual model fitting, including distribution-mixture-modeling and generalized-expectation-maximization implemented in the setting of 2D point clustering and classification. SOCR Analyses component (*Chu et al.*, 2009) provided hypothesis testing of both parametric and nonparametric models, data modeling (linear regression and ANOVA), and computation of power and sample size. SOCR Analyses was implemented using model-view-controller (MVC) software design pattern, allowing to decouple interactive visual representation from modeling techniques, such that the latter could be used separately as external computational library and easily extended. The SOCR Distributome project (*Dinov et al.*, 2016) addressed complementary computational modeling applications from the viewpoint of probabil-

ity distributions, including tools for simulation, analysis and inference, model-fitting, examination of the analytical, mathematical and computational properties of specific distributions, and exploration of the inter-distributional relations.

The suite of SOCR tools (*Dinov et al.*, 2008; *Dinov and Christou*, 2009) has been proven over time to successfully realize visual analytics workflows, in which an analyst interrogates the dataset in hand in the form of interactive dialogue motivated by an analytical question and supported by visualizations and data analysis components. Examples implementing such VA workflow include California ozone pollution case study using SOCR Charts and Analysis (*Dinov and Christou*, 2011) and visual analysis of big medicare, labor, census and econometric data with interactive SOCR Data Dashboard (*Husain et al.*, 2015). The original SOCR Java applets were open-source and generally scalable, but did not build on a common platform to enable component interoperability, resource sharing, and runtime interaction. SOCR infrastructure realizes a suite of web tools providing many features important for VA workflows, but became disconnected and hard to maintain due to the lack of common infrastructure. Moreover, most of SOCR applets along with other in-browser Java based visualization tools are becoming unsupported by major web browsers, which either removed or announced timelines for the removal of standards based plugin support, eliminating the ability to embed Flash, Silverlight, Java and other plugin based technologies, in part as a response to numerous vulnerability reports (*US-CERT*, 2013).

This experience shows limitations of the earlier Java-based SOCR tools, designs and implementations, which inhibits our efforts to further expand and maintain them. We introduce SOCRAT–a web-based scalable platform for in-browser interactive data analysis and visualization (*Kalinin et al.*, 2017). It relies on principles of multi-level modularity with central module control, re-usage, extension, and optimized interaction to address these challenges. This design broke down functional part of an integrated VA system into atomic functional parts and proposed a way to compose

them such that the whole system is flexible, extensible, and robust.

We show how combining multiple interactive visualization and analytical tools in SOCRAT allows employing custom in-house developed solutions together with third-party libraries to perform quick and easy exploration of extracted 3D morphometry measures. SOCRAT is used to investigate structures and patterns in the extracted morphometric dataset, from loading the feature tables into the toolbox, to wrangling missing and/or incongruent data values, to building interactive plots and using statistical and machine learning analysis tools.

3.1.4 Deep learning for 3D morphology classification

Deep learning describes a class of machine learning algorithms that are capable of combining raw inputs into layers of intermediate features (*Ching et al.*, 2018). While biomedical applications of deep learning are still emerging, they have already shown promising advances over the prior state-of-the-art in several tasks. Deep learning methods have transformed the analysis of natural images and video, and similar examples are beginning to emerge with cell images. We employ deep learning as an additional approach to 3D cell nuclear morphology classification.

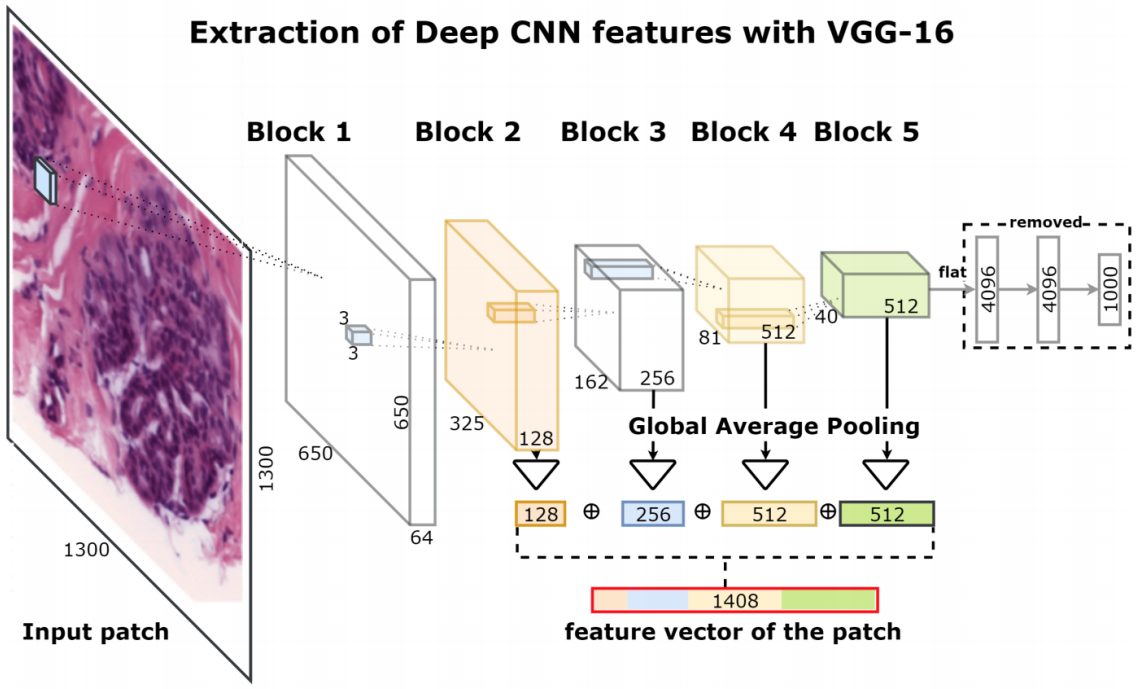
Conventional machine learning algorithms used in sections above are typically limited in their ability to process raw cell imaging data. Their performance heavily depends on the extraction of relevant morphological representations or morphometric features that require careful engineering and considerable domain expertise. Overall, limitations of conventional machine learning methods include the need for extensive human guidance, painstaking feature handcrafting, careful data preprocessing and the dimensionality reduction to achieve top performance. In contrast, deep learning methods model data by learning high-level representations with multilayer computational models such as artificial neural networks (ANNs) (*LeCun et al.*, 2015). While classic feed-forward artificial neural networks might serve as drop-in replacement for

other machine learning models and require input preprocessing and feature extraction, deep learning architectures, such as convolutional neural networks (CNNs), allow the algorithm to automatically learn features from raw and noisy data. Deep neural networks rely on algorithms that optimize feature engineering processes to provide the classifier with relevant information that maximizes its performance with respect to the final task. Such deep learning models can be thought of as automated feature learning or feature detection, which facilitates learning of hierarchical, increasingly abstract representations of high-dimensional heterogeneous data (*LeCun et al.*, 2015; *Kalinin et al.*, 2018b), also known as representation learning.

CNNs specifically have achieved great performance improvements in various computer vision tasks. CNNs are designed to process data that come in the form of multiple arrays, for example a colour image composed of three 2D arrays containing pixel intensities in the three colour channels. Many data modalities are in the form of multiple arrays: 1D for signals and sequences; 2D for images or audio spectrograms; and 3D for video or volumetric images. There are four key ideas behind CNN that take advantage of the properties of natural signals: local connections, shared weights, pooling, and the use of many layers (*LeCun et al.*, 2015). The architecture of a typical CNN is structured as a series of stages, 3.2. The first few stages are composed of two types of layers: convolutional layers and pooling layers. Units in a convolutional layer are organized in feature maps, within which each unit is connected to local patches in the feature maps of the previous layer through a set of weights called a filter bank. The result of this local weighted sum is then passed through a non-linearity such as a ReLU. All units in a feature map share the same filter bank. Different feature maps in a layer use different filter banks. The reason for this architecture is two-fold. First, in array data such as images, local groups of values are often highly correlated, forming distinctive local motifs that are easily detected. Second, the local statistics of images and other signals are invariant to location. In other words, if a motif can appear in

one part of the image, it could appear anywhere, hence the idea of units at different locations sharing the same weights and detecting the same pattern in different parts of the array. Mathematically, the filtering operation performed by a feature map is a discrete convolution, hence the name (*LeCun et al.*, 2015). One of the most popular deep CNN architectures is a 16-layer network, named VGG-16 (*Simonyan and Zisserman*, 2014). It consists of a number of convolutional blocks, each representing a stack of convolutional layers that employ filters with a very small receptive field (e.g., 3×3). VGG-16 has been applied to many tasks in biomedical image analysis (*Rakhlin et al.*, 2018; *Iglovikov et al.*, 2018; *Shvets et al.*, 2018b,a). Figure 3.2 demonstrates an example of applying VGG-16 to multi-scale feature extraction from pathology images.

Figure 3.2: Exemplar VGG-16 architecture



Notes. Each convolutional block consists of teo convolutional layers with a non-linearity, such as ReLU, and batch normalization (*Ioffe and Szegedy*, 2015) in between. Using global average pooling, it is possible to extract features from various levels of the network, capturing both low- and high-level representations. Adapted from *Rakhlin et al.* (2018) under CC-BY-ND 4.0 license.

Deep learning models have already been applied to 2D cell morphological profiling for both drug discovery (*Caicedo et al.*, 2018) and disease diagnostics (*Doan et al.*,

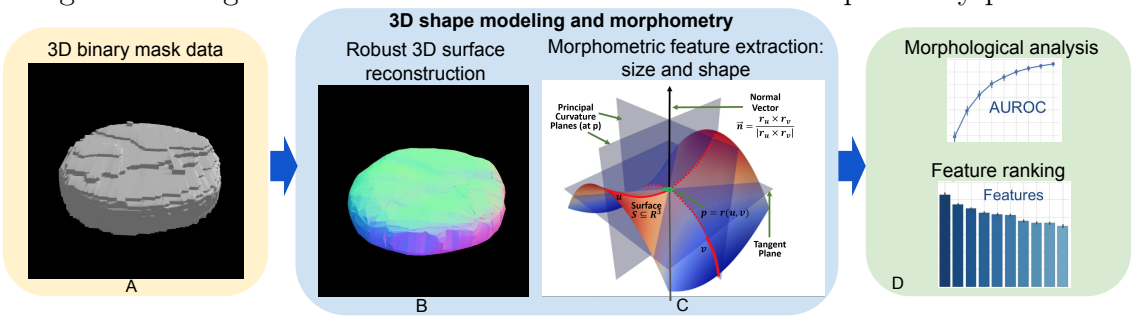
2018; *Uhler and Shivashankar*, 2018). However, efficiently scaling deep learning is challenging, and there is a high computational cost (e.g. time, memory and energy) associated with training neural networks and using them to make predictions (*Ching et al.*, 2018). Thus, directly applying deep learning models to 3D cell imaging data is inefficient due to the so called curse of dimensionality: the number of points on the grid grows exponentially with its dimensionality. In such scenarios, it becomes increasingly important to exploit data sparsity whenever possible in order to reduce the computational resources needed for data processing. Traditional convolutional network implementations are optimized for data that lives on densely populated grids, and cannot process sparse data efficiently. Alternatively, convolutional network implementations that are tailored to work efficiently on sparse data can alleviate that limitation (*Graham*, 2015).

Sparse CNNs can be thought of as an extension of the idea of sparse matrices (*Graham*, 2014, 2015). If a large matrix only has small number of non-zero entries per row and per column, then it makes sense to use a special data structure to store the non-zero entries and their locations; this can both dramatically reduce memory requirements and speed up operations such as matrix multiplication. However, if 10% of the entries are non-zero, then the advantages of sparsity may be outweighed by the efficiency which which dense matrix multiplication can be carried out (*Graham*, 2015) . Since as our object representation we propose surface reconstruction, the resulting meshes can be used to render highly sparse discrete nuclear shape voxel representations that can benefit from using sparse CNNs. We show how sparse 3D CNNs can be used for for accurate and efficient cell nuclear morphology classification and reach performance comparable to hand-crafted morphometric feature-based approaches.

3.2 3D surface morphometry

Figure 3.3 shows a high-level view of the end-to-end protocol. We start with a dataset of 3D binary nuclear and nucleolar masks. We modeled 3D nuclear and nucleolar boundaries by their surface reconstruction and extracted the derived morphometry measures. Finally, we computed statistical differences, identified shape morphometry-phenotype associations, and evaluated the results.

Figure 3.3: High-level schematic flow of the 3D surface morphometry protocol



Notes. Figure panels show: (A) 3D binary mask data; (B) mathematical representation and modeling of shape and size; (C) calculation of derived intrinsic and extrinsic geometric measures; and (D) machine learning based classification, feature ranking, and analysis.

3.2.1 Robust smooth surface reconstruction

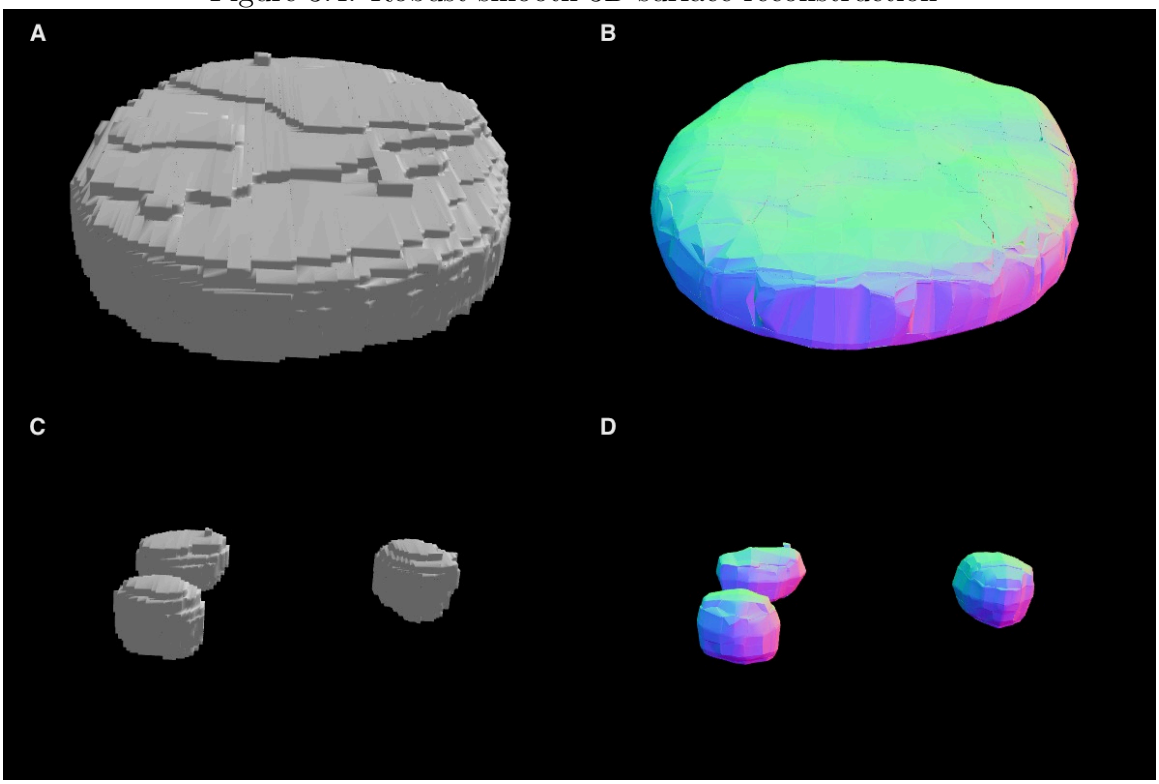
To model the 3D shape of cell nuclei and nucleoli, boundaries of their 3D masks extracted from the microscopy data are modeled as genus zero two-dimensional manifolds (homeomorphic to a 2-sphere S^2) (*Ferri and Gagliardi*, 1982) that are embedded as triangulated surfaces in $\mathbb{R}^3$, see figure 3.3. Our approach uses an iterative Laplace-Beltrami eigen-projection and a topology-preserving boundary deformation algorithm (*Shi et al.*, 2010). This algorithm performs robust reconstruction of the objects surfaces from their segmented masks using an iterative mask filtering process. First, a mesh representation is constructed from the boundary of an objects binary mask. Then, the boundary is projected onto the subspace of its Laplace-Beltrami eigen-functions (*Lévy*, 2006), which allows the algorithm to automatically locate the

position of spurious features by computing the metric distortion in eigen-projection. LB eigen-functions are intrinsically defined and can be easily computed from the boundary surface with no need of any parameterizations. They are also isometry invariant, and thus robust to the jagged nature of the boundary surface, which is desirable for biomedical shape analysis (*Niethammer et al.*, 2007). As previously shown in *Shi et al.* (2010), the discretized LB spectrum captures intrinsic shape characteristics (e.g., global shape transformations will preserve the spectral signature). The magnitude of the eigenvalues of the LB operator intuitively corresponds to the frequency in Fourier analysis, thus it provides a convenient mechanism to control the smoothness of the reconstructed surface. Using this information, the second step is a mask deformation process that only removes the spurious features while keeping the rest of the mask intact, thus preventing unintended volume shrinkage. This deformation is topology-preserving and well-composed such that the boundary surface of the mask is a manifold. The last two steps iterate until convergence and the method generates the final surface as the eigen-projection of the mask boundary, which is a smooth surface with genus zero topology (*Shi et al.*, 2010). These properties allow application of this algorithm to any shape, including, for example, crescent-shaped, multi-lobed, and folded, as long as shape topology is homeomorphic to a sphere. The exemplar results of this step performed on nuclear and nucleolar masks are shown in figure 3.4.

3.2.2 Morphometric feature extraction

In this study, we used six shape measures as features quantifying geometric characteristics of the 3D surfaces, see figure 3.5. To calculate these measures, first the principal (min and max) curvatures ($k_1 \leq k_2$) were computed using triangulated surface models representing the boundaries of genus zero solids *Terzopoulos* (1988). Then, shape morphometry measures can be expressed in terms of principal curvatures: mean

Figure 3.4: Robust smooth 3D surface reconstruction

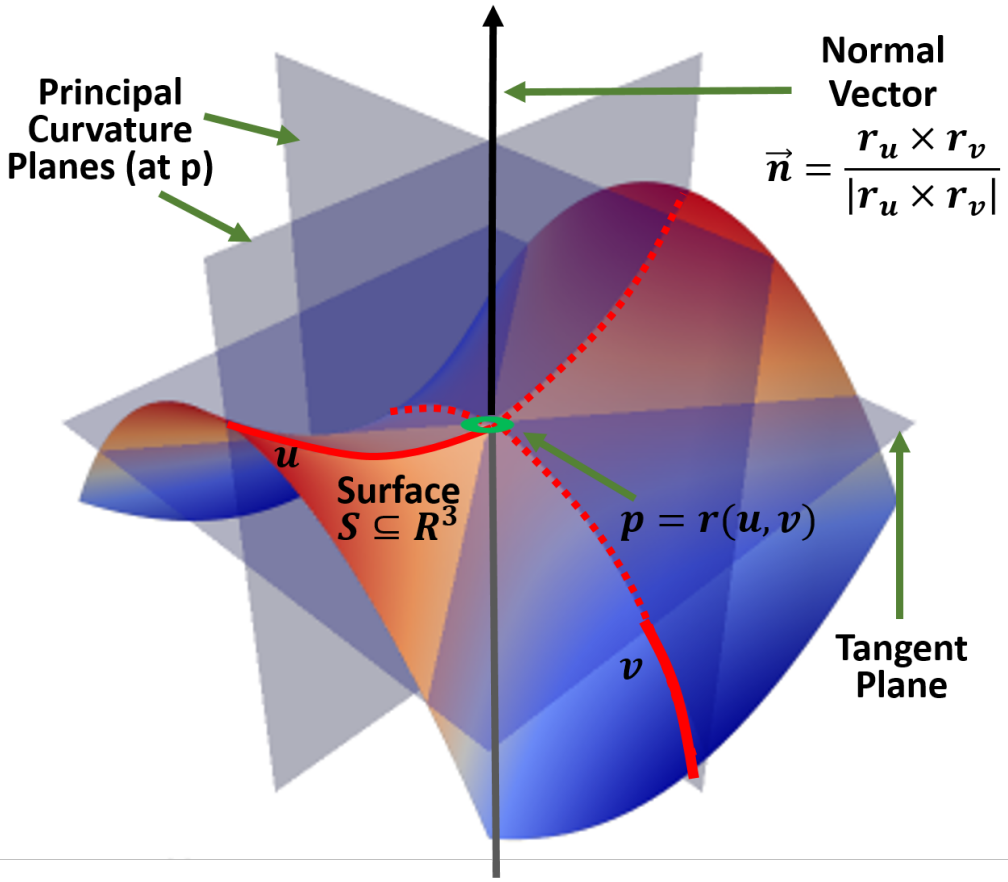


Notes. 3D visualization of: (A) a binary mask representation of a nucleus segmented from a Fibroblast cell image; (B) a mesh representation of a reconstructed smooth surface of a nucleus; (C) three binary masks for nucleoli segmented within this nucleus; and (D) three mesh representations of nucleolar surfaces, color-coded along the Z axis. Visualizations are produced with the SOCR Dynamic Visualization Toolkit web application (*SOCR*, 2018).

curvature as $MC = \frac{k_1 + k_2}{2}$, shape index as $SI = \frac{2}{\pi} arctan(\frac{k_1 + k_2}{k_2 - k_1})$, and curvedness as $CV = \sqrt{\frac{k_1^2 + k_2^2}{2}}$. The principal curvatures of a surface are the eigenvalues of the Hessian matrix (second fundamental form), which solve for k, $|H - kI| = 0$, where I is the identity matrix. If S is a surface with second fundamental form $H(X, Y)$, $p \in M$ is a fixed point, and we denote an orthonormal basis u, v of tangent vectors at p, then the principal curvatures are the eigenvalues of the symmetric Hessian matrix, $H = \begin{smallmatrix} H_{u,u} & H_{u,v} \\ H_{v,u} & H_{v,v} \end{smallmatrix} = H_{u,u}\partial u^2 + 2H_{u,v}\partial u \partial v + H_{v,v}\partial v^2$, a.k.a. shape tensor. Let $r = r(u, v)$ be a parameterization of the surface $S \subseteq \mathbb{R}^3$, representing a smooth vector valued function of two variables with partial derivatives with respect to u and v denoted by

r_u and r_v, see figure 3.5. Then, the Hessian coefficients $H_{i,j}$ at a given point (p) in the parametric u, v-plane are given by the projections of the second partial derivatives of r at that point onto the normal to S, $n = \dfrac{r_u \times r_v}{|r_u \times r_v|}$, and can be computed using the dot product operator: $H_{u,u} = r_{u,u} \cdot n$, $H_{u,v} = H_{v,u} = r_{u,v} \cdot n$, $H_{v,v} = r_{v,v} \cdot n$, see figure 3.5.

Figure 3.5: The (local) geometry of 2-manifolds



Notes. Per vertex definitions of curvature, relative to a local coordinate framework.

Volume is the amount of 3D space enclosed by a closed boundary surface and can be expressed as $V = \iiint_{\mathbb{R}^3} I_D(x, y, z) dx dy dz$, where $I_D(x, y, z)$ represents the indicator function of the region of interest (D) (*Larson and Edwards*, 2009). If $r(u, v)$ is a continuously differentiable function and the normal vector to the surface over the appropriate region D in the parametric u, v plane is denoted by $\overrightarrow{r_u} \times \overrightarrow{r_v}$, then S_Ω :

$r = r(u, v),\ (u, v) \in \Omega$, is the parametric surface representation of the region boundary (*Santal*, 2004). Then surface area can be expressed as $SA = \iint_\Omega |\overrightarrow{r_u} \times \overrightarrow{r_v}| dudv$. The fractal dimension calculations are based on the fractal scaling down ratio, ρ, and the number of replacement parts, N (*Mandelbrot*, 1982). Accurate, discrete approximations of these metrics are used to compute them on mesh-represented surfaces (*Ferri and Gagliardi*, 1982; *Jagannathan*, 2005). These discrete metrics were first introduced as a part of the shape analysis protocol (*Dinov et al.*, 2009) and were further applied in neuroimaging studies (*Fani et al.*, 2013; *Moon et al.*, 2015). Mathematical formulas and intuitive descriptions of size and shape measures are also given in table B.1 and table B.2, correspondingly.

The extracted 3D morphometric measures serve as features for training a number of machine learning algorithms in order to assess classification performance, see figure 3.3. The number of detected nucleoli per nucleus is included as an individual feature. We merged nucleoli-level features within each nucleus by computing sample statistics (e.g., average, minimum, maximum, and higher moments) for each morphometry measure as described in *Kalinin et al.* (2018d). These statistics are used to augment the signature feature vectors of the corresponding parent nuclei such that all feature vectors are of the same length. Correspondingly, nuclei that do not have any automatically detected internally positioned nucleoli were excluded from further analysis, such that for each nucleus there was at least one nucleolus.

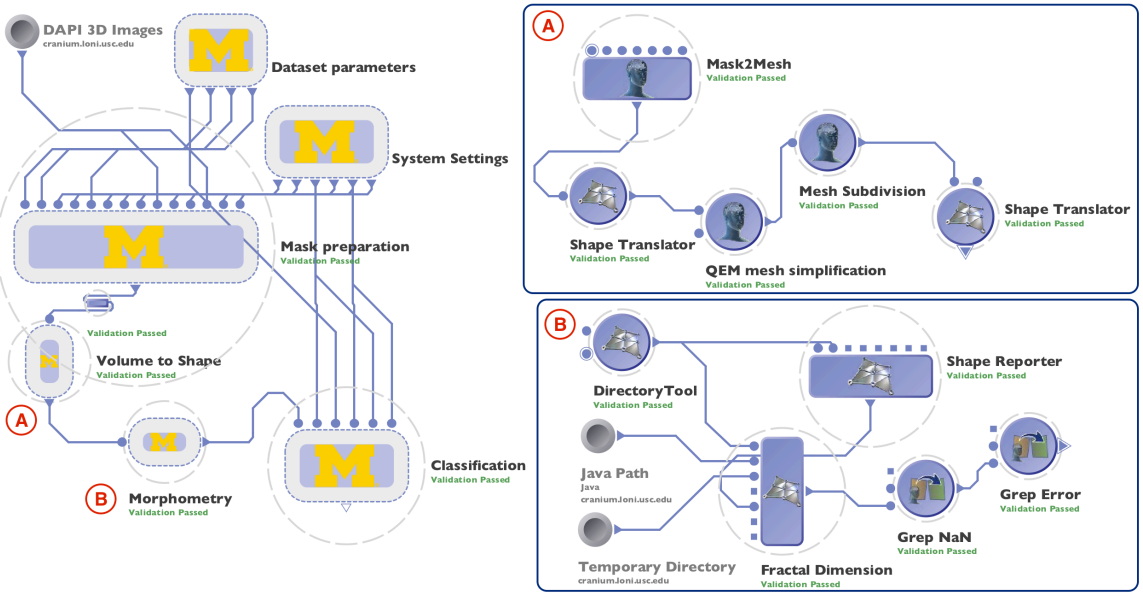
3.2.3 High-throughput workflow protocol

While the LONI Pipeline is a popular tool in neuroimaging and bioinformatics, it has not yet been utilized as widely by the bioimage analysis community. In this work, we implement a streamlined multi-step protocol that relies on a diverse set of tools and solutions seamlessly connected in the LONI Pipeline workflow, see figure 3.6. From a high-level perspective, every step of data processing and analysis protocol

is wrapped as an individual module in the workflow that provides input and output specifications that allow the Pipeline to automatically connect and manage atomic modules. The modular structure of our implementation makes it highly flexible and not limited to specific tools included in the workflow. It can be repurposed for a wide range of different experiments by adjusting parameters, adding, removing, or replacing individual modules, while preserving high-throughput capabilities, as presented in the Discussion section. Every module represents an independent component that can be used in a stand-alone fashion. As a result, a distributed, massively parallel implementation of our protocol makes it possible to easily process thousands of nuclei and nucleoli simultaneously. The workflow does not depend on the total number of 3D objects, biological conditions, or a number of running instances since its execution is completely automated once the workflow configuration is fixed, including job scheduling and resource allocation. During the execution, our workflow provides a researcher with real-time information about progress and allows the viewing of intermediate results at every individual step. In addition, failed modules may easily be restarted.

The workflow is configured in such a way that it can consume data in the specific format we used, i.e. $1024 \times 1024 \times Z$ 3D volumes in different channels as 16-bit 3D TIFF files. Each volume is processed independently, in parallel fashion, such that workflow automatically defines how many processes are needed to analyze all of the input data. 3D shape modeling and morphometric feature extraction are performed on individual masks independently, which allows us to simultaneously run up to 1,200 jobs on the cluster during our experiments, effectively reducing the computing time. Finally, the workflow collects morphometry information from each individual mask and combines them in the results table that is further used as an input to classification algorithm. These capabilities allow the user to take advantage of modern computational resources, lift the burden of low-level configuration from researchers,

Figure 3.6: Morphometry graphical workflow in the LONI Pipeline client



Notes. Screenshots of the exemplar graphical workflow in the LONI Pipeline client interface that include: (left) overview of the validated workflow protocol showing nested groups of modules; (A) expanded Volume to Shape group that includes modules that perform 3D shape modeling refinement; and (B) expanded Morphometry group that includes a module that performs morphological measure extraction.

make it easier to control the execution process, and improve reproducibility of the whole process.

3.3 Visual analytics with SOCRAT

We perform exploratory visual analysis of extracted morphometric features using SOCRAT, a web platform for interactive visual analytics. SOCRAT serves as a flexible platform for building powerful VA applications by providing a convenient way to seamlessly integrate custom and third-party components. In order to achieve that, we proposed a loosely coupled and centrally controlled platform architecture with modular structure (*Kalinin et al.*, 2017).

3.3.1 SOCRAT architecture

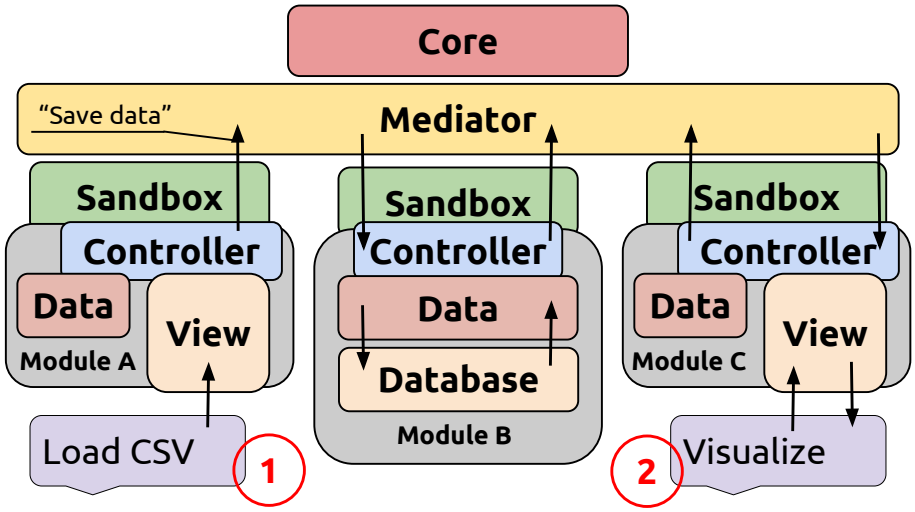
The proposed system is represented by a loosely coupled architecture (*Osmani*, 2011) with functionality broken down into independent modules with, ideally, no inter-module dependencies. Modules are single-purpose parts of a system with limited permissions. They are interchangeable in the sense that the system is capable of supporting, adding, removing or replacing modules without the rest of the modules in the system failing, which facilitates flexibility and robustness. In this decoupled setting, modules do not directly communicate with each other. Instead, they provide a means to communicate with the Core module. The Core module is a central control piece responsible for the initialization and internal communications of other modules, and for satisfying the interoperability requirement, similar to the Kernel module in (*Bender et al.*, 2000). If necessary, it performs module runtime validation and monitoring. Extensibility is implemented in the form of plug-in support, which enables wrapping third-party components as modules, providing them with a standard for other modules API.

The starting point for designing modular architecture was actually to break it down into independent functional pieces and then define their responsibilities. For web systems, it's common to design modules following MVC-like patterns (MV*), separating the data from the display, and organizing their interaction with the medium component. For VA application architecture, it's natural to initially separate visualization from data analytics and data storage, although, different views, on the contrary, should be made possible to combine. Thus, interactivity is more difficult to decouple into separate modules. Two simple options to approach this limitation are either to: (1) provide basic interactivity specifications within general intermediate display layer, that would be accessible by different modules, or (2) define more specific interactions individually in the scope of each module. The second option allows for more flexibility in terms of implementation, including easier third-party component

integration.

In suggested architecture, we do not impose strict module classification. Instead, we suggest that there are a few loosely defined types of modules: for example, modules that perform specific actions on data, such as calculations without implementing the UI (background modules) or modules that use data to populate interactive visualization specifications (visual modules). Modules with new visualization capabilities should be able to implement high-level predefined specifications as well as low-level control definitions, depending on the task. To combine such variety of approaches, the module's components should be modular as well. Intra-module component structure should allow for simple ways to circulate information within a module. For inter-module communication, module inner components can be exposed by opening access to them for other modules via Core, for example, providing calculations-as-a-service or visualization-as-a-service. In practice, however, it is reasonable to expect that a typical module of a web VA system will be a hybrid of these types.

Figure 3.7: General modular SOCRAT architecture



Notes. Human-computer and inter-modular interactions (via Sandbox-Mediator pattern) are shown by arrows. From left to right: (1) user uploads CSV file using module A, which broadcasts "Save data" message; Core module redirects the message to module B that saves the data into database; (2) then user requests visualization of data in module C, which requests data from module B, receives the data, and displays it in the view component.

Upon module initialization by the Core module, all modules are provided with Sandbox, an instance representing Facade software design pattern (*Osmani*, 2012), that hides inner Core structure from the module behind high-level messaging interface. This interface, in turn, is represented by a Mediator pattern (*Osmani*, 2012) that prevents modules from directly referring to each other and instead acts as an intermediary. Core can use Mediator to start, stop, and restart individual modules selectively in the runtime, without breaking the application. It is also responsible for answering module's request. For example, if a visual module requested specific data transformation that was outsourced to another module, which is currently not available, Core will use Mediator to provide visual module with negative result, such that it can display an appropriate error message or placeholder and/or try again later.

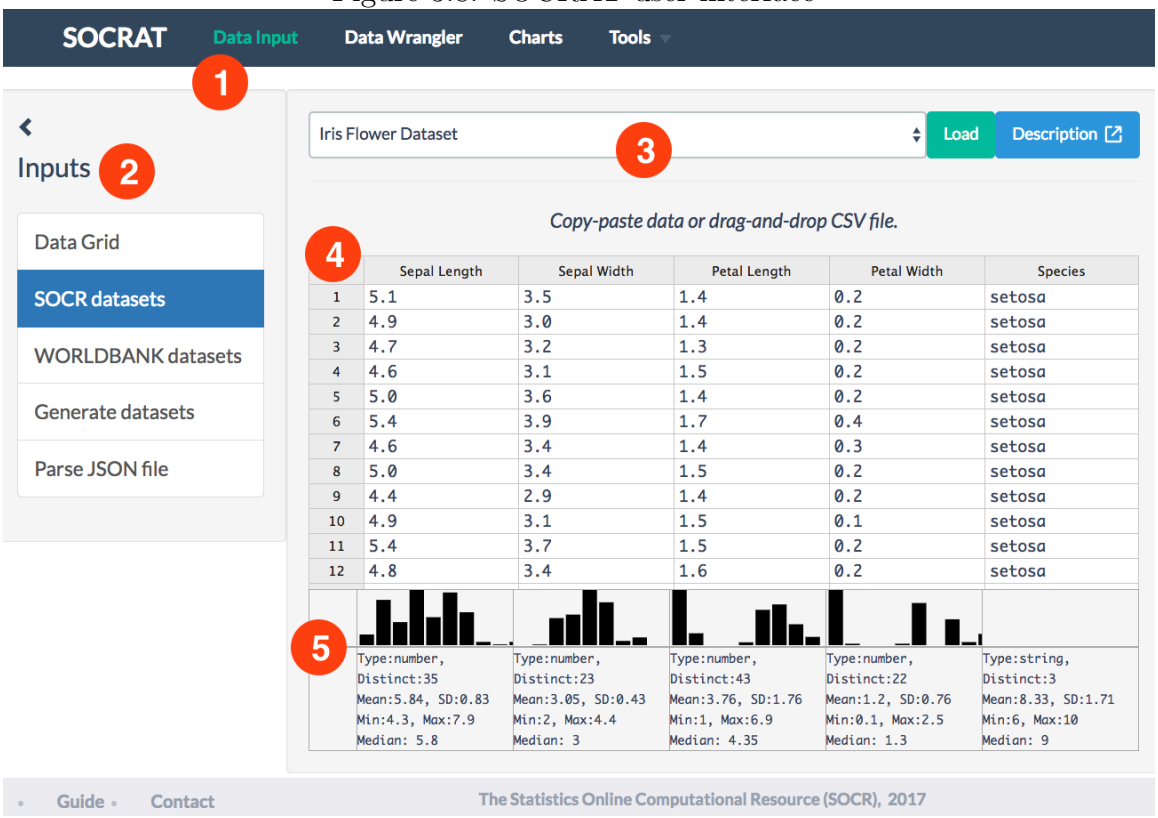
3.3.2 SOCRAT user interface

Similar to the UI designs of Tableau (formerly Polaris) *Stolte et al.* (2002), Voyager and PoleStar *Wongsuphasawat et al.* (2016, 2017), SOCRAT user interface (UI) consists of three top-level elements: main menu, sidebar and the central area, see figure 3.8. Main menu is used to provide access to core UI components. It is indicated in declarative fashion in every module definition along with its subcomponents.

On the applications start, SOCRAT Core recursively parses the module list and automatically adds links to all UI modules into main menu, while registering their URL in the routing scheme. This takes the menu-building burden from the developer, and allows to organize the main menu in many configurations, including nested dropdown sub-menus.

Sidebar is a common UI pattern that is typically used for auxiliary control placement. Both the main area and hideable sidebar are automatically initialized for every SOCRAT module, along with the methods for their interaction. Sidebar typically implements requests for current datasets and displays some of their properties, while

Figure 3.8: SOCRAT user interface



Notes. Overview of SOCRAT UI: (1) main menu indicating the currently active module, (2) sidebar with various data sources, (3) central panel includes module-specific data display and manipulation controls, (4) for example, dynamic, editable spreadsheet-like data grid that contains raw data values view and also allows to drag-and-drop CSV/TSV file to load the data, and (5) summary information panel below the data grid shows histogram and reports summary statistics for for each variable in the dataset.

central area can be used as a view for particular visualization specifications. This is the only UI restriction that SOCRAT imposes onto modules with UI; the developer can choose any visualization library in combination with any analytical methods that will be further used to build the application.

3.3.3 SOCRAT analytical capabilities

As a toolbox SOCRAT provide a wide range of VA capabilities, from data input, storage, management, and wrangling, to interactive visualizations, statistical analysis, and machine learning.

Spreadsheet-like live editor is used to display raw data values in a scrollable grid with dynamic loading of content, which allows the analyst to briefly glaze over the values in familiar Excel-like manner. Additionally, SOCRAT implements a module that wraps third-party data utility library Datalib (*Wongsuphasawat et al.*, 2016). Upon data loading we use Datalib for tabular data parsing, column type inference, and summary statistic calculation, including histogram generation. SOCRAT displays per-column summary statistics and histogram of values above each corresponding column of the dataset to improve efficiency of initial data exploration, see figure 3.8.

Data Wrangler (*Kandel et al.*, 2011; *Guo et al.*, 2011) allows highly interactive in-browser data cleaning and transformation supported by analytics and visualizations. It couples a mixed-initiative user interface with an underlying declarative transformation language. We created a module for data wrangling that embeds Wrangler into SOCRAT interface. As a result, Wrangler is represented as a separate SOCRAT UI module and declaratively specified in the SOCRAT platform configuration to appear in the dynamically built main menu.

Interactive visualizations in SOCRAT are based on D^3 (*Bostock et al.*, 2011) and vega-lite (*Satyanarayan et al.*, 2017) libraries. They provide over 30 various types of easily customizable chart configurations for in-depth analysis of multivariate tabular data based on histograms, scatter plots, line, area, bar, bubble, and pie charts. In order to assess the variability of extracted morphometry data, we include t-Distributed Stochastic Neighbor Embedding (t-SNE) (*Maaten and Hinton*, 2008) visualizations of the feature space generated by SOCRAT. Below we provides a formulation of t-SNE based on *Dinov* (2018).

The t-SNE technique represents a recent machine learning strategy for nonlinear dimensionality reduction that is useful for embedding (e.g., scatter-plotting) of high-dimensional data into lower-dimensional (1D, 2D, 3D) spaces. For each object (point in the high-dimensional space), the method models similar objects using nearby and

dissimilar objects using remote distant objects. The two steps in t-SNE include (1) construction of a probability distribution over pairs of the original high-dimensional objects where similar objects have a high probability of being paired and correspondingly, dissimilar objects have a small probability of being selected; and (2) defining a similar probability distribution over the points in the derived low-dimensional embedding minimizing the Kullback-Leibler divergence between the high- and low-dimensional distributions relative to the locations of the objects in the embedding map (*Dinov*, 2018). Either Euclidean or non-Euclidean distance measures between objects may be used as similarity metrics.

Suppose we have high dimensional data (ND): $x_1, x_2, ..., x_N$. In *step 1*, for each pair (x_i, x_j), t-SNE estimates the probabilities $p_{i,j}$ that are proportional to their corresponding similarities, $p_{j|i}$:

$$p_{j|i} = \frac{\exp\left(\frac{-||x_i - x_j||^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(\frac{-||x_i - x_k||^2}{2\sigma_i^2}\right)}.$$

The similarity between x_j and x_i may be thought of as the conditional probability, $p_{j|i}$. That is, assuming ND Gaussian distributions centered at each point x_i, neighbors are selected based on a probability distribution (proportion of their probability density), which represents the chance that x_i may select x_j as its neighbor, $p_{i,j} = \frac{p_{j|i} + p_{i|j}}{2N}$. The perplexity (*perp*) of a discrete probability distribution, p, is defined as an exponential function of the entropy, $H(p)$, over all discrete events: $perp(x) = 2^{H(p)} = 2^{-\sum_x p(x) \log_2 p(x)}$. t-SNE performs a binary search for the value σ_i that produces a predefined value *perp*. The simple interpretation of the perplexity at a data point x_i, $2^{H(p_i)}$, is as a smooth measure of the effective number of points in the x_i neighborhood. The performance of t-SNE may vary with the perplexity value, which is typically specified by the user, e.g., between $5 \leq perp \leq 50$. Then, the precision (variance, σ_i) of the local Gaussian kernels may be chosen to ensure

that the perplexity of the conditional distribution equals a specified perplexity. This allows adapting the kernel bandwidth to the sample data density – smaller σ_i values are fitted in denser areas of the sample data space, and correspondingly, larger σ_i are fitted in sparser areas. A particular value of σ_i yields a probability distribution, p_i, over all of other data points, which has an increasing entropy as σ_i increases. t-SNE learns a mapping $f : \{x_1, x_2, ..., x_N\} \longrightarrow \{y_1, y_2, ..., y_d\}$, where $x_i \in R^N$ and $y_i \in R^d$ ($N \gg d$) that resembles closely the original similarities, $p_{i,j}$ and represents the derived similarities, $q_{i,j}$ between pairs of embedded points y_i, y_j, defined by:

$$q_{i,j} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq i}(1 + ||y_i - y_k||^2)^{-1}}.$$

The *t-distributed* reference in t-SNE refers to the heavy-tailed Student-t distribution ($t_{df=1}$) which coincides with Cauchy distribution, $f(z) = \frac{1}{1+z^2}$ (*Dinov et al.*, 2016). It is used to model and measure similarities between closer points in the embedded low-dimensional space, as well as dissimilarities of objects that map far apart in the embedded space (*Dinov*, 2018). The rationale for using Student t distribution for mapping the points is based on the fact that the volume of an ND ball of radius r, B^N, is proportional to r^N. Specifically, $V_N(r) = \frac{\pi^{\frac{N}{2}}}{\Gamma\left(\frac{N}{2}+1\right)} r^N$, where Γ is the Euler's gamma function (*Wikipedia*, 2018), which is an extension of the factorial function to non-integer arguments. For large N, when we select uniformly random points inside B^N, most points will be expected to be close to the ball surface (boundary), S^{N-1}, and few will be expected near the B^N center, as half the volume of B^N is included in the hyper-area inside B^N and outside a ball of radius $r_1 = \frac{1}{\sqrt[N]{2}} \times r \sim r$. For example with $N = 2$, $\{x \in R^2 | \ ||x|| \leq r\}$ is representing a disk in a 2D plane (*Dinov*, 2018). When reducing the dimensionality of a dataset, if we used the Gaussian distribution for the mapping embedding into the lower dimensional space, there will be a distortion of the distribution of the distances between neighboring objects. This is

simply because the distribution of the distances is much different between the original (high-dimensional) and a the map-transformed low-dimensional spaces. t-SNE tries to (approximately) preserve the distances in the two spaces to avoid imbalances that may lead to biases due to excessive attraction-repulsion forces. Using Student t distribution $df = 1$ (aka Cauchy distribution) for mapping the points preserves (to some extent) the distance similarity distribution, because of the heavier tails of t compared to the Gaussian distribution. For a given similarity between a pair of data points, the two corresponding map points will need to be much further apart in order for their similarity to match the data similarity (*Dinov*, 2018). A minimization process with respect to the objects y_i using gradient descent of a (non-symmetric) objective function, Kullback-Leibler divergence between the distributions Q and P , is used to determine the object locations y_i in the map, i.e.,

$$KL(P||Q) = \sum_{i \neq j} p_{i,j} \log \frac{p_{i,j}}{q_{i,j}}.$$

The minimization of the KL objective function by gradient descent may be analytically represented by:

$$\frac{\partial KL(P||Q)}{\partial y_i} = \sum_j (p_{i,j} - q_{i,j}) f(|x_i - x_j|) u_{i,j},$$

where $f(z) = \frac{z}{1+z^2}$ and $u_{i,j}$ is a unit vector from y_j to y_i. This gradient represents the aggregate sum of all spring forces applied to map point x_i. This optimization leads to an embedding mapping that "preserves" the object (data point) similarities of the original high-dimensional inputs into the lower dimensional space. Note that the data similarity matrix $(p_{i,j})$ is fixed, whereas its counterpart, the map similarity matrix $(q_{i,j})$ depends on the embedding map. Of course, we want these two distance matrices to be as close as possible, implying that similar data points in the original space yield similar map-points in the reduced dimension (*Dinov*, 2018). Examples

of t-SNE applications to real biomedical data are available in *Dinov* (2018) and in SOCRAT *Kalinin et al.* (2017).

Additionally, SOCRAT analytical tools provide a number of tools for exploration of tabular data:

- statistical tests, such as t-test, Wilcoxon rank sum test, Kruskal-Wallis test, Friedman's test, ANOVA, ANCOVA, Generalized Linear Models, Contingency tables, Friedman's test and Fisher's exact test;
- regression analysis, including linear and logistic regression models;
- power and sample size analysis;
- regression analysis, including linear and logistic regression models power and sample size analysis;
- interactive clustering, including k-Means and spectral clustering algorithms
- dimensionality reduction with PCA decomposition and t-SNE algorithm with 2D and 3D interactive visualizations

3.4 Sparse 3D convolutional neural networks

Each layer of a CNN consists of a finite graph, with a vector of input/hidden units at each site. For regular two dimensional CNNs, the graphs are square grids. The convolutional filters are square-shaped too, and they move over the underlying graph with two degrees of freedom (*Graham*, 2015). Similarly, 3D CNNs are normally defined on cubic grids. The convolutional filters are cube-shaped, and they move with three degrees of freedom. In the interests of efficiency, we consider CNNs with a different family of underlying graphs (*Graham*, 2015). In 2D, we can build CNNs based on triangles. For each layer, the underlying graph is a triangular grid, and the

convolutional filters are triangular, moving with two degree of freedom. In 3D, we can use a tetrahedral grid and tetrahedral filters that move with three degrees of freedom (*Graham*, 2015). 3D tetrahedral sparse CNN is implemented by tweaking sparse CNN algorithm from *Graham* (2014) to work efficiently on general lattices. We employ 3D sparse CNN for nuclear morphology classification. Specifically, we use a CNN with the tetrahedral grid and tetrahedral filters that allow for a noticeable speed up when compared to cubic grid-based CNNs, although they may be less accurate at the small scales (*Graham*, 2015). We applied data augmentations during training using affine transformations. VGG-16 (*Simonyan and Zisserman*, 2014), shown in figure 3.2, was chosen as a network architecture, as it has shown to be a powerful deep learning model in many applications to biomedical image analysis (*Ching et al.*, 2018; *Rakhlin et al.*, 2018; *Iglovikov et al.*, 2018). We use 3D robust surface reconstruction (*Shi et al.*, 2010) to obtain mesh-based nuclear shape representations that are used as inputs to the network. At the input each shape is discretized in a tetrahedral grid and 10% of the training set is used for validation. Network is then trained for 500 epochs with stochastic gradient descent with momentum (*Ruder*, 2016).

3.5 Concluding remarks

In this chapter, we proposed a solution for 3D modeling, morphological feature extraction, analysis, and classification of cells by treatment conditions. Compared to other studies using 2D projections, this approach operates natively in 3D space and takes advantage of extrinsic and intrinsic morphometric measures that are more representative of the real, underlying nuclear and nucleolar geometry and allow easy human interpretation. Given the limitations of using 3D voxels for accurate shape representation, we employed 3D surface models to extract more informative size and shape measures to improve the morphology classification performance.

Our computational protocol implementation is highly parallel with throughput,

limited only by the number of available computing nodes, and it can process thousands of objects simultaneously with minimal human intervention. This pipeline workflow integrates a number of open-source tools for different steps of data processing and analytics. Every module in our workflow represents an individual component that can be easily modified, removed, or replaced by an alternative. Such modular software platform architectures have been shown to enable high reusability and ease of modification. This allows the user to use the same workflow or customize and expand it (e.g., specification of new datasets, swapping of specific atomic modules) for other purposes that require the analysis of a diverse array of cellular, nuclear, or other studies. The live demo available via the LONI Pipeline demonstrates the simplicity of use and high efficiency of parallel data processing. LONI also provides guest access (see Supplementary Information) and an opportunity to utilize a 4,500-core LONI cluster after applying for a collaboration account. Our computational approach is scalable and capable of processing complex big 3D imaging data, and is not limited to nuclear and nucleolar shapes. With some changes, it can be applied to other cellular and nuclear compartments of interest. More specifically, the robust smooth surface reconstruction algorithm can be directly applied to any 3D shapes, as long as their topology is sphere-like.

As one of the approach limitations, we pointed out that other geometric measures can be used to characterize shapes of interest, such as intrinsic shape context, compactness, symmetry, smoothness, convexity, etc. In the current representation, analyzable shapes are limited to genus zero surfaces, which is a fair assumption when modeling objects like nuclei or nucleoli. However, it might be not trivial when considering other nuclear structures, for example, chromosome territories or interchromosomal loops, since their topologies may not be homeomorphic to a sphere, or may not appear to be genus zero under some imaging conditions and modalities. It is also conceivable, yet not very likely for the discretized Laplace-Beltrami (LB) operator,

that 2 different shapes may have the same spectra. In this case, the algorithm may fail to detect the intrinsic differences between them due to false-negative error. Even though our workflow only requires little intervention (classifier selection and tuning), further improvements would involve adaptive implementations with even less manual intervention, as well as extraction of additional features. For example, 3D textural features could possibly increase discriminatory power of the method and provide more information on chromatin reorganization (*Kalinin et al.*, 2018a). Since nuclear deformation serves as a proxy to underlying processes, the importance of particular features and the methods ability to classify nuclei does not provide direct insight into the fundamental biological mechanism driving the observed morphometric differences between cell phenotypes or environmental conditions. The computational results should be further tested and externally validated using other experimental conditions and prospective data.

The goal of visual analytics is to support analytical reasoning and decision making with a combination of highly interactive visualizations and data analysis techniques. SOCRAT implements a visual analytics workflow that encompasses an iterative process, in which data analysts can interactively interrogate extracted morphometric measures in the form of interactive dialogue supported by visualizations and data analysis components. For example, in order to assess the variability of extracted morphometry data, SOCRAT includes such visualization methods as t-Distributed Stochastic Neighbor Embedding (t-SNE) (*Maaten and Hinton*, 2008) of the morphological feature space. It also can be used to demonstrate interactions between various morphometric features in order to assess their relationships. Finally, we demonstrated the ability to visualize volumetric images and extracted meshes online via SOCR Dynamic Visualization Toolkit web application (*SOCR*, 2018).

In general, correct classification of every single cell (type, stage, treatment, etc.) is a challenging task due to significant population heterogeneity of the observed cell

phenotypes. For example, the same sample may contain a close mixture of intertwined cancerous and non-cancerous cell phenotypes; or, both classes may include apoptotic cells exhibiting similar shapes or sizes. Given the nature of cell samples, culturing, preparation, and collection, we have considered classification of cell sets rather than single cells. The idea of classifying sets of cells, rather than individual samples, is not new and has been used in recent biomedical image classification studies (*Huang et al.*, 2014a; *Cheplygina et al.*, 2015). The rationale behind this is based upon the observation that even if an algorithm misclassifies a few cells in a sample, the final (cell set) label will still be assigned correctly, as long as majority of cells are classified correctly. Using this strategy, we performed classification on small groups of cells, ranging from 3 to 31 cells per set. During each fold of the internal cross-validation, these small cell sets were randomized by bootstrapping procedure with 1,000 repetitions. Random uniform sub-sampling was used to resolve the sample-size imbalance between the classes. Due to the possible presence of batch effects in data, we employed L2OGO cross-validation scheme.

The source code for image processing and derived data are made publicly available on the project web-page: `http://www.socr.umich.edu/projects/3d-cell-morphometry/data.html`. Additionally, extracted morphometric features are made available for interactive exploration and analysis online via our visual analytics platform SOCRAT (*Kalinin et al.*, 2017).

CHAPTER IV

Applications of 3D Nuclear Surface Morphological Analysis

4.1 Introduction

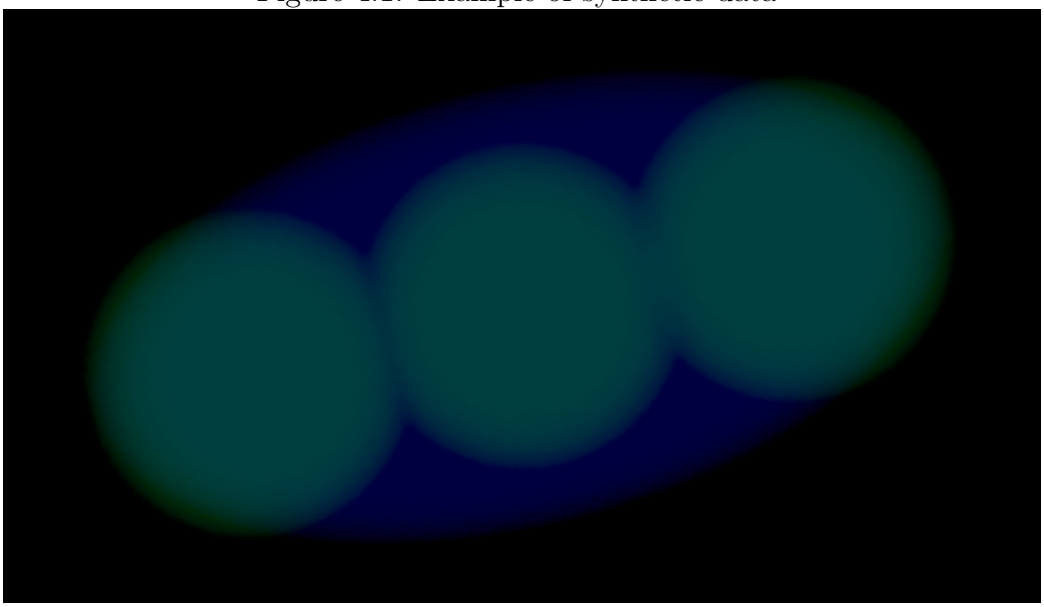
In this chapter we consider applications of 3D surface morphology modeling, morphometrics, visual analytics and deep learning to cell nuclear morphology classification and analysis. First, we validate accuracy of the proposed 3D surface morphometry method using synthetically generated data of various shapes. Then, we compare the use of nuclear morphometry using 3D robust surface reconstruction with spherical harmonics as features for morphological classification of fibroblast nuclei using a number of common machine learning algorithms. Sparse 3D CNN classification performance is also tested on fibroblast nuclear morphology classification. Then, we use 3D surface morphometry and SOCRAT visual analytics for more detailed analysis of both fibroblast and PC3 nuclear and nucleolar data. Finally, we introduce a new experiment in which we use 3D morphometry to observe morphological changes in astrocyte cells treated with a chromatin remodeling chemical compound.

4.2 Validation on synthetic data

To validate the shape morphometry metrics, we first applied them to synthetically generated 3D binary masks. We used the scikit-image Python library (*van der Walt et al.*, 2014) to create 3D solids representing cubes, octahedra, spheres, ellipsoids, and 3 overlapping spheres with linearly aligned centers , see figure 4.1. We processed these objects and compare the resulting shape morphometry measures. Specifically, we aimed to confirm the expected close relation between the analytically derived measures of volume and surface area computed using the corresponding shape parameters (e.g., radius, size), and their computationally derived counterparts reported by the processing pipeline workflow. Our results illustrate that for nucleus-like shapes, e.g., spheres and ellipsoids, the computational error is within 2%. For faceted objects, e.g., cubes and octahedrons, the calculation error is within 6%. The increased error in the latter case can be explained by the mesh smoothing the surface reconstruction algorithm applies at the shape vertices to resolve points of singularity (e.g., smooth but non-differentiable surface boundaries).

To demonstrate the detection of shape differences between different types of 3D objects, we also compared overlapping spheres against circumscribed ellipsoids. As expected, the average mean curvature and curvedness measures are lower and shape index values are higher for spheres compared to ellipsoids. We observed a similar trend when comparing changes in these shape morphometry measures for spheres, ellipsoids, and overlapping spheres. For example, average mean curvature and curvedness were highest for overlapping spheres and lowest for spheres, which is expected based on definitions of these measures. This simulation confirms our ability to accurately measure size and shape characteristics of 3D objects, which forms the basis for machine-learning based object classification based on boundary shapes. Exemplar results of synthetic data morphometry are available in table 4.1.

Figure 4.1: Example of synthetic data



Notes. Overlapping spheres, or beads, (R=30, overlap=5) within an ellipsoid (a=80, b=40, c=40) used to validate shape morphometry measure calculations and differences in morphometric features.

Primitive	AMC	SA	Volume	Curvedness	SI	FD
Cube	0.029	144438.190	4060563.200	0.323	0.012	2.181
Octahedron	0.036	42441.207	687556.250	0.349	0.017	2.016
Sphere	0.015	79633.500	2111214.800	0.713	0.009	2.095
Ellipsoid	0.026	33973.812	526245.700	0.633	0.013	2.097
Beads	0.029	29279.334	322649.560	0.612	0.035	2.147

Table 4.1: Morphometry of synthetic objects. Cube (a=160), octahedron and sphere (R=80). AMC–average mean curvature, SA–surface area, SI–shape index, FD–fractal dimensionality.

4.3 Fibroblast nuclear surface morphometry analysis

4.3.1 Comparison with SPHARM and sparse 3D CNN

Classification of single cell nuclear morphology from the fibroblast collection may be assessed using shape morphometry metrics as salient discriminatory features, which we compare against their corresponding SPHARM coefficients (*Dufour et al.*, 2015; *Ducroz et al.*, 2012). We used fibroblast binary nuclear masks to calculate both SPHARM and morphometric features.

Classifier	SPHARM, mean AUC ($\pm SD$)	Morphometry, mean AUC ($\pm SD$)
KNN	0.556($\pm$0.103)	0.629($\pm$0.204)
SVM	0.593($\pm$0.165)	0.677($\pm$0.354)
RBF	0.513($\pm$0.145)	0.682($\pm$0.264)
RF	0.619($\pm$0.175)	0.645($\pm$0.200)
AD	0.612($\pm$0.246)	0.663($\pm$0.252)
GBM	0.620($\pm$0.234)	0.674($\pm$0.229)

Table 4.2: Comparison of SPHARM coefficients and 3D surface morphometry descriptors for single cell fibroblast nuclei classification. KNN–k nearest nighbors, SVM–support vector machine with linear kernel, RBF–support vector machine with Gaussian kernel, RF–Random Forest, AD–AdaBoost, GBM–gradient boosting machines.

We used the popular SPHARM-MAT toolbox (*Shen*, 2010) with default parameters to compute SPHARM shape description coefficients as described in *Ducroz et al.* (2012) and used as feature vectors for classification.

Then we used machine learning classification methods on derived feature vectors with default parameters for each method. Performance was compared using the L2OGO cross-validation scheme and the area under the receiver operating characteristic curve as a performance metric. As shown in table 4.2, 3D shape morphometric measures not only demonstrate comparable discriminative performance to SPHARM coefficients, but outperform them using all tested algorithms.

Using the same fibroblast nuclear surface representations that were used to extract morphometric measures we test the performance of the sparse 3D tetrahedral CNN. At the input of the network, each 3D nuclear shape mesh is discretized in a tetrahedral grid with the size of 212×212×212. We train VGG-16-like network with the rate of 260 sec/epoch using a single NVIDIA GeForce GTX TITAN X Maxwell 12GB GPU. After 500 epochs the model reaches 2.96% error rate on training set and 21.12% error rate on the validation set (10% of original dataset). Although sparse 3D CNN outperforms both SPHARM and 3D morphometry classification, further investigation is required, due to the differences in cross-validation, parameter tuning, etc. The error rate on

the training set may indicate over-fitting, which may require more sophisticated data augmentation strategies (*Buslaev et al.*, 2018). Nevertheless, this result does show the potential for efficient and accurate applications of deep learning models for 3D morphological classification.

4.3.2 3D surface morphological classification

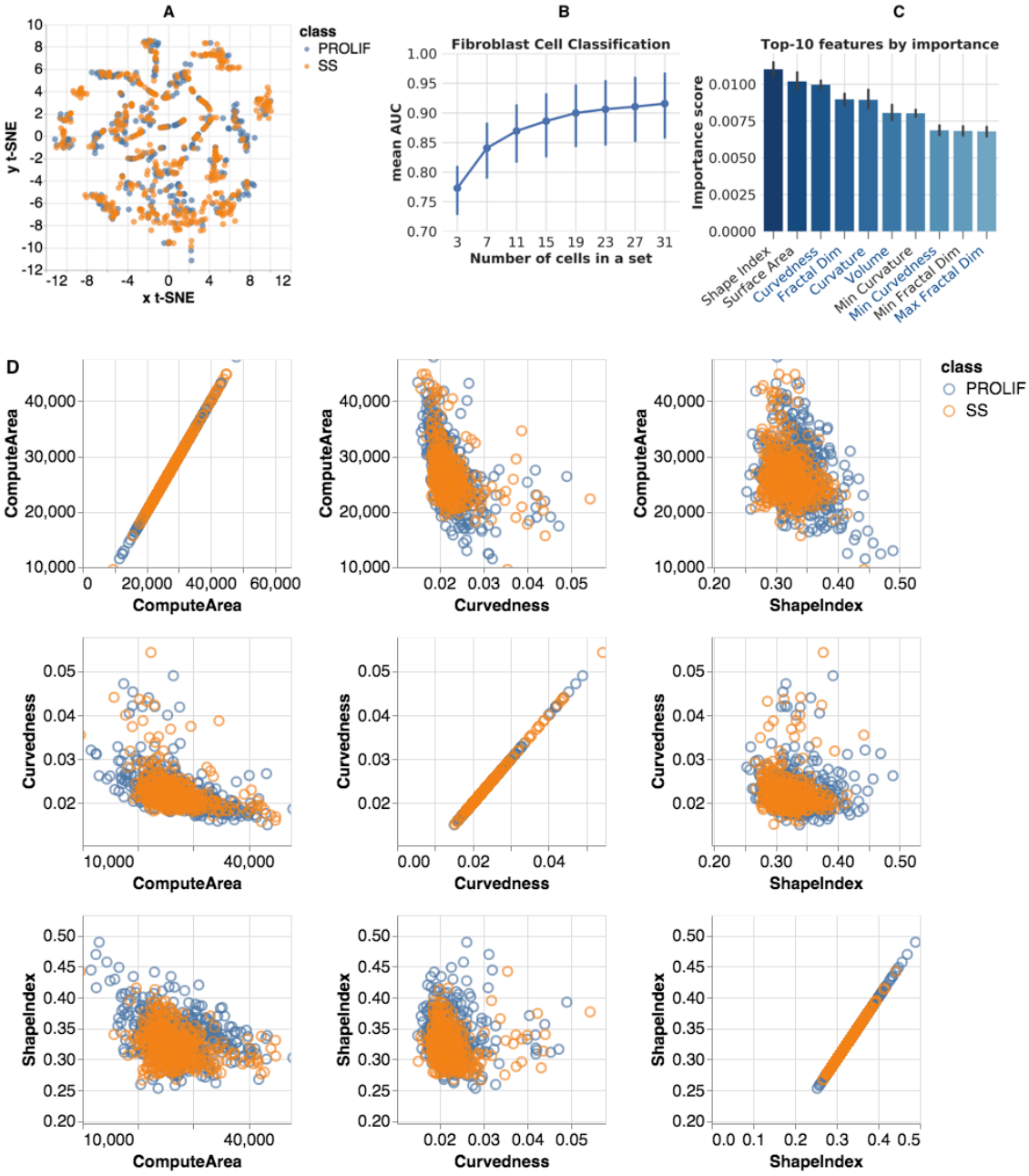
The full collection of fibroblast masks for binary classification consists of total 965 nuclei (498 SS and 470 PROLIF) and 2,181 nucleoli (1,151 SS and 1,030 PROLIF).

In order to assess the variability of extracted morphometry data, we include t-Distributed Stochastic Neighbor Embedding (t-SNE) (*Maaten and Hinton*, 2008) visualizations of the feature space generated by SOCRAT. Figure 4.2 demonstrates the variability of the extracted morphometry measures in a t-SNE projection visualized in SOCRAT. Although there was a small degree of grouping, there was no clear separation between classes.

The best result by a single classifier was achieved using a stochastic gradient boosting classifier with 1,500 base learners, maximum tree depth 8, subsampling rate 0.5. Hyper-parameters were fine-tuned using a cross-validated grid search. To evaluate these classification results, we measured accuracy, precision, sensitivity, and AUC over L2OGO cross-validation, which are presented in Table 2 for single cell and 19-cell-set classifications. Figure 5B shows mean AUC values for set sizes from 3 to 19 cells. A 90% mean AUC was reached when classifying sets with 19 cells and 92.5% for sets with 31 or more cells.

The gradient boosting classifier also computes and reports cross-validated feature importance, see figure 4.2. These allow us to evaluate which measures differ between two cell conditions, and potentially propose novel research hypotheses that can be tested using prospective data. Previous analysis has reported quantifiable changes in both nuclear size and shape under serum-starvation70. In our results, both nu-

Figure 4.2: Fibroblast morphometric analysis



Notes. Fibroblast morphometric analysis:(A) SOCRAT visualization of t-SNE projection of morphometric feature space; (B) mean AUC for various cell set sizes; (C) top-10 features for classification by importance score (right, nucleolar feature names start with Avg, Min, Max or Var, feature names that were also reported in top-10 for PC3 cells are shown in blue font); and (D): SOCRAT visualization of interactions between top-3 features.

Measure	single cell, mean ($\pm SD$)	19 cells set, mean ($\pm SD$)
Accuracy	0.699 ($\pm$0.076)	0.899 ($\pm$0.123)
Precision	0.701 ($\pm$0.075)	0.922 ($\pm$0.115)
Sensitivity	0.692 ($\pm$0.127)	0.874 ($\pm$0.224)
AUC	0.699 ($\pm$0.076)	0.899 ($\pm$0.123)

Table 4.3: Morphometry of synthetic objects. Cube (a=160), octahedron and sphere (R=80). AMC–average mean curvature, SA–surface area, SI–shape index, FD–fractal dimension.

clear (top-6, out of top-10) and nucleolar (4 of top-10) morphometric size and shape features are reported to be of high importance for distinguishing SS fibroblasts from PROLIF, see figure 4.2. We also visualized the relationship between top-3 features using SOCRAT figure 4.2. Visualizations suggest the smaller variation of morphometric measures in SS fibroblast nuclei compared to their PROLIF counterparts. This result may provide insight in further downstream analysis of potential underlying mechanisms that lead to these morphometric changes. We made the fibroblast morphometry data publicly available within SOCRAT for further analysis and validation (*Kalinin et al.*, 2017).

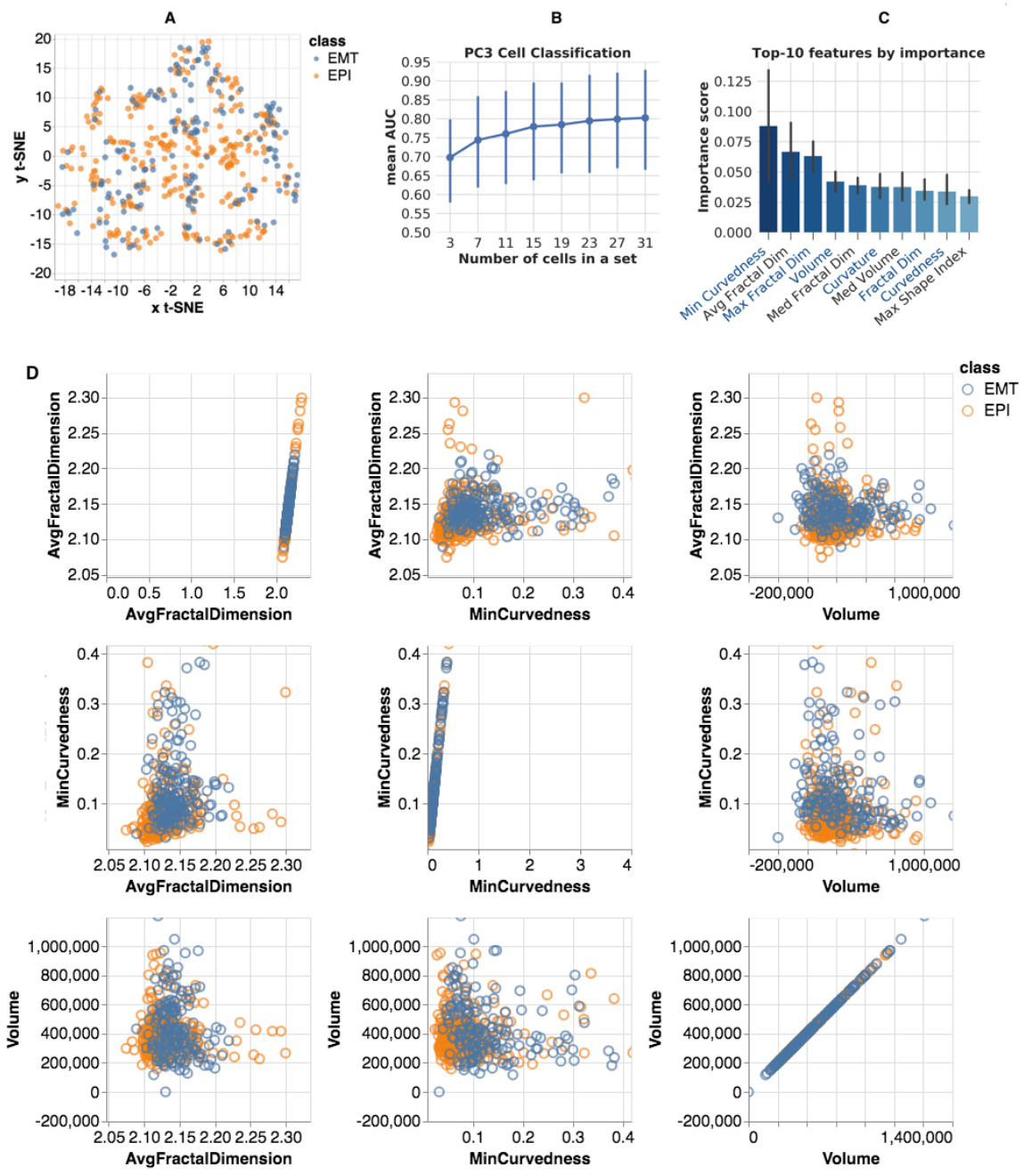
4.4 PC3 nuclear surface morphometry analysis

The second collection contains images of human prostate cancer cells (PC3). Through the course of progression to metastasis, malignant cancer cells undergo a series of reversible transitions between intermediate phenotypic states bounded by pure epithelium and pure mesenchyme (*Veltri and Christudass*, 2014). These transitions in prostate cancer are associated with quantifiable changes in both nuclear and nucleolar structure (*Montanaro et al.*, 2008; *Verdone et al.*, 2015). PC3 cells were cultured in: (1) epithelial (EPI), and (2) mesenchymal transition (EMT) phenotypic states. The collection includes 458 nuclear (310 EPI and 148 EMT) and 1,101 nucleolar (649 EPI and 452 EMT) 3D binary masks. Figure 4.3 demonstrates the

variability of the extracted morphometry measures in a t-SNE projection visualized in SOCRAT. Similar to fibroblasts, the projection of the PC3 morphometric feature space does not demonstrate clear separation between classes.

In this case, the best classification performance by single classifier is the result of applying a random forest model (1,000 trees, maximum tree depth 12, maximum number of features for the best split 40%). Hyper-parameters fine-tuning, accuracy metrics, and cross-validation procedures are identical to the ones reported in the previous fibroblast experiment. Classification of sets of 19 cells achieves a mean AUC of 76.2%, Table 3. Figure 6B reports the AUC for different group sizes to show how the classification performance increases with the cell-set size and reaches 80% for sets of 27 or more cells. In this experiment, we also examined the classifier-reported feature importance, Fig. 6C. The top-10 important features in this classification included nuclear (4 of top-10, which were also in Fibroblast top-6) and nucleolar (top-3, 6 out of top-10) shape morphometry features. Top feature interactions visualized using SOCRAT demonstrate the important changes in distributions of nucleolar morphometric measures, see figure 4.3. For example, it seems that the EPI nucleoli tend to have more variability in minimal curvedness and average fractal dimension, compared to EMT nucleoli. Previously reported PC3 morphological analyses (*Verdone et al.*, 2015) only used simple 2D nuclear form measures, such as diameter and the size of the bounding box. While we confirmed the importance of nuclear form in our results and suggested the need for further investigation of other highly ranked features, such as nucleolar curvedness, shape index, and fractal dimension, which may provide additional mechanistic insights. PC3 morphometry data are made publicly available within SOCRAT for further analysis and validation.

Figure 4.3: PC3 morphometric analysis



Notes. PC3 morphometric analysis: (A) SOCRAT visualization of t-SNE projection of morphometric feature space; (B) mean AUC for various cell set sizes; (C) top-10 features for classification by importance score (right, nucleolar feature names start with Avg, Min, Max or Var, feature names that were also reported in top-10 for Fibroblast cells are shown in blue font); and (D): SOCRAT visualization of interactions between top-3 features.

Measure	single cell, mean ($\pm SD$)	19 cells set, mean ($\pm SD$)
Accuracy	0.699 ($\pm$0.076)	0.899 ($\pm$0.123)
Precision	0.701 ($\pm$0.075)	0.922 ($\pm$0.115)
Sensitivity	0.692 ($\pm$0.127)	0.874 ($\pm$0.224)
AUC	0.699 ($\pm$0.076)	0.899 ($\pm$0.123)

Table 4.4: Morphometry of synthetic objects. Cube (a=160), octahedron and sphere (R=80). AMC–average mean curvature, SA–surface area, SI–shape index, FD–fractal dimension.

4.5 VPA-treated astrocyte morphometry analysis

4.5.1 Motivation and experiment description

Changes in nuclear morphology are associated with reorganization of chromatin architecture and related to altered gene regulation, cell function, differentiation and proliferation. One of the most important mechanisms in chromatin remodeling is the post-translational modification of the N-terminal tails of histones by acetylation (*Göttlicher et al.*, 2001). Histone deacetylation results in chromatin condensation and subsequent transcriptional repression while acetylation has an antagonistic effect leading to gene expression in cells (*Yang and Seto*, 2007; *Ganai et al.*, 2015). Acetylation of histones and other nuclear proteins plays an important role in cancer development and progression (*Kortenhorst et al.*, 2009). Therefore, inhibition of histone deacetylases (HDACs) through small-molecule inhibitors has gained significant attention in clinical research (*Ververis et al.*, 2013; *Ganai et al.*, 2015; *Eckschlager et al.*, 2017).

Valproic acid (VPA, 2-propylpentanoic acid) is an established drug in the long-term therapy of epilepsy, bipolar disorders, social phobias, and neuropathic pain (*Göttlicher et al.*, 2001; *Ganai et al.*, 2015). VPA has been shown to relieve HDAC-dependent transcriptional repression and to cause hyperacetylation of histones in cultured cells and *in vivo* (*Göttlicher et al.*, 2001). Most importantly, VPA induces differentiation, apoptosis, and autophagy of a variety of cancer cells and reduces tumor

growth and metastasis formation (*Göttlicher et al.*, 2001; *Montani et al.*, 2017). VPA is now together with other short chain fatty acids HDAC inhibitors tested in clinical studies as anticancer drugs (*Eckschlager et al.*, 2017). At the same time, it has been shown that VPA treatment results in quantifiable dose- and time-dependent changes in the nuclear structure of prostate cancer cell lines, reflecting change in chromatin remodeling dynamics in prostate cancer cells (*Kortenhorst et al.*, 2009).

Another potential application of VPA is related to its ability to promote neurogenesis and neuronal maturation by to enhancing the efficiency of cellular reprogramming mediated by HDAC inhibition (*Hsieh et al.*, 2004; *Zhang et al.*, 2015; *Gao et al.*, 2017; *Jang and Jeong*, 2018). Neuronal regeneration in adult mammalian brain is important for alleviation of brain injuries or neurodegenerative diseases. However, there are reports indicating that the regeneration capacity of adult brains may be limited and insufficient for brain repair (*Goldman*, 2016; *Li and Chen*, 2016; *Wang and Zhang*, 2018). Thus, cell replacement therapy using exogenous cells seems promising, including neuronal reprogramming from terminally differentiated somatic cells as a strategy to generate functional neurons (*Goldman*, 2016; *Gao et al.*, 2017). Astrocytes, the most abundant cell types in the brain, play important roles in maintaining brain homeostasis and modulating neural circuit activity (*Clarke and Barres*, 2013). Astrocytes developmentally originate from the same precursor cells as neurons, are capable of proliferating in response to brain damages, and therefore are considered as ideal starting cells to regenerate neurons (*Amamoto and Arlotta*, 2014). Astrocytes can be converted first into neuroblast cells and then differentiated into neuronal cells (*Niu et al.*, 2013; *Su et al.*, 2014). Many studies have already revealed that astrocytes of the central nervous system can be reprogrammed into induced neuronal cells by virus-mediated overexpression of specific transcription factors *in vitro* and *in vivo* (*Heinrich et al.*, 2010; *Niu et al.*, 2013; *Guo et al.*, 2014). However, application of this virus-mediated direct conversion is still limited due to concerns on clinical

safety *Cheng et al.* (2015). Compared to transcription-factor-based reprogramming, small molecules offer ease of use and a broader range of downstream applications *Zhang et al.* (2015). Recent studies have demonstrated the ability to directly reprogramming human astrocytes into functional neurons with a set of small molecules, including VPA (*Cheng et al.*, 2015; *Zhang et al.*, 2015; *Gao et al.*, 2017). Both *Cheng et al.* (2015) and *Zhang et al.* (2015) showed that removal of VPA from the the chemical small molecule cocktail reduced reprogramming efficiency. While *Cheng et al.* (2015) reported that VPA alone can induce astrocytes into neurons, *Zhang et al.* (2015) showed that if VPA is included in the reprogramming medium for more than 2 days it increased cell death. Such results are probably highly dependent on a specific protocol and concentrations of VPA, and moreover, may not serve as a good analogy to *in vivo* treatments. While corresponding changes in cell morphology were reported (*Cheng et al.*, 2015; *Zhang et al.*, 2015; *Gao et al.*, 2017), changes in nuclear morphology haven't been observed or quantified.

In this experiment we applied two different treatment protocols to human astrocytes culture with a goal to observe and measure the changes in 3D nuclear morphology that are reflective of chromatin reorganization. Both treatments went on for a week: cells first were imaged at the initial condition on day 0 and then treatment and imaging was conducted on days 3, 5, and 7. The first treatment protocol included only VPA. The second protocol used a small molecule cocktail treatment from (*Zhang et al.*, 2015) that also included VPA, but only on day 3. Full details of treatment are given in C.1. As a result, we obtained images of astrocytes in following conditions: control (CTRL), treated with VPA (VPA), and treated with small molecules (SM). We used DAPI as a nuclear stain, and the overall imaging protocol was similar to that for fibroblasts and PC3s, as desribed in A.3.

Treatment \ Day	0	3	5	7
CTRL	51	100	93	108
VPA	-	98	71	82
SM	-	86	98	139

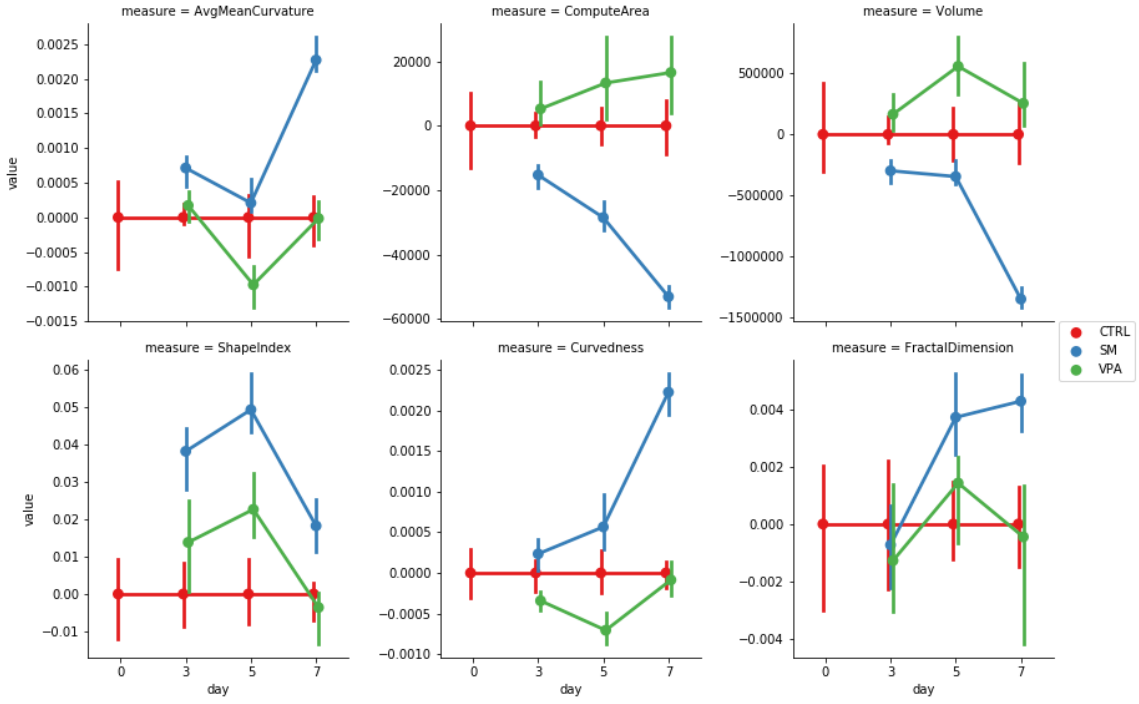
Table 4.5: Number of segmented astrocyte nuclei per treatment per day.

4.5.2 Morphological analysis of VPA-treated astrocyte nuclei

Each 3D volumetric image was re-sliced into a $1,024 \times 1,024 \times Z$ lattice ($Z = \{50, 80\}$), where regional sub-volumes facilitate the alignment with the native tile size of the microscope. All sub-volumes were saved as multi-image 3D TIFF volumes. For every sub-volume, accompanying vendor meta-data was extracted from the original data. Then, we segmented 3D nuclear binary masks from the original data sub-volumes with the Nuclear Segmentation algorithm from the Farsight toolkit (*Al-Kofahi et al.*, 2010; *Kalinin et al.*, 2018d), following the protocol described in Section 2.4.1. Table 4.5 shows the number of extracted 3D nuclear binary masks per cell condition per day. We then reconstructed surfaces of 3D binary nuclear masks using Laplace-Beltrami eigen-projection and topology-preserving boundary deformation (*Shi et al.*, 2010; *Kalinin et al.*, 2018c), as described in section 3.1.1. Using obtained surface representations, we computed intrinsic and extrinsic geometric metrics, including volume, surface area, mean curvature, curvedness, shape index, and fractal dimension (*Kalinin et al.*, 2018c) using a high-throughput computational workflow protocol introduced in section 3.1.2.

Figure 4.4 shows changes of morphometric measures over time for the control (CTRL) and two treatment cell conditions (VPA, SM). At each time point we calculate median values for each measure per treatment and then subtract the median value of the control group from medians of each of treatment groups. The results suggest that change is size and shape of nuclei occurred in astrocytes treated with the

Figure 4.4: 3D surface morphometry of VPA-treated astrocytes



Notes. Changes of morphometric measures over time relative to median value of the corresponding measure on control population.

small molecule cocktail were bigger than those in VPA-treated cells. Small molecule treatment results indicate that nuclear size was drastically decreasing as shown by both direct measures of nuclear surface area (ComputeArea) and volume (Volume) as well as indirect effect on extrinsic (not scale invariant) measures of shape, namely, mean curvature (AvgMeanCurvature) and curvedness (Curvedness). On the contrary, VPA-only treatment seems to be increasing the size of cell nuclei, especially on day 5. Scale invariant shape index (ShapeIndex) also indicated bigger changes of shape towards more round in cells treated with small molecules compared to VPA alone.

As a next step, we use extracted morphometric measures as features to train a Random Forest classification model (*Liaw and Wiener*, 2002) and evaluate pairwise discrimination between different treatments across time points using 3-fold cross validation. Mean AUC values shown in 4.6 demonstrate classification performance results

Treatments \ Day	3	5	7
CTRL vs VPA	0.677 (±0.093)	0.773 (±0.069)	0.692 (±0.088)
CTRL vs SM	0.737 (±0.070)	0.840 (±0.085)	0.948 (±0.027)
VPA vs SM	0.734 (±0.072)	0.826 (±0.065)	0.939 (±0.034)

Table 4.6: Pairwise classification performance of astrocyte nuclear morphologies, mean AUC.

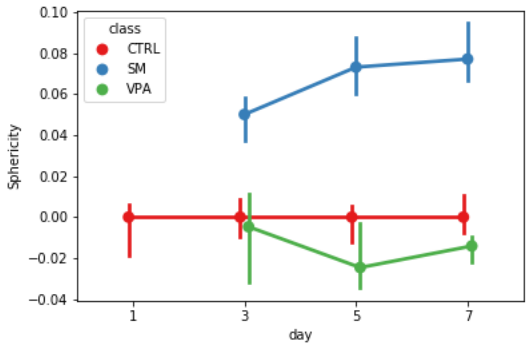
that are in general agreement with trends presented in figure 4.4. Both VPA and SM treatments introduce morphological changes in astrocyte nuclei that are indicated by the classification performance improvement from day 3 to day 6. However, on the day 7, nuclear morphologies of VPA-treated cells become more similar to those of the control population. Astrocytes treated with the small molecule cocktail exhibit the biggest difference in morphology from both control and VPA-treated groups on the last day of measurement.

Finally, to confirm that small molecule cocktail treatment results in larger, more round nuclear shapes, we computed voxel-based measure of 3D object sphericity. Sphericity is the measure of how closely the shape of an object approaches that of a mathematically perfect sphere. Sphericity of an arbitrary 3D object can be computed as a the ratio of the surface area of a sphere (with the same volume as the given object) to the surface area of the object (*Wadell*, 1935). We measure sphericity of a cell nucleus by fitting a 3D ellipsoid to the binary voxel mask using linear least squares (*Fitzgibbon et al.*, 1999) to calculate principal semi-axes: a, b, and c. We then compute the sphericity Ψ as following:

$$\Psi = \frac{\pi^{1/3}(6V_p)^{2/3}}{SA_p} \approx \frac{(a*b*c)^{2/3}}{(1/3*(a*b)^p + (a*c)^p + (c*b)^p)^{1/p}}, \tag{4.1}$$

where V_p is the volume of an ellipsoid and SA_p is the surface area, computed by the approximate formula using $p \approx 1.6075$ that yields a relative error of at most 1.061%

Figure 4.5: Nuclear sphericity of VPA-treated astrocytes



Notes. Changes of 3D sphericity of treated nuclei over time relative to median sphericity of control population.

(*Xu et al.*, 2009). Results shown in figure 4.5, confirm that nuclei of SM-treated cells become more spherical over time. VPA treatment slightly decreased roundness of nuclei on day 5, but then it was recovered on day 7 to that of control group. These results are in agreement with the 3D surface morphometry results shown in fugue 4.4.

Results of this experiment show that chromatin remodeling in astrocyte cell induced by VPA and small molecule cocktail treatments is reflected in quantifiable changes in nuclear morphology. 3D surface morphometry captures the trends in morphology changes over the course of treatment and allows to assess the difference between treatments and the control at every time point. As suggested by *Zhang et al.* (2015), VPA alone as a treatment *in vitro* seems to be not as effective as when combined with other small molecules. That is confirmed by characterization of VPA-only treated astrocytes that towards that end of the treatment course exhibited morphologies close to those of the control group. Small molecule cocktail treatment seems to be more effective and showed a robust deviation from controls in almost all measured morphological features. Using a machine learning model confirmed the better discrimination between controls and SM treatment compared to VPA. Finally, an additional voxel-based sphericity measure confirmed that SM treatment lead to larger, more spherical and round nuclear morphologies. These observations can pro-

vide further insight in details of chromatin reorganization in the astrocyte-to-neuron reprogramming process and relate them to underlying molecular mechanisms.

4.6 Concluding remarks

In this section we demonstrated applications of 3D surface morphology modeling, morphometrics, visual analytics, and deep learning to cell nuclear and nucleolar morphology classification and analysis. Robust surface reconstruction allows accurate approximation of 3D object boundaries that was validated on synthetic data. Suggested shape morphometric measures outperform another popular approach and demonstrated their universality across different cell types, conditions, and even domains. Using 3D reconstructed nuclear surfaces as inputs to a sparse 3D tetrahedral convolutional neural network have demonstrated a notable increase in the morphological classification performance, which indicates the high potential for applying deep learning models for this type of problems.

We tested our approach on the 3D cell nuclear morphology microscopy imaging dataset, which includes fibroblast and PC3 cells with a total of 1,500 nuclear and 2700 nucleolar masks. The classification results on these data comparing epithelial vs. mesenchymal human prostate cancer cell lines, and serum-starved vs. proliferating fibroblast cell lines, demonstrate the high accuracy of cell type prediction using 3D morphometry, especially when applied to sets of cells. Although different classification algorithms appear to be optimal for different experiments, we observed that both nuclear and nucleolar morphometric measures are important features for discriminating between treatment conditions or cell phenotypes. Similarly to the baseline results, tree-based ensemble classifiers have demonstrated best performance results, due to their ability to capture complex patterns in data. Reported feature importance ranks confirm and extend previously published results. Interestingly, there were 3 common morphometric features among the top-10 most important ones for both cell lines.

In the case of fibroblast classification, the results show the importance of nuclear morphometry, as reported in previous studies (*Seaman et al.*, 2015) Additionally, the number of nucleoli per nucleus and various internal nucleolar morphometric measures such as nucleolar curvature and nuclear fractal dimension appeared too be important for discriminating between serum-starved and proliferating fibroblasts, which may indicate possible effects of serum-starvation on the nucleolar structure.

For PC3 cells, the most important classification features are the moments of the distributions of various nucleolar morphometric measures, along with nuclear size and shape. This confirms previously reported observations (*Verdone et al.*, 2015) and suggests new important morphological characteristics, such as nuclear curvature and nucleolar fractal dimension. This demonstrates that our method extracts relevant information from cell forms to successfully classify cells using a combination of criteria. In addition, this also shows the importance of sophisticated shape metrics, compared to volume and surface area, that alone, were not the most informative features for the classification results.

The use of SOCRAT enables interactive interrogation of morphometric data in a visual manner, supported by analytical tools. This method of interactive visual analytics provides insight into feature dependencies and interactions, and can be used for result interpretation. We also demonstrated the visualization of 3D volumetric images and derived meshed surface representations using the SOCR Dynamic Visualization Toolkit web application (*SOCR*, 2018).

CHAPTER V

Conclusions

Cell nuclear morphology aids in the proper 3D organization of the genome and is perturbed across a wide spectrum of human diseases. 3D cell microscopy is a powerful technique that enables investigation of biological mechanisms related to morphological changes in the cell nucleus through analysis of changes in its size and shape. Quantification of nuclear morphology enables more subtle characterization of cellular phenotypic traits, which can be associated with functional changes coupled to underlying biological processes. In this chapter, we discuss how the ability to automate the processes of specimen collection, image acquisition, data pre-processing, computation of derived biomarkers, modeling, classification, and analysis can significantly impact clinical decision-making and fundamental investigation of cell nuclear deformation.

5.1 Main findings

In this work we contributed a scientific framework for 3D cell nuclear morphological analysis. This framework includes our knowledge-base regarding morphological microscopy imaging data collection, pre-processing, segmentation and analysis, which resulted in a publication of the biggest freely available dataset for 3D nuclear and nucleolar morphological analysis and classification. We have provided sample specific preparation protocols, imaging conditions, and summary information about the

dataset, along with baseline voxel-based morphometry classification results. We also suggested a specific cross-validation scheme that enables computing interval estimates for classification performance metrics while accounting for possible batch effects.

Next, we contributed an approach for accurate 3D surface morphometry. To our knowledge, this is the first attempt to combine 3D cell nuclear shape modeling by robust smooth surface reconstruction and extraction of shape morphometry measures into a highly parallel pipeline workflow protocol for morphological analysis of thousands of nuclei and nucleoli in 3D. Surface reconstruction protocol is based on Laplace-Beltrame eigen-projection method that has been shown to be more accurate and less susceptible to noise that widely used alternative spectral-based approaches, such as spherical harmonics. Validation on synthetic data confirmed the ability of the proposed technique to accurately represent and distinguish geometric characteristics of various 3D shapes. We also considered new powerful approach encompassing first ever application of sparse 3D deep convolutional neural networks to nuclear morphological classification that demonstrated state-of-the-art performance on fibroblast data. Coupled with the toolbox for interactive exploratory visual analytics, this approach allows efficient and informative evaluation of cell nuclear shapes in the imaging data and represents a reproducible technique that can be validated, modified, and repurposed by the biomedical community.

Finally, we demonstrated successful applications of proposed methodology to multiple nuclear morphology analysis problems. Our processing protocol was able to extract relevant information about 3D object morphologies from imaging data of different cell types in various conditions. There were two specific applications that have demonstrated the power and the efficiency of the proposed approach.

First, we compared morphologies of human prostate cancer cells (PC3) that undergo epithelial-to-mesenchymal transition. We were able to discriminate with higher confidence subtle morphological differences compared to the previously reported re-

sults due to the two main reasons: the flexibility and accuracy of the proposed pipeline and the use of both nuclear and nucleolar morphological measures. Moreover, we were able to not only confirm previously reported results regarding relevant morphological measures, but also obtained new ones that can provide insight in the underlying biological mechanisms and progress understanding of pathology of prostate cancer.

Second application involved monitoring nuclear morphologies of cell during drug-induced chromating remodeling via the inhibition of histone deacetylases in astroglial cells. Observing astrocyte nuclear morphologies over time allowed us to use our approach for building morphological trajectories or timelines that reflected the differences in drug-treated cell sub-populations compared to controls. In the presence of two different treatment we were not only able to distinguish them from controls, but also demonstrate the difference in their effects on cells, something that was previously only hinted in the literature.

Thus, the combination of suggested methods allows to perform both hypothesis testing as well as data-driven discovery in 3D nuclear morphology analysis. Moreover, these methods are universal and are not limited to limited to nuclear and nucleolar shapes. With some minor changes, it can be applied to other 3D cellular, sub-cellular, and sub-nuclear compartments and organelles of interest.

5.2 Future perspective: impact on basic research

There are multiple trends that indicate the importance of 3D morphological analysis in future investigations of fundamental investigation of cell nuclear architecture.

Results from the 4D nucleome program, funded by the National Institutes of Health (NIH), have led to the realization that significant molecular variation, which accounts for human differences in medication response and adverse reactions are likely based on the intricate organization of the spatial genome. Spatial and temporal morphological changes in the nucleus and nucleoli are associated with the underlying

reorganization of the chromatin architecture in 3D, as our results have confirmed. Thus, accurate quantitative measures of cell and nuclear morphologies will become of even more importance as a accessible and powerful proxy representation of underlying mechanistic transformations. At the same time, advances in imaging technologies allow capturing more different objects simultaneously, with higher resolution, and over time. From morphology point of view this progress lays out an important perspective of being able to quantitatively, on a level of cell population, obtain 3D cell and sub-cellular morphometries on a different hierarchical levels: from cell and nucleus, to nucleoli and other organelles, to chromosome territories and TADs. This will provide an opportunity for the multi-scale assessment of effects of underlying biological processes on cellular architecture. Moreover, together with molecular level techniques, such as Hi-C, such 3D shape morphometry workflow can form even more powerful combination for the investigation of DNA architecture in the spatial and temporal framework of the 4D nucleome. For example, measuring effects of various biomarkers and genetic and epigenetic variation on the chromatin re-organization (e.g. TAD-affecting regulatory SNPs), also captured by the morphometry from 3D imaging assays, would allow establishing a new mechanistic model of the processes that connect these modalities.

Another example of the many possible future applications of this workflow is to study asymmetric cell division (*Zheng et al.*, 2018). Stem and progenitor cells are characterized by their ability to self-renew and produce differentiated progeny. A balance between these processes is achieved through controlled asymmetric divisions and is necessary to generate cellular diversity during development and to maintain adult tissue homeostasis. Disruption of this balance may result in premature depletion of the stem/progenitor cell pool, or abnormal growth. In many tissues, dysregulated asymmetric divisions are associated with cancer. Whether there is a causal relationship between asymmetric cell division defects and cancer initiation is unknown.

Our shape analysis pipeline can be useful in studying the 4D nucleome topology of morphogenesis and cancer initiation.

Such deep synergies of genome- and phenome-level information will enable more effective two-sided approaches to uncovering the shape–function dynamics of the cell.

5.3 Future perspective: impact on clinical applications

Application of cellular and nuclear morphometry are already in use in both phenotypic drug discovery and diagnostics.

As discussed in the Introduction, pathologists have been using cell and nuclear morphology to detect various pathologies such as cancel for decades. While 3D imaging may not be always necessary for more straightforward tasks, there are cases when it is more beneficial. Specifically, Vision Gate company that produces Cell-CT platform (*Wilbur et al.*, 2015; *Meyer et al.*, 2015; *Pantanowitz et al.*, 2018), uses 3D cell morphometry for early detection of lung cancer from spitum samples that is not possible to confidently recognize from lung CTs. More globally, more diseases and conditions that are knows to affect cell, nuclear, and sub-nuclear shape and size will be possible to detect via morphometry as sample preparation, imaging technologies, and data analysis methodologies evolve. Powerful new analytics methods such as deep learning will enable better-than-human performance in such complex pattern recognition tasks without the need to hand-craft hundreds of features for every experiment. Furthermore, wider use of transfer learning will help to alleviate the need of collecting massive amounts of data, allowing instead to re-purpose learnt patterns from task to task. Our preliminary results showed a great promise for improved morphological discrimination using this type of models.

On the other hand, such companies as Recursion Pharmaceuticals are using morphological cell profiling to screen thousands of compounds against hundreds of disease models and find promising combinations (*Bray et al.*, 2016). They have a goal of

building a comprehensive database mapping the effects of tens of millions of genetic, chemical, and other biological perturbations on many cell types relevant to human disease. This would allow to predict potentially useful applications of already known treatments to new conditions, reducing the need for *de novo* experimentation.

Finally, these trends show the possibility of the personalized prediction of the treatment outcome, adverse events, and even dosage recommendation prediction, given the genomic, phenomic, ad other types of data and assisted by accurate morphometry. In the novel framework of pharmacoepigenomics, advances in cell morphological characterization will enable identification mechanisms assosiated with novel regulatory variants located in noncoding domains of the genome and their function; the mechanistic prediction of drug response, targets and their interactions.

5.4 Open science considerations

The very first challenged that we faced in this work was the lack of any public datasets for morphological analysis and classification. The sharing of high quality, labeled datasets is extremely valuable for the progress in this field; however, a clear asymmetry exists with government-sponsored academic researchers directed to share, while researchers in industry (e.g., from abovementioned companies) are often prohibited from sharing code, data and results due to proprietary and intellectual property protections. However, this situation is already changing in the machine learning and deep learning communities, that has witnessed acceleration of progress via public-posting of various datasets for benchmarking and software tools, including those developed and used in the industrial setting. Code-sharing and open-source licensing are also essential for continued progress in this domain. In order to promote the reproducibility of results, facilitate open-scientific development, and enable collaborative validation, we made the pipeline workflows, together with underlying source code, documentation, and derived data from this study, available online on the

project web-page: `http://www.socr.umich.edu/projects/3d-cell-morphometry`. Additionally, extracted morphometric features are made available for interactive exploration and analysis online via our visual analytics platform SOCRAT.

5.5 Open challenges and future directions

There are many remaining challenges related to different aspects of this work. For example, although the proposed cross-validation technique (L2OGO) allows to account for possible batch effects and calculate interval estimates of classification performance measures, in the case of imbalanced data or when the number of segmented nuclei varies a lot between images, it causes high variability of computed metrics. This can be possibly addressed by better class balancing, e.g. via subsampling or oversampling, or loss weighting during the each iteration of the cross-validation process.

Furthermore, one label per image is not always representative of all cell phenotypes in that image. Images can contain artifacts, debris, apoptotic and other non-target phenotypes. This can be addressed by using weakly-supervised methods that are robust to label noise or more advances curation steps that filter out such objects before the final classification step. Such filter could be tuned by observing the morphometric feature space and singling out objects who's metrics do not fit well into typical distribution for a given phenotype. Another option is to have a machine learning model trained to filter out objects that are not of interest, however, it may require a substantial amount of manual labeling.

In order to increase the number of extracted features and, this, the numeric shape representation, more geometric measures can be used to characterize object of interest, such as intrinsic shape context, compactness, symmetry, smoothness, convexity, etc. In the current representation, analyzable shapes are limited to genus zero surfaces, which is a fair assumption when modeling objects like nuclei or nucleoli. How-

ever, it might be not trivial when considering other nuclear structures, for example, chromosome territories or interchromosomal loops, since their topologies may not be homeomorphic to a sphere, or may not appear to be genus zero under some imaging conditions and modalities. It is also conceivable, yet not very likely for the discretized LB, that 2 different shapes may have the same spectra. In this case, the algorithm may fail to detect the intrinsic differences between them due to false-negative error. Combining features extracted from different object shape representations, e.g. voxel-based and surface-based can be helpful for improving the classification performance. Additionally, 3D textural features could possibly increase discriminatory power of the method and provide more information on chromatin reorganization. Since nuclear deformation serves as a proxy to underlying processes, the importance of particular features and the methods ability to classify nuclei does not provide direct insight into the fundamental biological mechanism driving the observed morphometric differences between cell phenotypes or environmental conditions. The computational results should be further tested and externally validated using other experimental conditions and prospective data.

Although deep learning based model applications have demonstrated superior classification performance, their interpretability and computational requirements remain main challenges. Unlike human-defined geometric features, weights of an artificial neural networks do not provide a direct mapping of their values to an intuitively understandable characteristics of the object. Interpretability of deep networks is an increasingly popular area of research that already produced a number of potential solutions that could be employed to address this challenge. Interesting prospect lies in relating deep neural network coefficients and hand-crafted geometric measures, which could lead to both improved discriminative performance and better interpretability.

SOCRAT implements a visual analytics workflow that encompasses an iterative process, in which data analysts can interactively interrogate extracted morphometric

measures in the form of interactive dialogue supported by visualizations and data analysis components. However, at this point, SOCRAT requires the user to enter data in a specific format ("long" or "tidy" data format). One of possible directions of future development there is to provide better user experience supported by both more convenient interface and smarter tools that can recognize data formats and help user to convert between different ones.

Overall, we expect biological image analysis approaches to become increasingly automated, accurate, and reliable. The capabilities of current solutions are already moving beyond simply automating what a biologist can do and are beginning to enable comprehensive analysis of all available information. This is especially important, given that imaging is not the only source of information in biology. Genomics, proteomics, transcriptomics, metabolomics, and other 'omics' all provide complementary views on biological processes of interest, and their combination with imaging will become a much richer source of knowledge than each field can offer individually. The abundance of data in all these fields poses major challenges in terms of standardized data storage and retrieval, but even more so for integrative data analysis. Although the development of methods for data analysis from each individual source remains important, methods that properly account for the relationships between types and modalities of data from heterogeneous sources have the potential to obtain results that that would be impossible to produce otherwise. Thus we anticipate research in this field to rely more on data-driven approaches and involve knowledge from different areas and disciplines. Meanwhile, current developments in biological image analysis are already equipping biologists with more automated, robust, and accurate tools and will prove indispensable in investigating the cellular and molecular organization in health and disease.

APPENDICES

APPENDIX A

Additional Information for Chapter II

A.1 Fibroblast sample preparation protocol

Fibroblasts (newborn male) were purchased from ATCC (BJ Fibroblasts CRL-2522 normal) and subjected to a G0/G1 Serum Starvation Protocol. They were recovered from cryogenic storage with growth in full media (MEM + 10% FBS Sigma-Aldrich + 1% MEM NEA Gibco Lot 1656019 + 1% Antibiotics Gibco Lot 1523692) for 48 hours. This was followed by a change to serum free media (0.1% FBS) for 4 days, followed by resuspension and growth on coverslips in serum free media for 24hrs. Aliquots were taken for fixation every 2hrs for 12hrs.

A.2 PC3 sample preparation protocol

PC3 EPI /EMT slides were prepared as follows:

1. 40K cells were seeded on slides and grown for 48 hours @ 37C, 5% CO2, and 90% RH

2. slides were washed in PBS @ RT

3. slides were fixed in 4% paraformaldehyde for 30 minutes @ RT

4. slides were washed in PBS @ RT

5. slides were dehydrated in 50% EtOH for 5 minutes @ RT

6. slides were dehydrated in 50% EtOH for 5 minutes @ RT

7. slides were dehydrated in 70% EtOH for 5 minutes @ RT

8. slides were dehydrated in 100% EtOH for 10 minutes @ RT

9. slides were allowed to air dry @ RT

A.3 Staining protocol

Before use, cells were rehydrated in descending ethanol concentration washes, 5mins each. Staining proceeded according to the following methods for each stain, with stains for up to five features applied jointly to the same sample in different colors to be imaged on different fluorescent channels:

1. Prolong Gold antifade reagent with DAPI (4',6-diamidino-2-phenylindole) (Invitrogen, Lot 168129) was applied to all samples according to manufacturer protocol to image the nucleus as a whole.

2. For EtBr imaging of nucleoli, Ethidium Bromide was applied by application of 5ul of EtBr working suspension onto a wet coverslip for 20 seconds, followed by a PBS wash, immediately prior to the application of DAPI.

3. For fibrillarin imaging of the nucleoli, antifibrillarin Alexa 448 label was purchased from ABCAM (EPR10823(B) Lot GR175169-1) and applied according to the manufacturers protocol prior to the application of DAPI.

A.4 Flow cytometry for fibroblast cells

Cell cycle profiles were confirmed for synchronized serum-starved and proliferating fibroblasts with flow cytometry. Serum-starved and proliferating cells were trypsinized, then fixed in PBS suspension with 100% cold ethanol added drop-wise while vortexing, followed by incubation for 20 minutes at -20C. Cells were then pelleted by centrifugation at 1000 rpm for 5-7 minutes, the excess ethanol decanted, and cells re-suspended in PBS. Approximately cells were filtered through a 40um cell strainer. Propidium Iodide (TOCRIS Biosciences (Batch 1A/170341) was prepared as a stock solution of 10mg/10mls in PBS, and diluted 1:20 to the working solution, with 1:1000 of RNAse Cocktail Enzyme (Ambion L/N 00268539) added. Resuspension was followed by incubation for 40 minutes at RT. Flow cytometry was performed by the University of Michigan Flow Cytometry core using Beckman Coulter CyAn ADP. Flow cytometry results show 92.6% of the cells in the G0/G1 phase for the synchronized serum-starved fibroblasts vs 69.5% for proliferating fibroblasts.

A.5 Imaging protocol

3D confocal imaging used a Zeiss LSM 710 laser scanning confocal microscope using a 63x PLAN/Apo chromate 1.4na DIC objective. Laser excitation for each channel proceeded with a Zeiss laser of the appropriate wavelength, followed by imaging in a set of wavelengths corresponding to the emission peaks of the fluorophore according to the Zeiss fluorophore database or manufacturer specifications, trimmed to avoid overlap between channels. Imaging proceeded with 1024×1024 pixels in a $128 \times 128 \mu M$ area. For 3D imaging, the confocal pinhole was approximately 0.5 Airy Units, with stacks of optical sections at 0.4 micron intervals, and stitchless no-overlap scanning across many frames of view in the XY plane, acquired by the LSM 710s automated scanning stage. Laser intensity was identical for each run using a particular

stain, while photomultiplier gain was adjusted to compensate for staining variability between samples. Further detail is provided in the metadata of our imaging dataset. Volume data in vendor-specific formats (e.g. Zeiss CZI) was archived in Omero (https://www.openmicroscopy.org/site/products/omero), the image repository of the Open Microscopy Environment (OME), and run through a series of pre-processing steps. The vendor-specific metadata stored in the file was parsed, extracted, saved in a local database and made available for use by other pipeline modules. This metadata included fields such as number of channels, X, Y, and Z size, and X, Y, and Z scaling factors. Typically, hundreds of metadata fields were extracted for each volume. The Bio-Formats Library (https://www.openmicroscopy.org/site/products/bio-formats) developed by the OME consortium was used to open and read the vendor-specific files and perform metadata extraction.

A.6 Segmentation details

Nuclear segmentation protocol included following steps:

- Convert each volume to 8-bit greyscale and apply despeckling using ImageJ (*Schindelin et al.*, 2015)
- Segment each volume in 3D using the Farsight toolkit's Nuclear Segmentation algorithm (*Al-Kofahi et al.*, 2010)
- Fill holes in derived 3D nuclear masks

Parameters for the Farsight toolkit's Nuclear Segmentation module were chosen as following:

- high_sensitivity: 0
- LoG_size: 30

- max_scale: 35
- xy_clustering_res: 2
- z_clustering_res: 1
- finalize_segmentation: 1
- sampling_ratio_XY_to_Z: 2
- Use_Distance_Map: 2
- refinement_range: 2

A.7 Curation data flow and post-processing modules

A.7.1 CurateCn

Inputs a segmented image file (1st parameter) and computes these voxel parameters for each segment ID and reported in the corresponding .log and .csv files.

- Centroid
- Voxel count
- Surface voxel count
- Spherical compactness
- Void count
- Edge count
- Bound true/false
- Adjacency matrix (only reported on .csv file)

Inputs the filter file (2nd parameter) and applies it to the computed voxel parameters. The filter file contains inclusive min/max values for these voxel parameters.

- Voxel count
- Void count
- Edge count
- Bound true/false
- Spherical compactness

Outputs:

- name_c0_gGGG_mask.log – log file
- name_c0_gGGG_mask.lst – contains the filenames of image files that passed the filter test
- name_c0_gGGG_mask.csv – spreadsheet readable file with voxel parameters

Notes:

- Processes a single .TIF file and can be executed in parallel with other images.
- For C2 images, the corresponding name_c0_gGGG_mask.lst file must have been processed

A.7.2 FilterC2Lst

Inputs these files given at parameters:

- name_c2_gGGG_statistics.txt – statistics file created by an earlier process
- name_c2_gGGG_connected_prelim.lst file - created after

- CurateCn processed at C2 image

Evaluates each nucleoli in the input .LST to see if it contains a non-zero amount of Fibrillarin (given in the .TXT tile). Outputs:

- name_c2_gGGG_connected.lst – contains filenames that passed the Fibrillarin test

Notes:

- Processes a single .LST file and can be executed in parallel with other images.
- The corresponding name_c2_gGGG_connected_prelim.lst file must have been processed.

A.7.3 MergeCn

Inputs all .LST files with names that match the 2nd command line parameter. The contents of all input .LST files is merged into a single file (name is the 1st parameter). Note the 1st parameter is traditionally the name of the working folder. Example to merge files in the run 0169 data set:

- java -cp . MergeCn ../0000169/0000169_c0.lst mask.lst
- java -cp . MergeCn ../0000169/0000169_c2_pre.lst prelim.lst
- java -cp . MergeCn ../0000169/0000169_c2.lst connected.lst

APPENDIX B

Additional Information for Chapter III

B.1 Definitions of morphometric measures

Table B.1: Size measure descriptions.

Geometric measure	Mathematical formulas	Interpretation
Volume	$\iiint_{\mathbb{R}^3} I_D(x,y,x)dxdydz$	The amount of 3D space enclosed by a closed boundary inside of a 3D solid which is quantified numerically in world coordinates. The volume of a solid represents the space capacity of the object.
Surface area	$\iint_{\Omega} \lvert\overrightarrow{r_u} \times \overrightarrow{r_v}\rvert dudv$	The surface area of a 3D solid object is the total area of its (curved) boundary (a 2-manifold). Surface areas of flat polygonal shapes must agree with their geometrically defined area. Volume and surface area are invariant under the group of Euclidean motions.

Table B.2: Shape measure descriptions.

Geometric measure	Mathematical formulas	Interpretation
Mean curvature	$MC = H = \frac{k_1 + k_2}{2}$	The only surface in $\mathbb{R}^3$ with constant positive mean curvature is the sphere. The curvature provides (local at each vertex) surface classification: • **Elliptical** both principal curvatures have the same sign and the surface is locally convex. • **Hyperbolic**: the principal curvatures have opposite signs, and the surface will be locally saddle shaped. • **Parabolic**: one of the principal curvatures is zero. Parabolic points generally lie in a curve separating elliptical and hyperbolic regions.
Shape index	$SI = \frac{2}{\pi}\, arctan(\frac{k_1 + k_2}{k_2 - k_1})$	Shape index is a qualitative measure of shape and can be sensitive to very subtle changes in surface shape, particularly in regions where the total curvature (or the curvedness) is very low.
Curvedness	$CV = \sqrt{\frac{k_1^2 + k_2^2}{2}}$	Curvedness is a function of the root-mean-square curvature of the surface, with flat areas of the surface having a low curvedness and areas of sharp curvature having a high curvedness.
Fractal dimension	$FD = \frac{log(N)}{\frac{1}{\rho}}$	Fractal dimension is a ratio providing a statistical index of complexity comparing how detail in a pattern (strictly speaking, a fractal pattern) changes with the scale at which it is measured.

B.2 Nuclear morphometric classification live demo

This demo is prepared for classification of serum-starved Fibroblast cells (SS, 160). This workflow take as an input original 16 $1024 \times 1024 \times Z$ 3D TIFF images (sub-volumes) in DAPI channel (c0) and metadata. It demonstrates nuclear binary mask preparation, 3D shape modeling, morphometric measure extraction, and classification running in distributed mode on a cluster using LONI Pipeline guest mode. It outputs .csv file with image-level output label, nucleus-level accuracy and average probability as well as labels and probabilities for individual nuclear masks that were segmented out of 3D input sub-volume, passed the curation, 3D shape modeling, feature extraction, and classification.

Instructions below describe how to use Pipeline in a guest mode. If you already have LONI Pipeline credentials you can just download Pipeline Client and log in using your username and password.

1. Download and install LONI Pipeline Client Web Start (requires Java)
2. Create "Try-It-Now" connection by clicking Connections icon at the bottom-right corner of the client to connect to the server without credentials (enter space for password)
3. Download workflow file and open in the Pipeline client
4. Click Run button at the bottom of the client – after workfklow validates the protocol, presence of input data, and availability of free nodes in cluster, it will start running jobs
5. Running the worflow take 2-3 hours on average, depending on availability of computing nodes in the cluster
6. After workflow is completed, right-click on Calculate Accuracy module in Classification group and download or view the output file from Output Files tab

You can double-click on group in the workflow at any moment to see individual modules inside. You can disconnect while the workflow is running – under Connections you will be able to see your unique GUEST-ID that you can use to reconnect later and check workflow status (enter space for password). Having your GUEST-ID you should be able to use LONI Pipeline Web App to reconnect to the same sessions (web app is still in Beta and might not work as expected). Workflow protocol can be ran multiple times to validate reproducibility of the morphometry results. Pipeline documentation, including instructions module definition, modification, and execution, is available on the official website.

APPENDIX C

Additional Information for Chapter IV

C.1 Astrocyte treatment protocol

Day 0:

- Collect 8 Day 0 samples
- Fix samples in 4% PFA for 10mins
- rinse 3 x 5mins in PBS
- store samples in PBS at 4deg

Replace media with 50% growth media and 50% N2 media (DMEM/F12 + 1X pen/strep, 1X N2 supplements)

Day 1: Completely replace media with N2 media containing: TTNPB (0.5 M), SB431542 (5 M,), LDN193189 (0.25 M,), and Tzv (0.5 M,) for SM treated samples:

- SM1 Treated samples: For 30ml of N2 media add 1.5ul TTNPB, 30ul SB431542, 1.5ul LDN193189, 3ul Tzv
- control samples: For 30ml of N2 media add 36ul DMSO

- VPA Treated (1.5mM VPA): For 30ml of N2 media add 450ul VPA

Day 3: Collect 6 Day 3 SM treated Samples, 6 Day 3 control samples and 6 Day 3 VPA samples:

- fix samples in 4% PFA for 10mins
- rinse 3 x 5mins in PBS
- store samples in PBS at 4deg

Replace with a different set of small molecules including CHIR99021 (1.5 M,), DAPT (5 M,), VPA (0.5mM,), and Tzv (0.5 M).

- SM2 Treated samples: For 30ml of N2 media add 4.5ul CHIR99021, 30ul DAPT, 150ul VPA, 3ul Tzv.
- Control samples: For 30ml of N2 media add 187.5ul DMSO.
- VPA Treated (1.5mM VPA): For 30ml of N2 media add 450ul VPA

Day 5: Collect 6 Day 5 SM treated Samples, 6 Day 5 control samples and 6 Day 5 VPA samples.

- fix samples in 4% PFA for 10mins
- rinse 3 x 5mins in PBS

Replace media with N2 media containing only CHIR99021 (1.5M), DAPT (5 M), and Tzv (0.5 M)

- SM3 Treated samples: For 30ml of N2 media add 4.5ul CHIR99021, 30ul DAPT, 3ul Tzv
- Control samples: For 30ml of N2 media add 37.5ul DMSO
- VPA Treated (1.5mM VPA): For 30ml of N2 media add 450ul VPA

BIBLIOGRAPHY

BIBLIOGRAPHY

Al-Aziz, J., N. Christou, and I. D. Dinov (2010), SOCR motion charts: An efficient, Open-Source, interactive and dynamic applet for visualizing longitudinal multivariate data, *J. Stat. Educ.*, *18*(3).

Al-Kofahi, Y., W. Lassoued, W. Lee, and B. Roysam (2010), Improved automatic detection and segmentation of cell nuclei in histopathology images, *IEEE Transactions on Biomedical Engineering*, *57*(4), 841–852.

Allis, C. D., and T. Jenuwein (2016), The molecular hallmarks of epigenetic control, *Nature Reviews Genetics*, *17*(8), 487.

Amamoto, R., and P. Arlotta (2014), Development-inspired reprogramming of the mammalian central nervous system, *Science*, *343*(6170), 1239,882.

Antoine, J.-P., and P. Vandergheynst (1999), Wavelets on the 2-sphere: A group-theoretical approach, *Applied and Computational Harmonic Analysis*, *7*(3), 262–291.

Arganda-Carreras, I., V. Kaynig, C. Rueden, K. W. Eliceiri, J. Schindelin, A. Cardona, and H. Sebastian Seung (2017), Trainable weka segmentation: a machine learning tool for microscopy pixel classification, *Bioinformatics*, p. btx180.

Batchelor, P. G., A. C. Smith, D. L. G. Hill, D. J. Hawkes, T. C. S. Cox, and A. Dean (2002), Measures of folding applied to the development of the human fetal brain, *IEEE transactions on medical imaging*, *21*(8), 953–965.

Bender, M., R. Klein, A. Disch, and A. Ebert (2000), A functional framework for web-based information visualization systems, *IEEE Transactions on Visualization and Computer Graphics*, *6*(1), 8–23.

Biggiogera, M., and F. F. Biggiogera (1989), Ethidium bromide-and propidium iodide-pta staining of nucleic acids at the electron microscopic level., *Journal of Histochemistry & Cytochemistry*, *37*(7), 1161–1166.

Booth, P., W. Hall, N. Gibbins, and S. Galanis (2014), Visualising data in web observatories, in *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion*.

Bostock, M., V. Ogievetsky, and J. Heer (2011), D^3 data-driven documents, *IEEE transactions on visualization and computer graphics*, *17*(12), 2301–2309.

Bray, M.-A., et al. (2016), Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes, *Nature protocols*, *11*(9), 1757.

Brechbühler, C., G. Gerig, and O. Kübler (1995), Parametrization of closed surfaces for 3-d shape description, *Computer vision and image understanding*, *61*(2), 154–170.

Buslaev, A., A. Parinov, E. Khvedchenya, V. I. Iglovikov, and A. A. Kalinin (2018), Albumentations: fast and flexible image augmentations, *arXiv preprint arXiv:1809.06839*.

Bustin, M., and T. Misteli (2016), Nongenetic functions of the genome, *Science*, *352*(6286), aad6933.

Caicedo, J. C., C. McQuin, A. Goodman, S. Singh, and A. E. Carpenter (2018), Weakly supervised learning of single-cell feature embeddings, in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9309–9318.

Caicedo, J. C., et al. (2017), Data-analysis strategies for image-based cell profiling, *Nature methods*, *14*(9), 849.

Chen, H., J. Chen, L. A. Muir, S. Ronquist, W. Meixner, M. Ljungman, T. Ried, S. Smale, and I. Rajapakse (2015), Functional organization of the human 4D nucleome, *Proceedings of the National Academy of Sciences*, *112*(26), 8002–8007.

Cheng, L., L. Gao, W. Guan, J. Mao, W. Hu, B. Qiu, J. Zhao, Y. Yu, and G. Pei (2015), Direct conversion of astrocytes into neuronal cells by drug cocktail, *Cell research*, *25*(11), 1269.

Cheplygina, V., D. M. Tax, and M. Loog (2015), On classification with bags, groups and sets, *Pattern recognition letters*, *59*, 11–17.

Chiang, M., et al. (2015), Analysis of in vivo single cell behavior by high throughput, human-in-the-loop segmentation of three-dimensional images, *BMC Bioinformatics*, *16*, 397, doi:10.1186/s12859-015-0814-7.

Ching, T., et al. (2018), Opportunities and obstacles for deep learning in biology and medicine, *Journal of The Royal Society Interface*, *15*(141), doi:10.1098/rsif.2017.0387.

Choi, H.-J., and H.-K. Choi (2007), Grading of renal cell carcinoma by 3d morphological analysis of cell nuclei, *Computers in Biology and Medicine*, *37*(9), 1334–1341.

Chu, A., J. Cui, and I. D. Dinov (2009), SOCR analyses - an instructional java web-based statistical analysis toolkit, *J. Online Learn. Teach.*, *5*(1), 1–18.

Clarke, L. E., and B. A. Barres (2013), Emerging roles of astrocytes in neural circuit development, *Nature Reviews Neuroscience*, *14*(5), 311.

Cole, R. W., T. Jinadasa, and C. M. Brown (2011), Measuring and interpreting point spread functions to determine confocal microscope resolution and ensure quality control, *Nat Protoc*, *6*(12), 1929–1941.

Cremer, T., et al. (2015), The 4D nucleome: Evidence for a dynamic nuclear landscape based on co-aligned active and inactive nuclear compartments, *FEBS letters*, *589*(20), 2931–2943.

Davis, J., and M. Goadrich (2006), The relationship between precision-recall and roc curves, in *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, ACM.

de Chaumont, F., et al. (2012), Icy: an open bioimage informatics platform for extended reproducible research, *Nat Methods*, *9*(7), 690–6, doi:10.1038/nmeth.2075.

Dinov, I., et al. (2009), Efficient, distributed and interactive neuroimaging data analysis using the loni pipeline, *Frontiers in neuroinformatics*, *3*, 22.

Dinov, I. D. (2006), SOCR: Statistics online computational resource, *J. Stat. Softw.*, *16*(11).

Dinov, I. D. (2016), Methodological challenges and analytic opportunities for modeling and interpreting big healthcare data, *Gigascience*, *5*, 12.

Dinov, I. D. (2018), *Data Science and Predictive Analytics: Biomedical and Health Applications using R*, Springer.

Dinov, I. D., and N. Christou (2009), Statistics online computational resource for education, *Teach. Stat.*, *31*(2), 49–51.

Dinov, I. D., and N. Christou (2011), Web-based tools for modelling and analysis of multivariate data: California ozone pollution activity, *Internat. J. Math. Ed. Sci. Tech.*, *42*(6), 789–829.

Dinov, I. D., J. Sanchez, and N. Christou (2008), Pedagogical utilization and assessment of the statistic online computational resource in introductory probability and statistics courses, *Comput. Educ.*, *50*(1), 284–300.

Dinov, I. D., K. Siegrist, D. K. Pearl, A. Kalinin, and N. Christou (2016), Probability distributome: a web computational infrastructure for exploring the properties, interrelations, and applications of probability distributions, *Computational statistics*, *31*(2), 559–577.

Dinov, I. D., et al. (2010), Neuroimaging study designs, computational analyses and data provenance using the loni pipeline, *PLoS One*, *5*(9), doi:10.1371/journal.pone.0013070.

Dinov, I. D., et al. (2011), Applications of the pipeline environment for visual informatics and genomics computations, *BMC Bioinformatics*, *12*, 304, doi:10.1186/1471-2105-12-304.

Doan, M., et al. (2018), Label-free assessment of red blood cell storage lesions by deep learning, *bioRxiv*, p. 256180.

Du, C.-J., P. T. Hawkins, L. R. Stephens, and T. Bretschneider (2013), 3d time series analysis of cell shape using laplacian approaches, *BMC bioinformatics*, *14*(1), 296.

Ducroz, C., J.-C. Olivo-Marin, and A. Dufour (2012), Characterization of cell shape and deformation in 3d using spherical harmonics, in *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*, pp. 848–851, IEEE.

Dufour, A. C., T.-Y. Liu, C. Ducroz, R. Tournemenne, B. Cummings, R. Thibeaux, N. Guillen, A. O. Hero, and J.-C. Olivo-Marin (2015), Signal processing challenges in quantitative 3-d cell morphology: More than meets the eye, *IEEE Signal Processing Magazine*, *32*(1), 30–40.

Eckschlager, T., J. Plch, M. Stiborova, and J. Hrabeta (2017), Histone deacetylase inhibitors as anticancer drugs, *International journal of molecular sciences*, *18*(7), 1414.

Eliceiri, K. W., et al. (2012), Biological imaging software tools, *Nat Methods*, *9*(7), 697–710, doi:10.1038/nmeth.2084.

Ellenberg, J., J. Swedlow, M. Barlow, C. Cook, U. Sarkans, A. Patwardhan, A. Brazma, and E. Birney (2018), A call for public archives for biological image data., *Nature methods*, *15*(11), 849–854.

Fani, N., et al. (2013), Fkbp5 and attention bias for threat: associations with hippocampal function and shape, *JAMA psychiatry*, *70*(4), 392–400.

Ferri, M., and C. Gagliardi (1982), The only genus zero n-manifold is sn, *Proc. Amer. Math. Soc.*, *85*, 638–642, doi:10.1090/S0002-9939-1982-0660620-5.

Fitzgibbon, A., M. Pilu, and R. B. Fisher (1999), Direct least square fitting of ellipses, *IEEE Transactions on pattern analysis and machine intelligence*, *21*(5), 476–480.

Flach, P., and M. Kull (2015), Precision-recall-gain curves: PR analysis done right, in *Advances in Neural Information Processing Systems*, pp. 838–846.

Ganai, S. A., S. Malli Kalladi, and V. Mahadevan (2015), Hdac inhibition through valproic acid modulates the methylation profiles in human embryonic kidney cells, *Journal of Biomolecular Structure and Dynamics*, *33*(6), 1185–1197.

Gao, C., et al. (2018), Model-based and model-free machine learning techniques for diagnostic prediction and classification of clinical outcomes in parkinsons disease, *Scientific reports*, *8*(1), 7129.

Gao, L., et al. (2017), Direct generation of human neuronal cells from adult astrocytes by small molecules, *Stem cell reports*, *8*(3), 538–547.

Goldman, S. A. (2016), Stem and progenitor cell-based therapy of the central nervous system: hopes, hype, and wishful thinking, *Cell Stem Cell*, *18*(2), 174–188.

Gonzalez-Sandoval, A., and S. M. Gasser (2016), On tads and lads: spatial control over gene expression, *Trends in Genetics*, *32*(8), 485–495.

Göttlicher, M., et al. (2001), Valproic acid defines a novel class of hdac inhibitors inducing differentiation of transformed cells, *The EMBO journal*, *20*(24), 6969–6978.

Graham, B. (2014), Spatially-sparse convolutional neural networks, *arXiv preprint arXiv:1409.6070*.

Graham, B. (2015), Sparse 3d convolutional neural networks, *arXiv preprint arXiv:1505.02890*.

Guo, P. J., S. Kandel, J. M. Hellerstein, and J. Heer (2011), Proactive wrangling, in *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11*.

Guo, Z., L. Zhang, Z. Wu, Y. Chen, F. Wang, and G. Chen (2014), In vivo direct reprogramming of reactive glial cells into functional neurons after brain injury and in an alzheimers disease model, *Cell stem cell*, *14*(2), 188–202.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten (2009), The weka data mining software: An update, *SIGKDD Explor. Newsl.*, *11*(1), 10–18, doi:10.1145/1656274.1656278.

Han, X., C. Xu, and J. L. Prince (2003), A topology preserving level set method for geometric deformable models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(6), 755–768.

Heer, J., and M. Agrawala (2006), Software design patterns for information visualization, *IEEE Trans. Vis. Comput. Graph.*, *12*(5), 853–860.

Heinrich, C., et al. (2010), Directing astroglia from the cerebral cortex into subtype specific functional neurons, *PLoS biology*, *8*(5), e1000,373.

Held, M., M. H. Schmitz, B. Fischer, T. Walter, B. Neumann, M. H. Olma, M. Peter, J. Ellenberg, and D. W. Gerlich (2010), Cellcognition: time-resolved phenotype annotation in high-throughput live cell imaging, *Nat Methods*, *7*(9), 747–54, doi:10.1038/nmeth.1486.

Higgins, G. A., A. Allyn-Feuer, S. Handelman, W. Sadee, and B. D. Athey (2015), The epigenome, 4D nucleome and next-generation neuropsychiatric pharmacogenomics, *Pharmacogenomics*, *16*(14), 1649–1669.

Higgins, G. A., A. Allyn-Feuer, P. Georgoff, V. Nikolian, H. B. Alam, and B. D. Athey (2017), Mining the topography and dynamics of the 4D nucleome to identify novel cns drug pathways, *Methods*, *123*, 102–118, doi:10.1016/j.ymeth.2017.03.012.

Hsieh, J., K. Nakashima, T. Kuwabara, E. Mejia, and F. H. Gage (2004), Histone deacetylase inhibition-mediated neuronal differentiation of multipotent adult neural progenitor cells, *Proceedings of the National Academy of Sciences*, *101*(47), 16,659–16,664.

Huang, H., A. B. Tosun, J. Guo, C. Chen, W. Wang, J. A. Ozolek, and G. K. Rohde (2014a), Cancer diagnosis by nuclear morphometry using spatial information, *Pattern recognition letters*, *42*, 115–121.

Huang, H., A. B. Tosun, J. Guo, C. Chen, W. Wang, J. A. Ozolek, and G. K. Rohde (2014b), Cancer diagnosis by nuclear morphometry using spatial information, *Pattern Recognition Letters*, *42*, 115 – 121, doi:https://doi.org/10.1016/j.patrec.2014.02.008.

Husain, S. S., A. Kalinin, A. Truong, and I. D. Dinov (2015), SOCR data dashboard: an integrated big data archive mashing medicare, labor, census and econometric information, *Journal of big data*, *2*(1), 13.

Iglovikov, V. I., A. Rakhlin, A. A. Kalinin, and A. A. Shvets (2018), Paediatric bone age assessment using deep convolutional neural networks, in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 300–308, Springer International Publishing, Cham, doi:10.1007/978-3-030-00889-5_34.

Ince, D. C., L. Hatton, and J. Graham-Cumming (2012), The case for open computer programs, *Nature*, *482*(7386), 485–8, doi:10.1038/nature10836.

Ioffe, S., and C. Szegedy (2015), Batch normalization: Accelerating deep network training by reducing internal covariate shift, *arXiv preprint arXiv:1502.03167*.

Jagannathan, A. (2005), Segmentation and recognition of 3d point clouds within graph-theoretic and thermodynamic frameworks : a thesis, Thesis (ph.d.), Northeastern University, 2005.

Jang, S., and H.-S. Jeong (2018), Histone deacetylase inhibition-mediated neuronal differentiation via the wnt signaling pathway in human adipose tissue-derived mesenchymal stem cells, *Neuroscience letters*, *668*, 24–30.

Jevtić, P., L. J. Edens, L. D. Vuković, and D. L. Levy (2014), Sizing and shaping the nucleus: mechanisms and significance, *Current opinion in cell biology*, *28*, 16–27.

Kalinin, A., and D. Lisitsin (2011), Robust estimation of qualitative response regression models, in *Applied Methods of Statistical Analysis. Simulations and Statistical Inference. AMSA-2011*, pp. 303–309.

Kalinin, A. A., S. Palanimalai, and I. D. Dinov (2017), SOCRAT platform design: A web architecture for interactive visual analytics applications, in *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*, HILDA'17, pp. 8:1–8:6, ACM, New York, NY, USA, doi:10.1145/3077257.3077262.

Kalinin, A. A., B. D. Athey, and I. D. Dinov (2018a), Evaluation of methods for cell nuclear structure analysis from microscopy data, in *Supplementary Proceedings of the Seventh International Conference on Analysis of Images, Social Networks and Texts (AIST 2018)*, CEUR-WS.

Kalinin, A. A., G. A. Higgins, N. Reamaroon, S. Soroushmehr, A. Allyn-Feuer, I. D. Dinov, K. Najarian, and B. D. Athey (2018b), Deep learning in pharmacogenomics: from gene regulation to patient stratification, *Pharmacogenomics*, *19*(7), 629–650.

Kalinin, A. A., et al. (2018c), 3D shape modeling for cell nuclear morphological analysis and classification, *Scientific reports*, *8*, doi:10.1038/s41598-018-31924-2.

Kalinin, A. A., et al. (2018d), 3D cell nuclear morphology: Microscopy imaging dataset and voxel-based morphometry classification results, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2272–2280.

Kamentsky, L., et al. (2011), Improved structure, function and compatibility for cellprofiler: modular high-throughput image analysis software, *Bioinformatics*, *27*(8), 1179–80, doi:10.1093/bioinformatics/btr095.

Kandel, S., A. Paepcke, J. Hellerstein, and J. Heer (2011), Wrangler: Interactive visual specification of data transformation scripts, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 3363–3372, ACM.

Kankaanpaa, P., L. Paavolainen, S. Tiitta, M. Karjalainen, J. Paivarinne, J. Nieminen, V. Marjomaki, J. Heino, and D. J. White (2012), Bioimagexd: an open, general-purpose and high-throughput image-processing platform, *Nat Methods*, *9*(7), 683–9, doi:10.1038/nmeth.2047.

Keim, D., G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon (), Visual analytics: Definition, process, and challenges, in *Lecture Notes in Computer Science*, pp. 154–175.

Khairy, K., J. Foo, and J. Howard (2008), Shapes of red blood cells: comparison of 3d confocal images with the bilayer-couple model, *Cellular and molecular bioengineering*, *1*(2-3), 173.

Khan, F., V. Foley-Bourgon, S. Kathrotia, E. Lavoie, and L. Hendren (2014), Using JavaScript and WebCL for numerical computations, *ACM SIGPLAN Notices*, *50*(2), 91–102.

Koenderink, J. J., and A. J. Van Doorn (1992), Surface shape and curvature scales, *Image and vision computing*, *10*(8), 557–564.

Kortenhorst, M. S., S. Isharwal, P. J. van Diest, W. H. Chowdhury, C. Marlow, M. A. Carducci, R. Rodriguez, and R. W. Veltri (2009), Valproic acid causes dose-and time-dependent changes in nuclear structure in prostate cancer cells in vitro and in vivo, *Molecular cancer therapeutics*, *8*(4), 802–808.

Lam, H.-C., and I. D. Dinov (2012), Hyperbolic wheel: A novel hyperbolic space graph viewer for hierarchical information content, *ISRN Computer Graphics*, *2012*, 1–10.

Langan, T. J., and R. C. Chou (2011), Synchronization of mammalian cell cultures by serum deprivation, *Cell Cycle Synchronization: Methods and Protocols*, pp. 75–83.

Larson, R., and B. Edwards (2009), *Calculus*, 10th ed., Cengage Learning, Boston, MA.

LeCun, Y., Y. Bengio, and G. Hinton (2015), Deep learning, *Nature*, *521*(7553), 436.

Lévy, B. (2006), Laplace-beltrami eigenfunctions towards an algorithm that" understands" geometry, in *Shape Modeling and Applications, 2006. SMI 2006. IEEE International Conference on*, pp. 13–13, IEEE.

Li, H., and G. Chen (2016), In vivo reprogramming for cns repair: regenerating neurons from endogenous glial cells, *Neuron*, *91*(4), 728–738.

Li, L., Q. Zhou, T. C. Voss, K. L. Quick, and D. V. LaBarbera (2016), High-throughput imaging: Focusing in on drug discovery in 3d, *Methods*, *96*, 97–102, doi:10.1016/j.ymeth.2015.11.013.

Liaw, A., and M. Wiener (2002), Classification and regression by randomforest, *R news*, *2*(3), 18–22.

Maaten, L. v. d., and G. Hinton (2008), Visualizing data using t-sne, *Journal of machine learning research*, *9*(Nov), 2579–2605.

Mandelbrot, B. B. (1982), *The fractal geometry of nature*, vol. 173, 1st ed., W. H. Freeman and Company, New York.

Meijering, E., A. E. Carpenter, H. Peng, F. A. Hamprecht, and J.-C. Olivo-Marin (2016), Imagining the future of bioimage analysis, *Nature biotechnology*, *34*(12), 1250–1255.

Meyer, M., M. Desbrun, P. Schrder, and A. H. Barr (2003), *Discrete Differential-Geometry Operators for Triangulated 2-Manifolds*, pp. 35–57, Springer Berlin Heidelberg, Berlin, Heidelberg, doi:10.1007/978-3-662-05105-4_2.

Meyer, M. G., M. Fauver, J. R. Rahn, T. Neumann, F. W. Patten, E. J. Seibel, and A. C. Nelson (2009), Automated cell analysis in 2d and 3d: A comparative study, *Pattern Recognition*, *42*(1), 141–146.

Meyer, M. G., J. W. Hayenga, T. Neumann, R. Katdare, C. Presley, D. E. Steinhauer, T. M. Bell, C. A. Lancaster, and A. C. Nelson (2015), The Cell-CT 3-dimensional cell imaging technology platform enables the detection of lung cancer using the noninvasive LuCED sputum test, *Cancer cytopathology*, *123*(9), 512–523.

Montanaro, L., D. Treré, and M. Derenzini (2008), Nucleolus, ribosomes, and cancer, *The American journal of pathology*, *173*(2), 301–310.

Montani, M. S. G., M. Granato, C. Santoni, P. Del Porto, N. Merendino, G. DOrazi, A. Faggioni, and M. Cirone (2017), Histone deacetylase inhibitors vpa and tsa induce apoptosis and autophagy in pancreatic cancer cells, *Cellular Oncology*, *40*(2), 167–180.

Montero, R. S., and E. Bribiesca (2009), State of the art of compactness and circularity measures, in *International mathematical forum*, vol. 4, pp. 1305–1335.

Moon, S. W., I. D. Dinov, S. Hobel, A. Zamanyan, Y. C. Choi, R. Shi, P. M. Thompson, A. W. Toga, and A. D. N. Initiative (2015), Structural brain changes in early-onset alzheimer's disease subjects using the loni pipeline environment, *Journal of Neuroimaging*, *25*(5), 728–737.

Niethammer, M., M. Reuter, F. E. Wolter, S. Bouix, N. Peinecke, M. S. Koo, and M. E. Shenton (2007), Global medical shape analysis using the laplace-beltrami spectrum, *Med Image Comput Comput Assist Interv*, *10*(Pt 1), 850–7.

Niu, W., T. Zang, Y. Zou, S. Fang, D. K. Smith, R. Bachoo, and C.-L. Zhang (2013), In vivo reprogramming of astrocytes to neuroblasts in the adult brain, *Nature cell biology*, *15*(10), 1164.

Object Refinery Limited (2017), JFreeChart, `http://www.jfree.org/jfreechart`.

Ollion, J., J. Cochennec, F. Loll, C. Escude, and T. Boudier (2013), Tango: a generic tool for high-throughput 3d image analysis for studying nuclear organization, *Bioinformatics*, *29*(14), 1840–1, doi:10.1093/bioinformatics/btt276.

Osmani, A. (2011), Patterns for large-scale javascript application architecture, `https://addyosmani.com/largescalejavascript/`.

Osmani, A. (2012), *Learning JavaScript Design Patterns: A JavaScript and jQuery Developer's Guide*, O'Reilly Media, Inc.

Pantanowitz, L., F. Preffer, and D. C. Wilbur (2018), Advanced imaging technology applications in cytology, *Diagnostic cytopathology*, doi:10.1002/dc.23898.

Papanicolaou, G. N., and H. F. Traut (1941), The diagnostic value of vaginal smears in carcinoma of the uterus, *American Journal of Obstetrics & Gynecology*, *42*(2), 193–206.

Pau, G., F. Fuchs, O. Sklyar, M. Boutros, and W. Huber (2010), Ebimage–an r package for image processing with applications to cellular phenotypes, *Bioinformatics*, *26*(7), 979–81, doi:10.1093/bioinformatics/btq046.

Pedregosa, F., et al. (2011), Scikit-learn: Machine learning in python, *Journal of Machine Learning Research*, *12*, 2825–2830.

Pegoraro, G., and T. Misteli (2016), High-throughput imaging as a versatile and unbiased discovery tool, *Methods*, *96*, 1–2, doi:10.1016/j.ymeth.2016.01.003.

Peng, H., A. Bria, Z. Zhou, G. Iannello, and F. Long (2014), Extensible visualization and analysis for multidimensional images using vaa3d, *Nat Protoc*, *9*(1), 193–208, doi:10.1038/nprot.2014.011.

Peng, T., and R. F. Murphy (2011), Image-derived, three-dimensional generative models of cellular organization, *Cytometry A*, *79*(5), 383–91, doi:10.1002/cyto.a.21066.

Pincus, Z., and J. A. Theriot (2007), Comparison of quantitative methods for cell-shape analysis, *Journal of Microscopy*, *227*(2), 140–156, doi:10.1111/j.1365-2818.2007.01799.x.

Rakhlin, A., A. Shvets, V. Iglovikov, and A. A. Kalinin (2018), Deep convolutional neural networks for breast cancer histology image analysis, in *International Conference Image Analysis and Recognition*, pp. 737–744, Springer.

Ramo, P., R. Sacher, B. Snijder, B. Begemann, and L. Pelkmans (2009), Cellclassifier: supervised learning of cellular phenotypes, *Bioinformatics*, *25*(22), 3028–30, doi:10.1093/bioinformatics/btp524.

Royer, L. A., M. Weigert, U. Günther, N. Maghelli, F. Jug, I. F. Sbalzarini, and E. W. Myers (2015), Clearvolume: open-source live 3d visualization for light-sheet microscopy, *Nature methods*, *12*(6), 480.

Ruder, S. (2016), An overview of gradient descent optimization algorithms, *arXiv preprint arXiv:1609.04747*.

Saito, T., and M. Rehmsmeier (2015), The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets, *PloS one*, *10*(3), e0118,432.

Santal, L. A. (2004), *Integral geometry and geometric probability*, 2nd ed., Cambridge University Press, Cambridge, UK.

Satyanarayan, A., D. Moritz, K. Wongsuphasawat, and J. Heer (2017), Vega-lite: A grammar of interactive graphics, *IEEE Transactions on Visualization and Computer Graphics*, *23*(1), 341–350.

Schindelin, J., C. T. Rueden, M. C. Hiner, and K. W. Eliceiri (2015), The imagej ecosystem: An open platform for biomedical image analysis, *Molecular Reproduction and Development*, *82*(7-8), 518–529, doi:10.1002/mrd.22489.

Schindelin, J., et al. (2012), Fiji: an open-source platform for biological-image analysis, *Nature methods*, *9*(7), 676–682.

Schneider, C. A., W. S. Rasband, and K. W. Eliceiri (2012), Nih image to imagej: 25 years of image analysis, *Nat Methods*, *9*(7), 671–5.

Seaman, L., W. Meixner, J. Snyder, and I. Rajapakse (2015), Periodicity of nuclear morphology in human fibroblasts, *Nucleus*, *6*(5), 408–416.

Seo, S., and M. K. Chung (2011), Laplace-beltrami eigenfunction expansion of cortical manifolds, in *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, pp. 372–375, IEEE.

Shen, L. (2010), Spharm-mat, `http://www.iu.edu/~spharm/`.

Shen, L., and F. Makedon (2006), Spherical mapping for processing of 3d closed surfaces, *Image and vision computing*, *24*(7), 743–761.

Shi, Y., R. Lai, J. H. Morra, I. Dinov, P. M. Thompson, and A. W. Toga (2010), Robust surface reconstruction via laplace-beltrami eigen-projection and boundary deformation, *IEEE transactions on medical imaging*, *29*(12), 2009–2022.

Shvets, A., V. Iglovikov, A. Rakhlin, and A. A. Kalinin (2018a), Angiodysplasia detection and localization using deep convolutional neural networks, in *Machine Learning and Applications (ICMLA), 2018 17th IEEE International Conference on*, IEEE.

Shvets, A., A. Rakhlin, A. Kalinin, and V. Iglovikov (2018b), Automatic instrument segmentation in robot-assisted surgery using deep learning, in *Machine Learning and Applications (ICMLA), 2018 17th IEEE International Conference on*, IEEE.

Simonyan, K., and A. Zisserman (2014), Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556*.

Singh, S., F. Janoos, T. Pécot, E. Caserta, K. Huang, J. Rittscher, G. Leone, and R. Machiraju (2011), Non-parametric population analysis of cellular phenotypes, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 343–351, Springer.

SOCR (2018), DVT: Dynamic visualization toolkit, `http://socr.umich.edu/HTML5/DViewer/`.

Solovyev, R. A., A. A. Kalinin, A. G. Kustov, D. V. Telpukhov, and V. S. Ruhlov (2018), FPGA implementation of convolutional neural networks with fixed-point calculations, *arXiv preprint arXiv:1808.09945*.

Steed, C. A., K. J. Evans, J. F. Harney, B. C. Jewell, G. Shipman, B. E. Smith, P. E. Thornton, and D. N. Williams (2014), Web-based visual analytics for extreme scale climate science, in *2014 IEEE International Conference on Big Data (Big Data)*.

Stephens, A. D., E. J. Banigan, and J. F. Marko (2018a), Separate roles for chromatin and lamins in nuclear mechanics, *Nucleus*, *9*(1), 119–124.

Stephens, A. D., P. Z. Liu, E. J. Banigan, L. M. Almassalha, V. Backman, S. A. Adam, R. D. Goldman, and J. F. Marko (2018b), Chromatin histone modifications and rigidity affect nuclear morphology independent of lamins, *Molecular biology of the cell*, *29*(2), 220–233.

Stolte, C., D. Tang, and P. Hanrahan (2002), Polaris: A system for query, analysis, and visualization of multidimensional relational databases, *IEEE Transactions on Visualization and Computer Graphics*, *8*(1), 52–65.

Su, Z., W. Niu, M.-L. Liu, Y. Zou, and C.-L. Zhang (2014), In vivo conversion of astrocytes to neurons in the injured adult spinal cord, *Nature communications*, *5*, 3338.

Tang, M., C. Gao, S. A. Goutman, A. Kalinin, B. Mukherjee, Y. Guan, and I. D. Dinov (2018), Model-based and model-free techniques for amyotrophic lateral sclerosis diagnostic prediction and patient clustering, *Neuroinformatics*.

Terzopoulos, D. (1988), The computation of visible-surface representations, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *10*(4), 417–438.

Thompson, P. M., C. Schwartz, R. T. Lin, A. A. Khan, and A. W. Toga (1996), Three-dimensional statistical analysis of sulcal variability in the human brain, *J Neurosci*, *16*(13), 4261–74.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Pearson College Division.

Uhler, C., and G. Shivashankar (2017), Regulation of genome organization and gene expression by nuclear mechanotransduction, *Nature Reviews Molecular Cell Biology*, *18*(12), 717.

Uhler, C., and G. Shivashankar (2018), Nuclear mechanopathology and cancer diagnosis., *Trends in cancer*, *4*(4), 320–331.

US-CERT (2013), Oracle java contains multiple vulnerabilities. alert (TA13-064A).

van der Walt, S., J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu (2014), scikit-image: image processing in python, *PeerJ*, *2*, e453, doi:10.7717/peerj.453.

Veltri, R. W., and C. S. Christudass (2014), Nuclear morphometry, epigenetic changes, and clinical relevance in prostate cancer, in *Cancer Biology and the Nuclear Envelope: Recent Advances May Elucidate Past Paradoxes*, edited by E. C. Schirmer and J. I. de las Heras, pp. 77–99, Springer New York, New York, NY.

Verdone, J. E., P. Parsana, R. W. Veltri, and K. J. Pienta (2015), Epithelial–mesenchymal transition in prostate cancer is associated with quantifiable changes in nuclear structure, *The Prostate*, *75*(2), 218–224.

Ververis, K., A. Hiong, T. C. Karagiannis, and P. V. Licciardi (2013), Histone deacetylase inhibitors (hdacis): multitargeted anticancer agents, *Biologics: targets & therapy*, *7*, 47.

Wadell, H. (1935), Volume, shape, and roundness of quartz particles, *The Journal of Geology*, *43*(3), 250–280.

Wang, L.-L., and C.-L. Zhang (2018), Engineering new neurons: in vivo reprogramming in mammalian brain and spinal cord, *Cell and tissue research*, *371*(1), 201–212.

White, M. J. D. (1977), *Animal cytology and evolution*, CUP Archive.

Wikipedia (2018), Gamma function, `https://en.wikipedia.org/wiki/Gamma_function`.

Wilbur, D. C., M. G. Meyer, C. Presley, R. W. Aye, P. Zarogoulidis, D. W. Johnson, N. Peled, and A. C. Nelson (2015), Automated 3-dimensional morphologic analysis of sputum specimens for lung cancer detection: Performance characteristics support use in lung cancer screening, *Cancer cytopathology*, *123*(9), 548–556.

Wilson, E. B. (1925), *Cell In Development And Heredity, 3rd. Rev*, Macmillan Company.; New York.

Wongsuphasawat, K., D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer (2016), Voyager: Exploratory analysis via faceted browsing of visualization recommendations, *IEEE transactions on visualization and computer graphics*, *22*(1), 649–658.

Wongsuphasawat, K., Z. Qu, D. Moritz, R. Chang, F. Ouk, A. Anand, J. Mackinlay, B. Howe, and J. Heer (2017), Voyager 2: Augmenting visual analysis with partial view specifications, in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM.

Xu, D., J. Cui, R. Bansal, X. Hao, J. Liu, W. Chen, and B. S. Peterson (2009), The ellipsoidal area ratio: an alternative anisotropy index for diffusion tensor imaging, *Magnetic resonance imaging*, *27*(3), 311–323.

Yang, X., and E. Seto (2007), Hats and hdacs: from structure, function and regulation to novel strategies for therapy and prevention, *Oncogene*, *26*(37), 5310.

Zhang, L., et al. (2015), Small molecules efficiently reprogram human astroglial cells into functional neurons, *Cell stem cell*, *17*(6), 735–747.

Zheng, G., A. A. Kalinin, I. D. Dinov, W. Meixner, S. Zhu, and J. W. Wiley (2018), Hypothesis: Caco-2 cell rotational 3d mechanogenomic turing patterns has clinical implications to colon crypts, *Journal of Cellular and Molecular Medicine*, doi:10.1111/jcmm.13853.

Zink, D., A. H. Fischer, and J. A. Nickerson (2004), Nuclear structure in cancer cells., *Nature reviews cancer*, *4*(9).