Title
Automated, High-throughput Analysis of Neurite Dynamics in Neurodegenerative Disease

Permalink
https://escholarship.org/uc/item/1kd747s8

Author
Ando, Dale Michael

Publication Date
2015

Peer reviewed|Thesis/dissertation

Automated, High-throughput Analysis of Neurite Dynamics in
Neurodegenerative Disease

by

D. Michael Ando

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biomedical Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Dedication and Acknowledgements

I am thankful for the many individuals who have contributed to my development during graduate training. I have been mentored by and worked with phenomenally intelligent, knowledgeable, and motivated individuals who have made the years fly by. There are several people who I would like to highlight for their significant role in my development.

The first person I would like to thank is my PI and mentor, Steve Finkbeiner. Steve has supported me through my highly non-linear path through graduate school from cell biology to bioimage informatics. Steve's focus on helping his students find their own path has been key to helping me merge my background in physics and my desire to tackle biomedical problems. Steve's advice has also been instrumental in helping me forward, through qualifying exams, a successful fellowship application, and an internship at Google.

The second person I'd like to thank is Eva Ladow. Eva is a fantastic scientist who taught and advised me innumerable times when I was working at the bench, and she has directly contributed to my graduation by carrying out the ALS-TDI screen, which I've used for most of my work. By virtue of being the first large-scale experiment on Roboticscope III, Eva was also part of a meta-experiment on how complete and robust the automation was, and I'm grateful for her understanding as we worked the kinks out of the system.

I also need to thank the two scientists who built the previous generation robotic microscope, Aaron Daub and Punita Sharma. Aaron shared my vision for automation in the aid of biology, and I enjoyed his thoughtful perspective on both science and technology. Punita took me under her wing after I just joined the lab, and my first introduction to robotic microscopy was writing a review paper together. From my discussions with them on the growing pains of their robotic microscope and image analysis pipeline, I was able to learn a lot about what was needed for the new microscope and analysis systems.

Abstract

Despite decades of research, there are no effective therapies for neurodegenerative diseases such as amyotrophic lateral sclerosis (ALS) and Parkinson's Disease (PD). These diseases are marked by a progressive loss of the neuronal processes referred to as neurites and the cell bodies. One obstacle to developing therapies is the uncertainty in translating improvements in disease model system into humans. The measures in disease models cannot match the scope of measures in a patient, and they are further limited by practical considerations on the number of timepoints and measurements to explore a larger space of possible therapeutic interventions. The failure of the previous approaches makes it important to find a new approach with the promise of more relevant measures. We have developed a fully-automated system for high-throughput robotic microscopy and an automated image analysis pipeline for our longitudinal datasets that incorporates supervised machine learning. We use this system to extend our measurements to include the quantification of neurite dynamics. The positional accuracy of our robotic microscopy combined with supervised machine learning enables us to measure neurite extension and retraction over time. In an ALS model, we have shown that the decrease in neurite area is mainly due to an increase in neurite retraction. We compared our measures of neurite dynamics to cell survival and demonstrated that measures of neurite dynamics are superior at separating the healthy and disease conditions, and they can quantify differences between the conditions that would not be detectable using survival analysis. We investigated the ability of measures of neurite dynamics to distinguish between an ALS and PD disease model. While accuracy of the PD model was lower than the ALS model, the parameters that contributed to the accuracy were very differently ranked, indicating there is not a single useful neurite measure for differentiating diseases. We believe the robustly and longitudinally detecting neurites will give us valuable insights into mechanistic and therapeutic studies of neurodegeneration.

Table of Contents

List of tables

List of figures

Chapter 1: Introduction

Patients with neurodegenerative diseases lack treatments that significantly delay or halt the progression of the disease. Multiple factors contribute to the difficulty and risk in creating effective therapies for these diseases; however, a critical factor is the uncertainty of translating mechanisms found in disease models to effective therapies in humans (Finkbeiner, 2010). One reason for this uncertainty is that the biology of the disease model may be fundamentally different than that of the human disease due to differences in species, how the disease is introduced into the model, or in the local environment of the cell, including its cellular neighbors. Because we cannot easily, and nondestructively, peer into the molecular arrangement of living humans, we must overcome the inherent uncertainty of disease models. We can do this by validating that the measurements from the disease model correspond to features of the disease in humans. Our lab has validated several models of neurodegenerative disease, including Huntington's disease (HD) (Arrasate et al., 2004), Parkinson's disease (PD) (Skibinski et al., 2014), and amyotrophic lateral sclerosis (ALS)/frontotemporal lobar dementia (FTLD) (Barmada et al., 2010). For example, our model for HD demonstrated features of the disease that were later confirmed *in vivo* (Miller et al., 2010).

Another reason for this uncertainty is that the simpler biology of a model compared to a human patient restricts our ability to assess measures of disease. This reason is especially true for neurodegenerative diseases, which involve the most complex organ, the brain. In specific regions of the brain, distinct neuronal subpopulations tend to have an exaggerated vulnerability that causes cognitive and motor deficits characteristic of each disease (Roselli and Caroni, 2015; Saxena and Caroni, 2011). Even in animal models that contain organs with complex functions and interactions, we can only take surrogate measures for a subset of sign in human neurodegenerative diseases. Thus, the uncertainty lies in how well a measure in a model translates to human disease. This uncertainty is exacerbated by the basic measures used in

almost all large-scale studies of compounds or genes that rescue a disease phenotype. Typically, studies in *in vitro* models only take snapshots of fixed cells at a single timepoint of one or a few measurements, such as cell viability, protein aggregation, or organelle function (Jain et al., 2012; Varma et al., 2008; Zhang et al., 2014). While investigators would likely prefer a richer set of measurements to increase confidence that they are treating the root cause of the disease rather than a particular sign, more measures cost experimental time and/or money. Thus, they often resort to the barebones approach to find drugs and genes in search of an effective therapy.

Despite decades of research, the barebones approach has been ineffective at identifying a successful therapy, which motivates efforts to find better ways to measure disease. A valuable measurement for quantifying relationships between initial conditions and eventual fate is time. We can use time to resolve cellular responses that start at different times during disease progression. By studying the relationship between the response magnitude and cell fate, we can determine whether the response is pathogenic or compensatory. In searching for better measures of disease models, we ideally would obtain longitudinal measures that help us quantify the course of the disease, including the early stages when we likely have more leverage to prevent its progression.

Our lab has pioneered a high-throughput longitudinal microscopy platform that evaluates and follows single cells over long time courses lasting days to weeks (Arrasate and Finkbeiner, 2005). With this platform, we can capture the full range of a disease-associated phenotypes, from the early stages to cell death. For example, by following a single cell over time, we can use the same statistics as clinical trials to determine whether factors positively or negatively impact cell health, and then rank them by effect size. This platform settled a controversial debate in HD with its first discovery. Originally, researchers did not know if inclusion bodies in HD were related to worse, the same, or better outcomes for the cell (Ross, 1997; Saudou et al., 1998; Sisodia, 1998). With our platform, we demonstrated that the inclusion bodies were rather a coping

response that improved the survival of neurons expressing mutant huntingtin compared to cells that maintained a diffuse distribution of the mutant protein (Arrasate et al., 2004). Since that first discovery, we have improved the automation of our platform and sophistication of our statistical analysis and live-cell reporters to increase the power of the system (Daub et al., 2009; Sharma et al., 2012; Skibinski and Finkbeiner, 2013).

Here, we will describe my work that focused on increasing the richness of longitudinal measures that we can gather without affecting experimental time. Chapter 2 describes the fully-automated robotic microscope we developed to enable high-throughput, longitudinal imaging. Chapter 3 describes the automated image analysis pipeline we developed to handle the large number of images from our robotic microscopes. This pipeline incorporates supervised machine learning to identify difficult to quantify features such as dim neurites and enables novel dynamic neurite measures. Chapter 4 describes the application of our microscope and analysis pipeline to the patterns of disease caused by TAR DNA-binding protein 43 (TDP-43), which is involved in amyotrophic lateral sclerosis (ALS), based on a primary screen funded by the ALS Therapy Development Institute (ALS-TDI). We also investigate how our results may extrapolate to other neurodegenerative diseases.

Chapter 2: Roboticscope III

Overview of Robotic Microscopy

Robotic microscopy (RM) differs from other forms of automated microscopy in one key way: it returns to image precisely the same location at each timepoint for experiments that can span days to weeks. Imaging the same location over time allows us to follow individual cells, enabling powerful statistics for relating a cell's state to its fate. The relationships uncovered using this longitudinal analysis would be challenging or impossible to uncover with single snapshot experiments. Thus, RM is a powerful tool for understanding disease mechanism and developing therapeutics.

RM enables longitudinal imaging without the sit-and-stare approach, which requires the same plate to stay on the microscope for the duration of the experiment. The sit-and-stare approach would be too low-throughput for survival studies that typically last a week. RM only requires the plate be on the microscope during imaging, and the plate is stored in a cell culture incubator between imaging timepoints. A naive implementation of this approach would have little or no overlapping area present in all images across the experimental timecourse. Substantial displacement can be introduced by the positioning of the microtiter plate in the stage's plate holder and hysteresis in the mechanical stage over time. We need the same physical origin for the stage at each timepoint to ensure that the stereotyped movements of the automated stage return to the same physical locations.

The first generation robotic microscope achieved this by requiring a manual alignment step before the start of imaging each timepoint. A scientist would move the stage until the image of a stable mark on the microtiter plate, termed the fiducial image, was superimposed on the fiducial image from the first timepoint. The second generation RM had three main innovations over the first: a PerfectFocus system that was superior in reliability and speed to software-based

autofocus, an environmental chamber that held temperature and CO2 during imaging, and a robotic arm that could access several microtiter plate stacks that were enclosed within the environmental chamber. However, it still relied upon manual alignment before the start of each timepoint, and it lacked humidity within the environmental chamber, limiting its ability to do fully-automated long-term imaging.

Evolving RM Experimental Needs

The first and second generation RMs were focused on survival analysis of primary rodent neurons where the presence of the neuron's cell body was tracked daily for approximately a week. Several developments in the lab spurred the development of the next generation RM. The first is that scientists were performing more experiments that required higher temporal resolution than was necessary to measure cell survival, such as studies measuring protein or organelle half-life. These were not feasible on the run-times on the first and second generation robotic microscopes that required approximately an hour to image a single plate. The second is that the number of experiments was increasing as more external academic and industry collaborators desired time on the instrument. Scientists adapted to this increased experimental load by multiplexing experiments with different imaging routines on the same plate, but this required a work-around where the plate was effectively run again for each different imaging routine on the plate. The third is that we were moving toward IPSC models where we needed to capture more area to get the appropriate number of correct-looking cells for analysis.

Thus a new solution was required that could support the many new needs: faster imaging of single plates, flexible imaging intervals of seconds to daily, and experiments lasting days to weeks. flexible imaging requirements so that each well's conditions could be completely independent, and the ability to start imaging a plate without a human operator present.

Design Considerations for Roboticscope III

A major tension in designing a microscope system is the inevitable trade-off between speed, resolution, and sensitivity. Improving the speed of imaging a single plate was a driving force behind the creation of the new RM. An analysis of time spent imaging during a run on the second generation RM showed that the majority was spent moving a component of the system into the correct position (e.g. filter wheel or stage) with the second largest time going to the cumulative time for image exposure. High resolution imaging is important for neurons, especially the processes and subcellular organelles. The processes of neurons provide a lot of information about the health status of the neuron, and their presence or absence often determines whether a scientist will call a cell dead or alive. This prohibits us from just imaging at lower magnification because we wouldn't have the resolution to correctly call when a cell dies, the critical factor even for survival analysis. Improving the sensitivity was also becoming more of a driving concern because of the new half-life experiments. These studies suffer from low signal since only a small fraction of the total photoswitchable protein population can be switched without adversely affecting cell health. We were also moving to fluorescent proteins (FP) in the blue and far red portion of the spectrum in order to increase the number of channels, and FPs at those wavelengths have much lower intrinsic brightness.

Another practical concern was the software used to control the microscope. We desired the ability to program in a commonly used programming language to make it easier to get experienced programmers. We also valued having access to the source code in the event we needed to make a modification to support our experiments that may be difficult to convince a company to implement.

Description of mechanical components for Roboticscope III

Camera- The camera is a back-illuminated Andor iXON 888 electron multiplying (EM) CCD that provides six advantages over the previous generation's camera: increased sensitivity

over a broad range of the spectrum, low read noise at high read speeds, reduced thermal noise, a larger field of view, a deeper pixel well depth, and increased bit-size of the images. The back-illuminated CCD has >90% QE (vs ~60% QE for the previous generation camera), which improves signal for dim fluorescent signals. The EM gain reduces the read-noise 4-5 fold even at the fastest read noise that both improves S/N and throughput. The cooling to -80C (vs -30C for the previous generation camera) significantly reduces thermal noise. The physical size of the sensor for the Andor 888 is roughly 3 times larger than for the previous camera with an equal gain in the FOV. Since we have identified stage and component movement as the main bottleneck in imaging throughput, reducing the number of movements by 1/3 substantially decreases experiment run time. The combination of a deeper well depth (80,000 vs 16,000 for the previous generation camera) and high digitization bit-size (14 bit vs 12bit for the previous generation camera) increases both the dynamic range and contrast of the images, important properties for intensity-based image analysis.

A downside is that the pixel size of the Andor iXON 888 EMCCD is 13 microns by 13 microns, four-fold the size of the previous generation camera. While we recognized that this would lead to a reduction in resolution, we decided that the experiments that require higher resolution could be performed with an objective with higher magnification and NA with the trade-off of less imaging area captured.

Microscope Base- We upgraded our microscope base from the Nikon TE-2000 to the Ti-E that provides three key advantages: improved PerfectFocus, optimal placement of the emission filters in the optical path, and faster component movements.

The PerfectFocus has been shrunk and integrated into the nosepiece unit for the objectives and the laser wavelength used for alignment has been shifted from 770nm to 870nm. Shifting the laser to a longer wavelength has increased the useable red-spectrum for fluorescent imaging, a valuable property for multiplexing fluorescent proteins given the emergence and development of far-red fluorescent proteins. Integrating into the nosepiece has freed up space

to move the emission filter wheel into the infinity space of the microscope that was previously taken up by the PerfectFocus. Emission filters are designed to be in the infinity space as the bandpass peak will shift depending on the angle of incident light on a filter that means that the bandpass may differ for light from the edge on an image as compared to the center. Mechanical improvements in the speed of changing components such as emission wheels, filter wheels, and objectives as well in reduction in communication overhead between the microscope controller and components will increase throughput compared to the older TE-2000

Objectives- The standard 20x and 40x ELWD objectives were updated with improved chromatic correction in the near-ultraviolet and near-infrared portions of the spectra, another improvement for the multiplexing of fluorescent proteins. We integrated a non-ELWD objective, the 20x Plan Apochromat VC objective, with a substantially higher numerical aperture than the 20x ELWD objective. This allows for a roughly 8-fold higher image brightness and 2-fold increase in resolution, both very valuable properties for dim fluorescent reporters of biological activity and organelle-specific fluorescent proteins.

Filter System- We upgraded these to be hard-coated for greater longevity.

Light Source- We have upgraded from the Lambda LS to the Lambda XL light source that provides three key advantages: longer bulb life, higher maximum lamp intensity, and controllable lamp output. The previous bulbs have an average bulb lifetime of 1000 hours, which is just 41 days when imaging 24/7 as our automated system allows. The new Lamba XL bulbs last 10 fold longer lifetime that will mean less downtime for the microscope, fewer abrupt changes in lamp intensity due to bulb replacement, and significant cost-savings.

Robotic Arm- We are using a robotic arm with the same capabilities as before. We are modifying the gripper to work with the microscope in a new manner.

Hotel Incubator- We have upgraded from microtiter plate stacks to a Liconic automated hotel incubator with three advantages: random access of plates, improved environmental control for plate-storage and integration of a bar-code reader. Random access of plates is useful for

imaging experiments that occur at different biological timescales (e.g. several hours vs several days) and thus need different imaging intervals. An automated hotel incubator allows us to store the plates at high humidity, which was previously impossible due to electronics malfunctions at a humidity higher than 50%. This allows us to do automated imaging runs over even longer timespans. We also integrated a bar-code reader that is valuable for ensuring that the correct plate with the correct imaging method is run and is useful for organizing downstream data analysis.

Environmental Chamber- Our environmental chambers will hold temperature and CO_2 equivalent to the hotel incubator. Imaging runs are short enough that evaporation is not a significant concern during imaging.

Host Computer- Our host computer has been upgraded to a 64-bit operating system from a 32-bit and has been upgraded with faster central and graphic processing power, larger RAM and storage drive memory, and faster networking capabilities. In addition to being able to reliably handle the large amount of data generated from the new microscope, the increased processing power allows for on-the-fly analysis as supported by our new control program discussed below.

Description of the High-Throughput Screening (HTS) Plugin

The previous generation's procedural script built within a proprietary imaging program was replaced with multi-threaded plugin that integrated with the open-source Micro-Manager software (Edelstein et al., 2014). The plugin is written in Java rather than a niche language. The HTS Plugin communicates with GreenButtonGo (Biosero, San Diego, CA), the laboratory automation scheduling software that coordinated the microscope with the robotic arm and automated incubator.

The major goal of the HTS plugin was automating the steps that normally required a human to start an imaging run. The crucial step for automation was the alignment of the fiducial

images before each run. For the opaque 96-well microtiter plates we used, we determined that the intersection of four wells was a good fiduciary feature. We used a pyramid intensity-based algorithm in the StackReg ImageJ plugin to perform the alignment (Thevenaz et al., 1998). This was superior to a feature-based algorithm because the features visible on an opaque plate would be from the dynamic cells in the well rather than a rigid structure on the plate.

Imaging parameters for experiment were encoded in a comma-delimited file called the Imaging Template. This file allowed the user to specify the wells, channels, exposures and EM gain for the experiment as well as the date and time for each timepoint. The order of wells was determined by the order of the wells in the Imaging Template, enabling any arbitrary pattern across the plate. The channel order was determined by semi-colon separated values where each channel could have its own exposure and EM gain setting. Because each well could have a distinct set of channels, exposures, and EM gains, it is now easy to multiplex experiments on a single 96-well plate.

Another option to increase the throughput of the system was the ability to set whether all of the stage positions of a certain channel were imaged before switching channels. This was helpful because switching emission filters and potentially polychroic cubes was quite slow. Imaging all stage positions in a channel first minimized the amount of time spent switching components. However, it could also introduce small misalignments between images in different channels due to slew in the mechanical stage. Thus, this was set as an option in case absolute pixel alignment was required. Because higher throughput was the driving goal, the default was imaging all stage positions in a channel first.

Previously the components had been controlled serially so the minimum delay between consecutive images was the sum of the delays of moving all of the separate components. This could add significant time to an experiment multiplexing three fluorescent proteins and imaging a large area in each well. The new control program now simultaneously controls each component such that the delay is equal to the delay of the slowest component, usually stage

movement. This makes it reasonable to gather significantly more area per well and thus increase the number of cells sampled per condition.

Retrospective of Completed System

The completed third generation RM we call Roboticscope III has been operating since Fall 2012. We have our goal of a fully-automated system where the scientist can put a plate in the incubator and instruct the system to perform the appropriate imaging routines at the specified times. We have achieved our goal of high-throughput: the system can image a 96-well plate with one channel over 11% of the area of each well in 15 minutes. This allows us to capture 83,000 images in a single day. Roboticscope III has supported both basic science as well as screening projects, generating terabytes of data in the process. The HTS plugin that was developed for Roboticscope III is also being using by Roboticscope II and Roboticscope IV, representing the first time the same microscope control software has been used on multiple RM. All of the subsequent images analyzed in my dissertation are from Roboticscope III.

Chapter 3: Automated Image Analysis Pipeline of Neurons

Introduction

The adage that "a picture is worth a thousand words" captures the imbalance between the ease of capturing images and the difficulty of fully describing them. This imbalance has been exacerbated by advances in automated microscopy, which has created a flood of data that requires automated techniques to analyze—the amount of time needed for a scientist to determine the outcome of their experiment would be prohibitively slow for scientific progress. Large amounts of data also create unique problems for reliable human analysis. Humans get tired and inattentive over long periods of monotonous work, analysis rubrics can differ between scientists working on the data in parallel, and even the same scientist may have an evolving rubric based on increased familiarity with the types of objects in the images. With an automated solution, we can remove the need for human analysis and study biological findings in our microscopy images with a rapid, unbiased, and consistent technique.

In the Finkbeiner lab, we have developed a unique technology called robotic microscopy, which tracks individual cells to relate their initial state to their resulting fate. We have applied this technology to understanding neurodegenerative disease. Microscopy images are a rich source of information on neuronal characteristics that are encoded within the shape, position, and fluorescence intensity of the cell. By tracking individual neurons over days to weeks, we can expand that information to include dynamic measurements. Unfortunately, by imaging live cells longitudinally, we face distinct challenges in accurately and consistently extracting biological details from the images, largely due to limits on increasing the intensity of dim objects and variations in object intensity over time. Furthermore, creating an algorithm to extract biological details is deceptively difficult due to the ease with which humans recognize objects (Meijering,

2012). By overcoming these challenges to automating image analysis, we can increase the accuracy, sensitivity, and sophistication of our robotic microscopy platform.

Within our lab, we have developed several software programs that automate some or all of the neuron image analysis. The first fully automated analysis pipelines that we created focused on neuron survival (Daub, 2013; Sharma et al., 2012). Briefly, these pipelines had three stages, as shown in Figure 1A. The first stage pre-processed the images by correcting the background, montaging the tiles, and then aligning the montages. The second stage identified the neuron regions using histogram adjustment, image filters, and thresholding. The third stage tracked the neuron regions over time to specify the lifetime of each neuron. These automated pipelines for survival analysis offered a tremendous time advantage over manual or semi-automated analysis, but they suffered from three key deficiencies. The first is that the tracking algorithm had a strict requirement on overlap of the cell body between timepoints. Even though neurons are largely immobile, this requirement may not be met, which would lead to a premature end of the neuron track and an underestimate for the lifespan of that neuron. The second is that the algorithms could confuse cellular debris with a live cell body. For example, algorithms that solely use intensity to identify neurons may identify bright fluorescent debris as a neuron, which would lead to incorrect continuation of the neuron track and an overestimate for the lifespan of that neuron. The third is that these pipelines did not conveniently record how the scientists curated the experimental datasets. This record would be valuable for training new scientists to identify biological objects in the images and for comparing rubrics between scientists or over time.

Improving automated survival analysis has clear benefits for the traditional experiments with robotic microscopy. However, a larger advancement would be moving beyond measuring just the cell body and toward a higher dimensional space where measurements are taken simultaneously. This higher dimensional space can capture a more detailed picture of the cellular status of our models of health and disease. For example, beyond cell survival, we can

13

measure the shape of axons and dendrites to define the state of the neuron. These axons and dendrites, collectively referred to as neurites, form neural circuits, and neurite dynamics, such as extension or retraction, are important for understanding normal and dysregulated function. By taking automated measures of neurite extension and retraction with longitudinal imaging, we can gain insight that normally requires impractical and tedious manual analysis.

Our automated pipeline for image analysis that we describe here takes a probabilistic approach based on machine-learning models trained from ground-truth data provided by scientists. In contrast to the hand-tuned approach of the previous pipeline, our new pipeline provides a modular approach for identifying and quantifying new features in biological images. We've successfully applied this pipeline to the difficult challenges of identifying dim neurites and live neurons in our longitudinal imaging datasets. With this pipeline, we can provide the basis for creating a rich set of relevant biological measurements for defining the fingerprint of cell state.

Enlighten: Using supervised machine learning to analyze neurons imaged longitudinally

The impetus for Enlighten arose when we recognized that the same root problem was shared between the confusion of cellular debris with a live cell body and the extension of our measurements beyond just the cell body: correctly classifying each pixel. With techniques from the field of machine learning, we can identify the best set of parameters to classify pixels out of the large number of possible parameters. In contrast, humans are slow and ineffective at searching through the possible parameters, which has limited the success of more historical methods for automating image analysis. Enlighten uses supervised machine learning (SML) to accurately classify previously unseen pixels from examples where the class (e.g., neuron, neurite) of a pixel is known. The new workflow of Enlighten is shown in Figure 1B.

SML has its own set of challenges, including its need for a large number of annotated examples. We developed two pieces of software to generate labeled training data: one for

labeling each pixel in a small patch from an image and one for labeling regions while curating survival analysis. With the resulting training examples, we select an appropriate machine-learning algorithm and train a model for use in our analysis pipeline. SML also has the long-term benefit of holding the potential for continual improvement. To evaluate the performance of the SML model on a new experiment, the scientist can label a small number of patches. Thus, each new experiment can contribute new training examples, and these examples can accumulate over time to train better models.

From a computational perspective, our experiments exhibit high data parallelism. For example, every experiment has at least main divisions. First, an experiment may be composed of multiple 96-well microtiter plates, and second, each plate can be divided into individual wells. Each division creates a self-contained unit of data that can be processed without external input, a necessary criteria for parallel computation. Thus, if sufficient computing resources are available for computing all experimental wells at once, then an entire experiment can be processed in the same amount of time it takes to process a single well.

Our current unit of analysis is at the well level. The image data for a well is from several timepoints that vary between several hours to a day over an experimental timeframe that can last a week or more. For each timepoint, there are images in one to three fluorescence channels. These images are uploaded with experimental metadata about the source and history of the cells in the experiment, any chemical or genetic manipulations done on each well, and the image acquisitions parameters. At present, these metadata are entered by the experimenter at the start of the experiment, but we are looking toward a future where we can automatically generate the metadata with by further integrating robotics in the culturing of the cells. By pairing the image with metadata, we can create an expanding database of healthy and disease-model neuron images. This database could be used for training future SML models or re-analyzing experiments as the algorithms that extract biological features improve.

Background Correction and Normalization

Microscopy images often have inhomogeneous illumination patterns due to dust and uneven illumination. Large sensors are especially prone to uneven illumination because of the difficulty in uniformly filling the field-of-view with light. We can correct for these illumination patterns using the equation below, which converts the raw, inhomogeneous images to corrected flat-field images.

$$\text{Equation 1:} \quad Corrected = (Raw - Darkfield)/(Brightfield - Darkfield)$$

The darkfield image has the same imaging parameters as the raw image, but without illumination. This corrects for artifacts such as thermal noise; however, with our EMCCD camera, thermal noise only negligibly contributes, so we ignore this correction.

The brightfield image has the same imaging parameters as the raw image, but there should be no objects in the imaging plane. We don't capture this image, because we use every well on our microtiter plate. Instead, we can use our imaging dataset to approximate this image. To do this, we consider that our neuron images never have 100% pixel occupancy. This means that for each pixel location, several images have a background pixel intensity value at that same location. We can then use a quantile value calculated for each pixel position from all, or a subset, of the images for a channel and timepoint. We have found a subset of 100 images and the 5th quantile work well in approximating the brightfield image. Finally, we apply a smoothing function, a mean filter with a radius of five pixels, on the resulting brightfield image. This approach reduces local variations in pixel intensity that occur on a spatial scale smaller than the scales of both dust and uneven illumination.

We divide the raw image by the brightfield image for the appropriate timepoint and channel with a 32-bit float precision rather than the original 16-bit integer range. We convert the

32-bit float images back to 16-bit integer range based on two practical issues: less software support for 32-bit float tiff images compared to 16-bit tiff images and the doubled image size of 32-bit float images. We perform this conversion by scaling the 32-bit float values using a scaling factor that is constant for a channel, which ensures that images from all timepoints are scaled to the same range. The scaling factor is a product of the background intensity for that channel at the first timepoint and the fold-difference in the bit depths of the image (16-bits) and the camera (14-bits). This approach ensures that the full 16-bit range keeps as much precision from the 32-bit images as possible.

Probability Images: Representations of our understanding of our images

Enlighten diverges from the previous pipelines by extensively using probability images in which each pixel value represents our confidence in the class (e.g., foreground, cell body, neurite) for that pixel. These probability images replace the images of freely diffusible fluorescent protein (FDFP) that highlight the morphology of the neuron and show the large dynamic range of their intensities (Figure 2A). For example, the cell bodies of neurons are very bright, often at the edge or even above the dynamic range of the camera. The borderline, however, requires that neurons be over-exposed, because the signal from neurites is often little above the noise floor. This dramatic difference in signal is driven by the large difference in volume for the FDFP between the large, circular cell body and the long, thin neurites.

By using probability images, Enlighten improves the performance of the subsequent stages of the analysis pipeline, such as montaging, aligning, and, of course, quantifying biological features. These improvements are largely influenced by two neurite-related factors: the dramatic enhancement in signal-to-background for dim features, as shown in Figure 2B, and the ability to classify neurite pixels, as shown in Figure 2C. Neurites dominate the image area compared to cell bodies, so we can improve both the montaging and aligning stages by having more points to compare between two images.

Enlighten is also unique because it preserves the probability image rather than using a threshold, such as the traditional cut-off of 50%, to create a binary mask. For dim features, such as neurites, in which the measurement values are near the noise floor, it is difficult to consistently boost the pixel probability above 50%. If the neurite confidence fluctuates close to the threshold, one could see large swings in the measured neurite area for small shifts in neurite confidence. Thresholding also limits the amount of information available for downstream analysis, such that a rejected pixel in a binary image may have had 49.9% or 0% probability with a 50% threshold. By preserving the probability images, we can use a hierarchical approach to segment images as shown in Figure 3A. We can also compute the probability image for neurons using measurements on filters applied to both the original fluorescence-intensity image and the neurite-probability image. The advantages of probability images for neurites will be further discussed and quantified in the next chapter.

This new probabilistic approach to our image-analysis pipeline led us to reassess how even simple segmentation decisions are made, such as the distinction between what is foreground or background. Each decision is made by a machine-learning model that was trained using labeled examples from scientists. This approach has tremendous advantages over hand-tuned segmentation algorithms. For example, the same workflow for training models is applicable when creating models for progressively more complex features. Thus, this approach gives us a modular means of achieving our two main goals: accurately extracting the biologically relevant features to define signatures of cell state and continually improving our ability to identify and measure a wider range of biological features. The next section will describe the individual steps that we follow to achieve these goals (Figure 4A).

Pixel Labeling

The most crucial piece of software that supports the probabilistic approach is a pixel labeler. We developed a web-based labeler for the pipeline (see screenshot in Figure 5A), which

provides basic information about pixel coordinates and the label names that are needed for the SML techniques to find the distinguishing criteria for each label. Because we are doing pixel-level predictions and starting from scratch on training examples, we used "dense" labeling, in which every pixel in the labeling region should have a label. Dense labeling is advantageous because it forces the scientist to label the boundaries between objects. However, scientists are likely to disagree on the exact pixels that constitute the boundary between regions of different classes (e.g., neurite and background). The supervised machine-learning model will learn the range of expert beliefs to best classify each pixel.

Each montage of a single well at a single timepoint contains several million pixels, and it would take an exorbitantly long time to densely label each one. For a given number of training examples, we would also prefer to capture variations in images across time, experiments, and conditions rather than oversample from a few images. To minimize oversampling, we randomly label several square patches from each image. The dimensions of the square patch can vary depending on the frequency of the feature to be labeled. For instance, examples of foreground and background are quite common, so a small patch is likely to provide pixels for both categories. On the other hand, features such as debris or neuron bodies are less common, so we use a larger patch and relax the requirement that the labeling be dense throughout the entire patch. This pragmatic approach ensures that the scientist doesn't excessively label a single category or skip over a large number of less-informative image patches.

We have found that, especially for small patches, it is valuable to also include the adjacent area of the patch, although this area is not intended for labeling. Scientists may need this adjacent to place the pixels of the patch into the appropriate context, because they are often accustomed to looking at larger scales. Restricting the available contrast options can also be helpful. For example, the high dynamic range in the image data cannot be fully represented on a monitor, which is particularly problematic when trying to label dim features. Allowing the scientist full control over contrast could lead to inconsistencies in labeling if different scientists choose

different contrast levels. To keep the time for labeling each patch low, the number of labeling options is also limited. At a minimum, the options are foreground or background, and at maximum, the options are background, neurite, neuron body, and debris.

The pixel labeler produces an output of several folders containing images. The first folder holds the original fluorescence-intensity images, which are required for computing the filtered images that train the SML model. The other folders represent the possible labels. Within each label folder is a binary image where the non-zero pixels indicate whether that pixel is an example of that label.

Model Training

Probability images can be generated with a variety of machine-learning models. We chose a logistic-regression model for several reasons. First, the model is easy to share. It can be described as a single equation of parameters and their associated weights, plus a potential intercept term. Thus, a table of the parameters, their weights, and the intercept is sufficient to replicate the model. Second, the filtered images that make up the parameters do not all need to be present in RAM (random access memory) at the same time when evaluating the pixel probability. The equation can be evaluated term-by-term by storing the partial sum in an "accumulator" image. Thus, the RAM requirements can be limited to the accumulator image and the filtered image that is being calculated, regardless of the number of parameters in the model. Third, there are established methods for reducing the logistic-regression model to the minimum number of parameters needed for accurate predictions. Fewer parameters will result in faster computation of probability images. Thus, we can try to train the model with a large number of filters without worrying about an excessive evaluation time during analysis.

The first task for model training is generating the measurements from the filter images. The filters and the variable ranges for each filter used for Enlighten are shown in Table 1. These filters were chosen because they are commonly used in computer vision and they can respond

to line-like objects, such as neurites. The variable ranges for the filters were chosen to approximate the range of object dimensions in our images and keep the training dataset a reasonable size. The size of the training dataset is represented by the equation below:

Equation 2:

$$\text{Training Dataset Size} = (\#\text{ of pixels}) * (\#\text{ of input images}) * (\#\text{ of filter possibilities})$$

This size can rapidly grow, especially if care is not taken to contain the number of filter possibilities. For instance, if we consider the difference of Gaussian filters when each radius could range between 1 and 100, the possible number of filters is 4950. Then, if each measurement has 32-bit precision, and we have two input images and 1 million pixels, then the dataset size is ~37GB.

Table 1. Filters applied to input images during model training

Filter	Variable Range
Gauss	Sigma: 1, 3, 5, 7, 9, 11, 13, 15, 17, 19
Difference of Gaussians	Sigma_Low: 1, 3, 5, 7, 9, 11, 13, 15, 17, 19
	Sigma_High: 3, 5, 7, 9, 11, 13, 15, 17, 19
	Low must be lower than High
Mean of Laplacian of Gaussians	Sigma_Low: 1, 3, 5, 7, 9, 11, 13, 15, 17, 19
	Sigma_High: 3, 5, 7, 9, 11, 13, 15, 17, 19
	Low must be lower than High
Scharr	N/A
Gabor	Sigma: 1, 3, 5, 7, 9, 11, 13, 15, 17, 19
(min, max, max-min, mean, SD, CV)	Lambda: 2, 6, 10, 14, 18, 22, 26, 30, 34, 38
	Gamma: 0.1, 0.5, 1.0

	Psi: 0, 0.5

Before calculating a filtered image, we first apply the natural log to the intensity of each pixel, which decreases the separation between the extremely high pixel values and the rest of the pixel values so that the histogram more closely approximates a normal bell curve. This approach has two benefits for model training. The first is that we want a model that is robust to variation in the input images. Because all pixel values in our images are greater than 1, the nominal difference between pixel values on a logarithmic scale will be smaller than the difference on a linear scale. This difference shrinks the range of variation between images. The second is that we want a model that distinguishes between features at different extremes of the dynamic range, such as neuron cell bodies and neurites. By shrinking the extreme values, all features will be in a more similar intensity range.

To create the training examples for model training, the filter measurements are combined with the label information from the pixel labeler. To minimize the size of the training dataset, it is important to only keep filter measurements for the small area of each image that is annotated. As the number of examples grows, it may also be necessary to fit only a subset of the training data into RAM.

The training data are formatted in columns for the label data and for each filter measurement respectively. Because the input image could be the original fluorescence image and/or a probability image, the column name for the filter measurements includes a prefix to denote the input image to which the filter was applied. For instance, original_gauss_4 would be a gaussian blur of radius 4 applied to the original fluorescence image, and probforeground_dog_1_4 would be a difference of gaussians (dog) of radius 1 and 4 on the foreground probability image.

Before training a model, the training dataset must be split into two subsets: one for training the model and the other for testing it. The test dataset ensures that the model

accurately calculates the training subset without "memorizing," or overfitting, the training subset. Overfitting is especially problematic for models with many parameters. We use an 80%/20% split for the training/testing subsets. When splitting the data, we ensure that the training examples in the two subsets are as independent as possible. Otherwise, the testing subset may not detect overfitting of the model. For this reason, we forced all annotations from a specific well, regardless of timepoint, to be part of the same training or testing subset. In the future, when there are more annotations available, all wells from a specific experimental plate should also be forced into the same subset.

Another way to minimize the chance of overfitting, while keeping the final analysis time short, is to constrain the number of parameters in the model. However, determining the minimum set of filters is challenging. We have broadly selected filters and variable ranges that should be informative, but for any set of filters, we need to empirically determine the optimal variable settings. We can't directly use these settings with all filters included to determine the most important filters. For instance, the magnitude of the weight is not a reliable proxy for how important the filter is, because the influence of the filter is the product of the weight and the intensity values in the filtered image. An important filter could have very large intensity values and thus need a smaller weight to "normalize" its contribution. This classic conundrum is referred to as variable selection, and it can be solved by finding a method that shrinks insignificant parameters to zero so they can be distinguished from important parameters with small weights. We examined two methods for variable selection: logistic regression with an L1 penalty and markov chain monte carlo (MCMC) simulations with a spike-slab prior.

We used the glmnet package in R to perform logistic regression with an L1 penalty (Friedman et al., 2010) (Friedman et al., 2015). In contrast to an L2 penalty that penalizes the square of the weights, the L1 penalizes the absolute value of the weights. Given a certain "budget" for the weights, the L2 penalty tends to shrink the large weights the most, while the L1 penalty will operate equally on both large and small weights. Thus, L1-penalized models will

tend to have more terms at zero. The pure L1-penalty model did not converge for the foreground vs background model. Glmnet is unique because it incorporates an "elastic net" that allows different ratios of L1 and L2 to incorporate as the penalty. Applying 90% L1 and 10% L2 allowed our model to converge. The resulting model successfully removed many of the terms, but could still leave several hundred terms. We could further reduce the number of terms by increasing the L1/L2 penalty, but at the cost of decreased accuracy.

Because Enlighten needs several models to isolate features such as neurites or neurons, we need to further shrink the number of parameters in the models. As an alternate to glmnet, we examined a spike-slab prior with MCMC using the BoomSpikeSlab package in R (George and Mcculloch, 1997; Scott, 2014; Tüchler, 2008). A spike-slab prior has a high value with no variance at zero as the "spike" and a low-amplitude gaussian with high variance centered at zero as the "slab." This prior captures the belief that most terms we put into the model are not important and should have a weight of zero. However, we don't want all weights to be zero, so the slab allows the terms with sufficient evidence to have a non-zero posterior weight. This approach corresponds to our desire for a concise model. Although the training for spike-slab took longer than glmnet, the resulting models consistently had equivalent accuracy with fewer terms. These results may be due to the co-linearity of many of the filter measurements: filters with very similar variable settings will have similar intensity values. The small amount of L2 penalty would tend to group these correlated filters, thereby increasing the number of filters with non-zero weights. Spike-slab can grab just one representative filter from a group of related filters, keeping accuracy high without introducing extraneous filters. Spike-slab is the method we used for training all of the logistic regression models for Enlighten.

Alignment and Montaging

Aligning images across time and stitching them into a montage are critical steps for longitudinal analysis. The per-pixel measurements of neurite dynamics described in the next

chapter only heighten the importance of these steps. By misaligning neurite regions via inaccurately stitching a montage or aligning images from consecutive timepoints, a positive pixel in one image could appear negative in the next (or vice versa) simply due to misregistration. The automated algorithm could erroneously mistake this for neurite extension or retraction. We made several improvements in Enlighten to minimize the potential impact of these issues.

We align adjacent images in a montage and images from consecutive timepoints with an intensity-based algorithm in OpenCV (Bradski and Kaehler, 2008). This algorithm minimizes the squared difference between the two images as shown below:

$$R(x, y) = \sum_{x', y'} (T(x', y') - I(x + x', y + y'))^2$$

The intensity-based algorithms performed better than feature-based measures, such as the scale-invariant feature transform (SIFT) on our images. We believe this effect is due to the lack of rigid objects in our images and the similarity in features across the image. Even when there wasn't translation between two timepoints, the feature-based method would erroneously detect an offset.

We align images between adjacent timepoints as an example of how we perform intensity-based alignment. Here, we treat the first timepoint image as the starting point for alignment, and then each later image is aligned to the most recent earlier image. Before alignment, we trim a large border from the later image and a small border from the earlier image. In Enlighten, we use border widths of 225 and 25 pixels for the later and earlier images, respectively. The trimmed version of the later image is then slid across the trimmed version of the earlier image, and the squared difference is calculated for each position of the later image on the earlier image. With the center of the trimmed earlier image and the position of the later image, where the squared difference is a minimum, we calculate the offset of the later image to

the earlier image. We then shift all of the pixels of the later image by this offset. Pixels that are shifted outside of the dimensions of the image do not contribute to the final image, and pixels that do not have a new value shifted to their position are left with a value of zero. After the images from all time points are aligned, the stack of aligned images is trimmed to remove all zero-valued borders.

In the previous pipeline, a source of errors was from using the fluorescence-intensity images of the freely diffusible fluorescent protein for the intensity-based alignment. For accurate alignment, the objects we use to align must reliably be in the same position over time. The fluorescence intensity of reliable features, such as neurites, is quite low, while the intensity of unreliable features, such as cellular debris, is quite high. Even bright objects such as living cell bodies are not ideal, because there are only 30–50 present in an image montage, and they may move or die. Because neurite pixels are the most common type of foreground pixel in our images, we now use the neurite-probability images for alignment in Enlighten. The neurite probability images have a substantially improved signal-to-background ratio, so our alignment can be performed using a reliable feature with numerous bright pixels.

The sensitivity of the per-pixel measurements of neurite dynamics also revealed new issues. By inadvertently rotating the camera axes with respect to the microscope stage axes, we could create an angle in the overlap between adjacent images. This angle can introduce small shifts in the calculated position for each image in a montage, and small shifts contribute to erroneous measures of neurite dynamics. Because this montage error is difficult to detect, we introduced a per-tile analysis stage before the montaging step to capture neurite dynamics without this possible error (Figure 1C). This per-tile stage of analysis includes alignment, calculation of probability images, and neurite measurements. The pipeline also includes the alignment, neurite measurements, and cell tracking after montaging. The duplication of alignment and neurite measurements allows us to chose which measurement to use depending on whether we see a rotational mis-alignment in our images without having to rerun the analysis

or make due with the smaller overlap area of per-tile alignment. While not currently implemented, the alignment step after montaging could use the alignment offsets calculated for the per-tile alignment as a starting point.

Segmentation of Neurons

We can detect pixels that belong to neuron cell bodies with the SML approach described above. However, the neuron-probability images are not ideal for defining the area of the neuron region, because the SML approach cannot classify all pixels within the cell body with equal confidence. Instead, neuron regions in the neuron-probability images are characterized by a fluctuating higher probability at the edges of the cell and a lower probability at the center. The neurite attachment points on the boundary of the cell body are the best indicator that the bright region is not debris, which increases the neuron probability along edge sections of the neuron cell body near neurites. The center of the neuron cell body may look similar to the center of debris, which decreases the probability that the program assigns the center of the neuron of being a neuron. Thus, by thresholding the neuron probability image, we may obtain several small regions or a U-shaped region rather than capturing the full extent of the neuron's cell body.

Fortunately, we do not need to use the neuron-probability images to define the neuron area. Instead, we can use a probability image that will fully capture the neuron area, the non-neurite-probability image. This image is composed of the circular regions of cell bodies and debris. Then, with the regions that result from thresholding the non-neurite-probability image, we can assess the number of pixels within the region that are highly likely to belong to a neuron. We then use the mean value of neuron probability within the region to assign a probability that the region is a neuron.

Neuron Tracking

Tracking cells is a hard problem that is often solved by increasing imaging frequency. However, more frequent imaging will also decrease the number of experiments that can be performed in a day. Fortunately, neurons don't move very much; but they can sometimes be found close to each other, so even small movements can confound the results. Additionally, the longer timescales of our longitudinal analysis can add some complications. For example, some neurons start expressing the fluorescent protein long after the start of the experiment. Neurons also die over time. Thus, if two neurons are visible at the earlier timepoint and only one at the later timepoint, then the tracking algorithm must be able to confidently determine which one died and which one survived, or if two cells died and a new cell appeared. The sensitivity and accuracy of the survival analysis rests upon the ability of the pipeline to accurately track a cell to its true fate in a crowded image where cells are moving, dying, or appearing.

The previous pipeline performed tracking by requiring that the cell body had at least 80% overlap across two timepoints. This criterion was true most of the time, but it could easily truncate measurements of lifespans prematurely. This algorithm also assumed that all of the objects were neurons. We improved upon the previous algorithm in two stages.

The first improvement was to relax the requirement of a strict overlap and move to a greedy algorithm for tracking neurons. This algorithm matched an object at the earlier timepoint to an object at a later timepoint using a hand-tuned similarity metric based on how far away the objects were and the difference in their areas as shown in the equation below:

Equation 3: $Tracking\ Score = (20,000 / Distance) + (10,000 / Area\ Difference)$

This algorithm sequentially removes the best matches to stop that object from potentially matching with other objects. Because not all objects may be matched between two timepoints

due to appearance or death, the tracking score is set to zero for any objects that are greater than 100 pixels from each other. However, this algorithm is limited because it still assumes that all of the objects were neurons, which we improved with the second stage.

The second improvement introduced the possibility that a region was not a neuron. As discussed above, neuron-probability images enable Enlighten to assign a probability that a region is a neuron. In addition, we can capitalize on our longitudinal data by integrating information before and after each timepoint to help predict cell types and track continuity. This approach is most useful for borderline cases in which the cell-type probability is close to a user-chosen threshold, such as 50%. For instance, an object could have a 55% probability of being a neuron on the first and third timepoints, but a 49% probability on the second timepoint. If we had to assign cell type without looking at the adjacent timepoints, we would naively say that the second timepoint object is debris, breaking the track. However, we do not need to use a threshold on cell type before determining cell tracks. Instead, we can use a combinatorial optimization algorithm that includes the cell-type probability when determining the optimum matches between two timepoints. To account for the appearance of new neurons, the best match must be above an empirically-derived probability for new neurons of 0.5%. The equations for the probability that an object in an earlier timepoint is linked to an object in a later timepoint are given below:

$$\text{Equation 4:} \quad P(Debris\,after)=1-P(Neuron\,after)$$

$$\text{Equation 5:} \quad Same\,Type=P(Neuron\,before)*P(Neuron\,after)+P(Debris\,after)$$

$$\text{Equation 6:} \quad \beta=\ln(Same\,Type)+(-2.095908*\ln(Distance)) \\ +(-1.794061*\ln(Area_{after}/Area_{before}))+9.371554$$

$$\text{Equation 7:} \quad Probability\,of\,Link=e^{\beta}/1+e^{\beta}$$

In our previous example, the probability of a link between the objects in the second and third timepoints will only be slightly lower due to the decrease in the magnitude of *Same Type* rather than a broken track. Thus, we can use our cell-type predictions from both timepoints to inform the link probability.

We use the Hungarian algorithm to determine the best matches between two timepoints (Albalate and Minker, 2011). To prevent jumping, we also introduce a distance cut-off at 100 pixels, so that objects that are further apart have a link probability of zero.

Type and Track Annotation

Given the importance of accurately tracking survival analysis, we also wanted to collect data on how scientists would correct the type and tracks predicted by Enlighten. Our goal was to build a database of neuron tracks that could be used to test if changes to Enlighten would improve its tracking predictions. To accomplish this, we built a web-based tool that labels survival tracks as shown in Figure 5A. In this tool, the user is presented with a table of regions found by the automated segmentation. They can then select the correct type for an object, break a track if it has been connected to the incorrect object in a subsequent timepoint, and create a link to an object in the next timepoint if the track has ended prematurely.

By virtue of being web-based, our annotation application can be used by many different neuroscientists, potentially simultaneously, to correlate results from images across many different experimental and imaging conditions. However, correcting every neuron track in an experiment is an unreasonable burden on the scientist. We came up with two methods that reduce the amount of manual annotation required per-experiment. The first method is to use a "spot-checking" approach in which only a few wells of each condition are densely labeled. The results from these wells would ensure that the accuracy of the automated pipeline was similar enough to human annotation for the automated analysis to be considered valid for that experiment. This approach also supports a steady accumulation of cell type and track

information that can be useful for training or testing future algorithms. The second method is to automatically recalculate the ideal tracks after the user breaks or creates a link. Correcting a cell track can have a waterfall effect. For example, when two cells (cells A and B) are present in the previous timepoint, and one cell (cell C) is present in the next timepoint, then blocking the link between cells A and C will increase the probability that the cells B and C do have a link. Rather than forcing the user to both break the incorrect link and form the correct link, we can use our prediction algorithm to try to form the correct link and omit any link the user specified was invalid. A similar procedure is followed when the user sets a link to an object: the user-specified link is set as valid, any previous link to the object is invalidated, and the prediction algorithm resolves any other potential changes.

To keep the task of annotating a well manageable, we selected a minimum set of annotation labels shown in Table 2. By restricting the possible labels to the broad categories we are interested in (e.g., health status, cell type), the user won't need to annotate every object in the image, which will reduce the total time needed for annotation.

Two states which could be set for any class were also available for labeling (Table 3). The certainty of the labeling is also important to capture. For example, in instances of cell death, it can be challenging to know precisely when a cell has changed from living to dead. One could imagine only wanting to train based on really certain examples, or at least to include it as a weighting. A caveat is that the uncertain state has an ambiguity about whether the person would be uncertain about cell type or health status. This ambiguity can be highlighted for particularly interesting examples with the important state, the second state available for any class.

For our minimum set of annotation labels, we focused on our immediate goal of matching human survival. Here, we considered two broad distinctions: does this region contain a single living/dead cell, or is there a reason this region should be excluded? We'd prefer to reject rather than give a label that would dilute the "purity" of a category. Some reasons to

exclude a region could be multiple neurons in a mask, a blurry image, or incorrect montaging that created a fractured view of a neuron.

Table 2. Classification options available for labeling cell regions

Label	Meaning
Neuron Body	This region contains a single living neuron body containing the nucleus.
Neurites	This region contains only a neurite and not the cell body of a neuron. This label is useful for demonstrating when the segmentation picked up a bright neurite rather than a cell body. This label can be removed as a category and grouped into the forget label if this quality control is not desired.
Glial	This region contains a single living glial body containing the nucleus. While glial cells are rare, this label can help ensure that an anomalously high number of glial are not present in an experiment, which could change the outcome. This label can be removed as a category and grouped into the forget label if this quality control is not desired.
Debris	This mask contains all or a part of a dead neuron body.
Forget	Exclude this region, because it would not be useful training data.

Table 3. Toggle options available for labeling cell regions

Toggle	Meaning
Uncertain	The user is unsure of this classification.
Important	Highlights example regions for later review.

Validation with human annotation

To validate neuron tracking, we could see the appropriate divergence in the risk of death between healthy controls and several disease models, including Hungtinton's Disease (HD) and

amyotrophic lateral sclerosis (ALS). The ALS model will be discussed further in the next section. For the HD experiments, we replicated the previous results of poly-glutamine length-dependent increases in cumulative hazard, and dose-dependent increases in cumulative hazard of the mutant form of huntingtin (Figure 6A, C, D). The results from the automated analysis also corresponded well to the human-annotation of the same experiment (Figure 6 B, E).

Anomalous Image Discovery

Images with bubbles or drugs that are not colorless can create regions with unusually high fluorescence intensity. On the other hand, empty or out-of-focus wells will have lower intensity images. By identifying these outlier wells, we can better determine conditions that should be excluded from analysis and/or re-run to obtain higher quality images. However, in an experiment that can have tens of thousands of images, or a screen that can have hundreds of thousands images, we can't feasibly inspect all of the images manually. Instead, we can automate the discovery of images with anomalously high or low intensity fluorescence, which has two valuable benefits. For screens, automation mitigates the concern of false negatives due to poor image quality rather than a true absence of biological effect. For both screens and basic science, automation can reduce loss of time and valuable experimental data due to the presence of a mechanical issue with the microscope that causes failures in imaging.

We also discovered an unanticipated benefit of probability images: they provide a means of automatically flagging unusual images. In particular, they help us define outliers using a different feature than we are measuring for the assay. For instance, the neurite-probability image is a useful way to determine neurite dynamics, although it can be fooled by bright anomalies such as bubbles or colored drugs. However, we do not want to use high neurite area to detect bright anomalies, because a bubble in a well with a few neurites may have a similar area to a well with a lot of neurites. Thus, by using the same measure to both find true hits and

33

remove anomalies, we introduce a potential trade-off between sensitivity of hits and selectivity against erroneous results.

Enlighten helps us bypass this trade-off, because we have probability images for other features. The models for both neurites and non-neurite probabilities can be fooled by bright anomalies. Thus, we can use the non-neurite-probability image to identify bright anomalies without sacrificing the wells with high neurite area. By looking at the change in area of the non-neurite-probability image at different confidence thresholds, the anomalous images were distinguished via a high slope (Figure 7A). The bright anomaly images had the sloped region at the higher confidence-threshold values (Figure 7B), while the dim anomaly images had the sloped region at the lower-confidence threshold values (Figure 7C). We interpret this high slope as showing that the model does not understand how to segment those images, because it was not trained on images with those anomalies. An extension to using the non-neurite-probability images would be to train a model to directly recognize each anomaly.

Chapter 4: Neurite dynamics during neurodegeneration

ALS, TDP-43, and Neurites

ALS is a disorder that largely affects the motor neurons that control voluntary movement in the cerebral cortex and the spinal cord. Nearly ~20,000–30,000 people in the US are affected by ALS, and each year, another ~5,000 people are diagnosed. This makes it the most common motor-neuron disease. People of all ages can be affected, but most are between 40- and 70-years-old, with an average age of 55 years. ALS is fatal, and death normally occurs 3–years after the onset of symptoms. As patients lose muscle control, their diaphragm and chest wall muscles can atrophy, leading to respiratory failure. Present treatments are largely centered on alleviating symptoms of the disease. The only exception is riluzole, an FDA-approved drug that decreases glutamate signaling. This drug only extends survival in patients by 2–3 months and comes at a relatively high price (Miller et al., 2012). A more effective therapy for ALS is clearly needed.

Much of our knowledge of the molecular mechanisms of ALS comes from studying genes with mutations linked to autosomal dominant ALS. Unfortunately, the five most common genes with ALS-linked mutations only cover ~15% of the total cases (Lattante et al., 2015); however, they form the basis for our disease models due to their monogenic nature. One gene with a common mutation linked to ALS is TDP-43 (Rutherford et al., 2008), whose pathology is found in almost all ALS patients. Thus, TDP-43 may affect disease progression even without a genetic mutation (Arai et al., 2006; Mackenzie et al., 2007; Neumann et al., 2006). For example, wild-type TDP-43 appears to have intrinsic-adverse properties that are aggravated by mutations, including a tendency to cause toxicity by aggregating (Johnson et al., 2009) or mislocalizing from its normal position in the nucleus into the cytoplasm (Barmada et al., 2010), and also by decreasing mitochondrial density and function. (Wang et al., 2013)

ALS toxicity is measured by the loss of neurons (i.e., loss of cell volume or presence of markers of cell death) in both histopathology of human patients and studies in disease models. However, some therapeutic interventions that rescue the loss of neurons in a disease model cannot prevent the onset of disease symptoms and eventual death. For example, when an SOD1-G93A mouse model of ALS was crossed with a BAX knock-out mouse to protect the motor neurons from apoptotic-cell death, SOD1-mediated death of motor neurons was completely rescued, but time to symptom onset and lifespan were only modestly improved (Gould et al., 2006). In a similar study, SOD1-G93A mice were unilaterally injected with neural stem cells expressing glial-derived neurotrophic factor (GDNF), which improved motor-neuron survival through all stages of disease progression, but not the onset of disease symptoms (Suzuki et al., 2007).

By focusing exclusively on cell survival, we may create therapies that successfully rescue the cell body but still fail to prevent disease onset and progression. Fortunately, researchers found a measure that better correlated with the severity of disease symptoms: loss of axonal processes of the motor neuron from its connection on the motor endplate. By expressing GDNF in the muscle instead of the spinal cord, using transgenic muscle–specific expression of GDNF or transplantation of mesenchymal stem cells that express GDNF, several researchers showed a significant improvement in survival of the ALS mouse model (Li et al., 2007; Suzuki et al., 2008). Strikingly, GDNF treatment restricted to the muscle rescued *both* cell bodies and the innervation of the neuromuscular junction. A similar effect was seen in earlier studies. For example, a modest improvement or lack of improvement in functional deficits and survival was correlated with a modest or lack of improvement in axon counts, respectively (Gould, 2006) (Suzuki, 2007).

In addition to better correlating with functional deficits, denervation of the neuromuscular junction may also be an early stage of ALS progression. A study in the same mouse model more closely tracked the progression of disease symptoms related to both muscle-endplate innervation. The onset of measurable motor deficits occurred at 78-85 days while frank cell loss

was not visible until 100 days (Fischer et al., 2004). However, loss of neuromuscular innervation was an even earlier measure of disease than the motor deficits with visible divergence between healthy and diseased animals measurable at 47 days. Thus, denervation can define an initial period of disease progression that would have been missed by solely studying cell survival. Including denervation as a measure of disease will enable us to develop therapies that target an early stage of disease. Targeting these early stages of the disease offer the most hope for a successful intervention that can delay the onset and/or progression of disease symptoms.

Neurite Segmentation

Measuring neurites may enrich our insights into the mechanisms of neurological diseases and improve our search for efficacious therapies. For example, the morphology of neurite branching is more complex and intricate than that of the simple cell body, and thus, neurite responses, rather than cell survival, may offer a much greater dynamic range and sensitivity for assaying health. In addition, neurites are clearly correlated with neuron function, and they can indicate disease progression *before* the onset of neurological symptoms. With our longitudinal platform, which exposed the true mechanism underlying inclusion bodies in HD by evaluating cells in real-time (Arrasate et al., 2004), we extended neurite measures to include biologically important dynamics, such as how neurites extend or retract through the stages of disease.

When introducing a new measure, especially a longitudinal one, we must confront two practical obstacles: additional time for experiments and for analysis. Imaging throughput is very sensitive to increases in experimental time. Longitudinal experiments already take longer than single-timepoint experiments to collect valuable data on how the measurements change over time, so we must minimize any further increase. With low costs for computational time, our concern shifts from additional computational analysis time to how much

a human expert must be involved. Scientists can easily create a bottleneck during analysis because of the many demands on their time and the sequential nature in which they process data.

Ideally, a new measure would add no additional time to an experiment and would not need to be performed more often than the daily imaging done to analyze cell survival. With our platform, we can achieve this ideal for measures of neurites, because all experiments are performed with a freely diffusible fluorescent protein that outlines the morphology and marks both the cell body and neurites. Although small-scale neurite changes can occur on the order of minutes to hours, daily imaging is sufficient for capturing the large-scale changes in neurites. Thus, we can successfully and longitudinally measure neurites under these conditions without costing additional experimental time.

Due to the growing popularity of automation in biological microscopy experiments, and the abundance, heterogeneity, dimensionality, and complexity of the images we can now collect, the time needed to analyze images is creating a bottleneck (Meijering, 2012). Even for experiments evaluating single timepoints, the number of images that can be generated in a screen preclude manual analysis and even restrict the percentage of data that a human can practically review for quality control. For example, consider a longitudinal experiment of a 96-well plate imaged daily over a week. Even if a scientist could trace all of the neurites in just five minutes per image, it would still take 504 hours or roughly thirteen weeks of full-time work to analyze a single experiment. This imbalance between how much data we can collect and how much we can manually analyze compels us to create automated solutions analyze data in a more objective, explicit, and unbiased manner.

Humans can easily analyze an image, which starkly contrasts with the difficulty in finding computational algorithms that can perform the same task with the same accuracy (Meijering, 2012). This contrast has been recognized in many contexts with images that contain biologically relevant features; however, identifying neurites in wide-field fluorescence images accurately

stands out as one of the most difficult problems. The root of this problem is quite simple: neurites are long, thin processes, which means that they have a lower volume for a given area in a wide-field image compared to the cell body of the neuron. The difference in volume can be two or more orders of magnitude, causing the signal-to-noise ratio to differ by an equal range. The DIADEM (Digital Reconstruction of Axonal and Dendritic Morphology) competition, established by the Allen Institute for Brain Science and Janelia Farm Research Campus of the Howard Hughes Medical Institute, exemplifies the importance and effort put into extracting accurate neuron tracks (Brown et al., 2011). While this example focused on reconstructing neurite arbors of individual neurons from sparse, single timepoint images, it also highlighted the difficulty in designing an algorithm to choose the dim neurites in an image complicated by other objects. Also, despite some automated solutions, the motivation for the DIADEM competition was that most measurements on neurites are still performed by hand.

Almost all large-scale screens for therapeutic interventions are experiments on fixed cells. As a result, the existing automated algorithms for extracting neurites are optimized for conditions substantially different from those of longitudinal imaging of live cells. In some cases, the neuron-plating densities are purposefully kept sparse so that the experimenter can easily visualize the arbor or individual process of a single neuron (Misiak et al., 2014). The same approach is used for longitudinal experiments that use a low-throughput sit-and-stare approach to tracking the same location over time (Fanti et al., 2011; Kaech et al., 2012). The experimenter may use staining that is incompatible with live cell imaging to highlight the cell body or neurites (Huang et al., 2010; Jain et al., 2012; Pani et al., 2014). The algorithms developed for these experimental set-ups, even those that rely on sophisticated divisions such as graph-based segmentation, assume the experimenter can determine a single threshold value to distinguish pixels of biological objects we care about from background (Charoenkwan et al., 2013; Ho et al., 2011; Kandaswamy et al., 2013; de Santos-Sierra et al., 2015; Wu et al., 2010).

We initially pursued a similar threshold-based method to identify neurites in our longitudinal images; however, we found inaccuracies and inconsistencies over time and across experiments. In retrospect, we realized that there are several reasons why a threshold technique would not work well for live-cell imaging. First, live-cell fluorescence imaging often has a lower signal-to-noise ratio than immunofluorescently labeled fixed cells. The underlying biology controlling the expression and degradation of the biosensor determines the intensity of the fluorescence signal, and a dim signal cannot be corrected solely by increasing exposure time because that can cause phototoxicity. Second, unlike fixed cells that can use secondary antibodies to multiply the number of fluorophores that tag a protein, long-term imaging of live cells require expression of fluorescent proteins with a single fluorophore attached to it. (Shorter-term imaging can use small molecule dyes, but they are usually only useful for a few hours after loading and are often toxic over longer timescales.) This approach restricts expression to lower levels, because at very high levels, overexpression of some freely diffusible fluorescent proteins can induce toxicity. Third, while fixed cells only need simple media formulations, such as phosphate buffered saline, living cells also require molecules such as amino acids, flavins, and other metabolites, which cause background fluorescence (Billinton and Knight, 2001). Some special treatments, such as sodium borohydride or proprietary cocktails like MaxBlock, can reduce autofluorescence in fixed cells, but they do not support long-term culturing of live cells.

The most important reason that thresholding is inappropriate for longitudinal neurite segmentation is that biological features change the fluorescence intensity over time. For example, if production is greater than degradation of a constitutively expressed fluorescent biosensor, then the fluorescence intensity of the biosensor will increase over time. The amount of biosensor production and degradation may also change over time, potentially leading to a steady-state level if production and degradation reach a balance. Freely diffusible fluorescent proteins that highlight neuron morphology exhibit this monotonic increase over time. This increase means that bright objects will become more prevalent over time, and when an object is

sufficiently bright, it will brighten, or "bloom," adjacent pixels due to its diffraction pattern from light passing through the objective and its light scattering off adjacent objects. The composition of the image can also undergo a dramatic shift. In healthier wells, the criss-crossing web of axons and dendrites grow more close-knit, while in less healthy wells, neurites grow sparser and neurons are replaced by debris from cell death. These unique challenges must be overcome to unlock valuable insights into how neurites extend and retract in health and disease.

To solve these challenges, we developed an analysis system built on supervised machine learning. This system uses a simple workflow that replaces analysis by hand-crafted algorithms with human annotation. This workflow is built on images that represent the probability that each pixel could belong to a certain category, such as background, neurites, neuron-cell bodies, or debris. By transforming the image from arbitrary units of fluorescence to a confidence probability for each pixel, we can evaluate how much our conclusions differ with more or less stringent interpretations of the data—a distinction that separates image processing from image analysis (Danuser, 2011). This probabilistic-image approach also refines the category using a hierarchical approach that proceeds from the distinction between foreground and background towards more complex features. Using images from our longitudinal microscopy platform, which positionally aligns images through time, we can investigate how neurites change within and between models of neurodegenerative disease, with a focus on resolving early indicators of disease. We believe that robustly and longitudinally detecting neurites will give us valuable insights into mechanistic and therapeutic studies of neurodegeneration.

Probability images generated by supervised machine learning can capture dim neurites while still excluding background

Faithfully identifying neurites in living cells over time with conventional approaches to segmentation is challenging, because there is no single imaging condition we can optimize. For

example, to accurately and consistently distinguish between a background and foreground pixel across the entire image at all imaging timepoints, we must consider the low signal-to-noise ratio, the increasing fluorescence intensity over time, and the bloom from bright objects. Indeed, the background pixels around bright objects will be significantly brighter than those around dim objects like neurites. Thus, by picking a single number to threshold an image, the size of bright objects will be overestimated or dim objects will be falsely labeled as background and missed. In addition, we need to determine whether a foreground pixel belongs to a neurite, a neuron, or cellular debris. Ideally, to remap the intensity to the most plausible identity of that pixel, we need a pixel-by-pixel method that is influenced by the fluorescence intensities of a pixel and its neighborhood.

We describe an image where each pixel value represents the probability that the pixel belongs to a certain category, such as background or foreground, as a probability image. This image is similar to (Meijering et al., 2004) that used for a five-term hand-crafted algorithm that predicts "neuriteness" to aide human tracing of neurites. In contrast, we use a true probability image calculated from a learned model of neurites involving measurements that include local spatial information. These images are not transitory images computed before classification by a threshold. Rather, we use them extensively in our analysis pipeline to help refine pixel categorization, unify our interpretation of images over time, and provide insight into the robustness of a measurement.

While our approach can classify many biological features, we will focus on the first stage of separating background from foreground. For the first step, scientists are shown random image patches, in which they label every pixel as either background or foreground. Then, for each image, several hundred filtered images are computed to highlight features, such as intensity, size, and texture. Each pixel from an image patch is then assigned a label and a vector of values from the filtered image, which are fed into an algorithm that finds the smallest set of filters and weighs each filter to correctly separate foreground and background pixels. The

output of this algorithm is the probability that a pixel is foreground, such that 100% definitely represents foreground and 0% definitely represents background. This probability can be interpreted as the range of expert beliefs about a pixel (i.e., the percent of time a scientist would label a pixel like this as foreground). Once each pixel in an image has a foreground probability associated with it, we can create an entire image where the original intensity has been replaced with the foreground probability. As shown in Figure 8, this foreground probability image can be both sensitive and selective, capturing the dim neurites without also mislabeling background pixels as foreground.

Biological features can be identified using a decision tree of probability images

While any machine-learning model can output the probability of classification to generate a probability image, we use a hierarchical approach to find nested biological features in an image (Figure 3). Each decision made by a model is binary, so we do not need to consider all potential multi-class distinctions initially, and the probability images are both the output and a potential input for the subsequent models. An example of a nested feature is a pixel belonging to a neuron body. Here, we can trace the decision tree that will specify the probability that the pixel belongs to the neuron class. First, we assess the probability that the pixel is foreground rather than background. Next, we assess the probability that this pixel, if we assume it is a foreground pixel, belongs to the non-neurite class. This class has several potential refinements: it could be a neuron body, a glial cell, or debris. The final probability of the pixel being a neuron body is represented by the equation below.

Equation 8:

$$P(\text{Neuron Body}) = P(\text{Neuron Body} \mid \text{Non-neurite}) * P(\text{Non-neurite} \mid \text{Foreground}) * P(\text{Foreground})$$

Each point of the decision tree is a model created with labeled training data. For dim objects, such as neurites, the first prediction of foreground is an important distinction that helps us use local spatial information more effectively because a dim neurite pixel may become a bright foreground pixel. We can predict the neurite probability of that pixel by combining spatial information from both the foreground-probability image and the fluorescence-intensity image. In our hands, the most important parameters in models trained using both a probability image and the fluorescence intensity image will include parameters from both images. The superiority of this approach is also evident from the increase in accuracy compared to using only the original fluorescence-intensity or foreground-probability images (Figure 9).

Another advantage of this hierarchical approach is that we can re-use the same workflow to generate each model in the decision tree. This advantage means that we can use the same software to classify and label pixels and train machine-learning models to distinguish between simple and complex data, such as foreground versus background and neurons versus glia, respectively. With our method, we can also easily extend the pixel classification to make more sophisticated ones such as classification of mitochondrial fragmentation. The only change is the labeled data that gets fed for training purposes, and this saves considerable time that would otherwise be devoted to crafting new image analysis algorithms.

Probability images can selectively and dramatically increase the signal-to-background ratio of neurite pixels

Despite including local spatial information in predictions of neurite pixels, the predicted probabilities will be lower for dim neurites than those of brighter neurites. Dim neurites will have lower pixel values in the filtered images, and this difference means that the machine-learning model will have difficulty assigning dim neurite pixels a high-confidence prediction close to 100%. While we can use the hierarchical approach (i.e., using both the original fluorescence-intensity image and the foreground-probability image) to predict the neurite-probability image

and increase the prediction probability for dim neurites, in general, dim neurites still do not have a high-confidence prediction. However, neurite pixels in the probability images do show improved separation from the background pixels that contain no signal from any fluorescent object, referred to as the signal-to-background ratio (S/B). We can assess this improved separation by comparing the neurite S/B between the probability images and the original fluorescence-intensity images. This approach will help us quantify the increased separation enabled by the probabilistic approach.

For a fluorescence-intensity image and a neurite-probability image of the same well and timepoint, we identified regions without any objects and calculated the mean from those regions to define the background value of the image. We then divided each image by its appropriate background value to create an S/B image for the fluorescence-intensity and the neurite-probability images. To create an image that represented the fold change of the S/B from the probability image to the fluorescence image, we divided the pixels of the neurite-S/B image by the fluorescence-S/B image. To isolate neurite pixels, we created a mask from the 50%-confidence threshold of the neurite-probability mask and then graphed the cumulative histogram for the neurite pixel

The fold-change in neurite-probability image S/B compared to fluorescence-intensity image S/B was above one for greater than 99% of neurite pixels from both early and late timepoints, which indicates that neurite pixels are selectively enhanced (Figure 10). The median improvement in S/B is greater than 10-fold, and approximately 23% of neurite pixels had greater than a 20-fold improvement for the early imaging timepoint with more dim neurites. This finding demonstrates that the neurite-probability images can generate the higher S/B images normally associated with single snap-shot techniques, without the drawback of killing cells before imaging and using a sister plate for each experimental timepoint.

The improved separation between neurite and background pixels in the neurite probability images is also evident in comparisons of thresholding the original fluorescence-

intensity images and the neurite-probability image, as shown in Figure 8. For the dim neurites in fluorescence images, the lower threshold values can include the dim-neurite pixels, but they will also permit more background pixels. This effect makes the lower threshold on the fluorescence image more sensitive, but less selective. By raising the threshold value, we can exclude almost all background pixels, but at the expense of also excluding dim neurites. This approach reverses the effect, making the higher threshold on the fluorescence image less sensitive, but more selective. In contrast, a single threshold on the probability image can capture dim-neurite pixels while still excluding background pixels, demonstrating that the enhanced separation of the neurite-probability image can make our analysis both sensitive and selective.

Per-pixel probability analysis generates more robust segmentation than conventional thresholding techniques over multiple timepoints

The intensity units of the probability image have a defined meaning, the confidence in the predicted category, which is an important advantage over the original fluorescence images, especially for longitudinal studies. Fluorescence-intensity values can be meaningful over time, for instance, when estimating changes in protein expression. However, the relative definitions for foreground- and background-pixel intensities are defined by the local pixel neighborhood, which contributes to the significant trade-off between sensitivity and selectivity when thresholding fluorescence images. This trade-off is exacerbated if one wants to define an equal threshold across all imaging timepoints of a longitudinal experiment, when the number of dim neurites and amount of bloom from bright objects is changing over time. The probability images are created by machine-learning models that are trained through examples of a wide range of timepoints; thus, the intensity of the probability image can be more consistently compared over time.

The intensity of neurites changes over time as the expression of freely diffusible protein marking the morphology increases (Figure 11A). If we choose the image mean of the first timepoint as the threshold value for all timepoints, we can see that the mean is too low to accurately classify background near the bright cell body at the last timepoint and too high to capture all of the neurites at the first timepoint (Figure 11C). Conversely, if we choose the image mean at the last timepoint as the threshold value for all timepoints, we would fail to capture dimmer neurites at the first time point (Figure 11D). However, a confidence threshold for the foreground-probability image set at the first timepoint can capture the dim neurites at the first timepoint while still selecting against the background pixels at the last timepoint (Figure 11E). This approach shows that a single probability value, instead of a single fluorescence intensity, can consistently, sensitively, and selectively discriminate foreground from background over time.

If the fluorescence-intensity threshold determined at a single timepoint is not consistently accurate across all timepoints, we can also evaluate the performance when calculating the threshold value per timepoint. To determine the accuracy of the per-timepoint mean and the median of fluorescence-intensity thresholds compared to a 50% threshold on the foreground-probability image, researchers can draw pixel masks of foreground and background images over 7 days. We assessed the sensitivity and selectivity of each method at each timepoint. Table 4 shows the mean sensitivity and selectivity values and the standard deviation for each measure (in parenthesis) over time.

Table 4. Comparison of neurite sensitivity and selectivity of thresholding between fluorescence and neurite-probability images.

	Sensitivity Foreground Accuracy (Std. Dev. over time)	Selectivity Background Accuracy (Std. Dev. over time)
Intensity Image:	59.0%	95.8%

Mean	(17.2%)	(1.3%)
Intensity Image: **Median**	88.2% (7.4%)	**74.3%** (7.4%)
Foreground Probability Image: **50%**	**90.9%** (2.1%)	91.3% (2.7%)

Similar to the results shown in Figure 8D-G for a single timepoint, the per-timepoint mean threshold is very selective, but not very sensitive, while the per-timepoint median threshold is more sensitive, but less selective. Both of these methods are out-performed by a single confidence threshold on the foreground-probability image that demonstrates both high sensitivity and selectivity. The small variation in the sensitivity and selectivity for the foreground-probability image indicates that the probability images can provide a consistent basis for analysis over time. Thus, the trade-off between sensitivity and selectivity for thresholding fluorescence intensity is present for comparing both single and multiple timepoints, and probability images are superior when performing sensitive and selective longitudinal analysis. *Next, we investigated if our longitudinal imaging platform, which returns to precisely the same location at each timepoint, could create novel measures of neurite remodeling.*

Imaging with precise-positional correspondence over time coupled with per-pixel predictions of neurite classification enables dynamic measures of neurites

Neurite changes can be an early and significant indicator of disease progression and severity. Thus, by studying neurite extension and retraction, we can better quantify, understand, and therapeutically target these important changes early in disease. Fortunately, we do not need to adapt our longitudinal microscopy platform to collect this data. By returning to the same location over time and highlighting the neuronal morphology with a freely diffusible fluorescent

protein, we can create an imaging dataset that contains information on both neuron-cell survival and neurite dynamics. We can also use our novel pipeline for image analysis on these datasets to selectively and robustly measure neurites of living neurons over time, without additional costs to experimental time. This provides a high-throughput means of moving beyond solely focusing on the cell body to evaluating potential therapies and improving our understanding of how disease conditions evolve.

Previous studies used several techniques to quantify neurite dynamics. For example, manual measurements or kymographs may be used for sparser images and generally shorter and higher frequency imaging. Neither of these methods can scale up well for high-throughput automated microscopy, because the combination of large image datasets from automated microscopy, especially when imaging large montages and multiple timepoints, and the time-consuming nature of manual measurement makes these methods unworkable for analyzing entire experiments. Kymographs are also poorly matched for high-throughput imaging, because they are used to zoom into a particular neurite to study its dynamics on much shorter timescales than the several days used in our neurodegenerative model. Imaging a plate often enough to capture these shorter timescales comes at the cost of overall imaging throughput, and focusing on a single neurite is poorly matched to the high density of neurites in our images.

The most common measure for looking at neurites over time is the total neurite length. This measure involves segmenting neurites at each timepoint into a binary mask, creating a single pixel–wide "skeleton" from that mask, and then adding up the number of non-zero pixels. Because these neurite images are collected on either sister plates or automated microscopes that do not return to the same location, the neurites that are observed at each timepoint are not the same. For this measurement, extension and retraction are just net changes in the total area within a field rather than a direct measurement of specific neurites within the field.

While our image-analysis pipeline can measure the neurite area and length at each timepoint, we can use the positional alignment in our images to measure neurites at per-pixel

levels. We can detect when an empty pixel becomes occupied by a neurite and when an occupied pixel becomes empty (Figure 12A). However, we cannot presently relate neurite pixels to their neurite arbor or originating neuron. The neurites from all neurons have the same fluorescence channel, so we cannot assign ownership in the dense web of neurites. Instead, we can only measure the per-pixel changes in neurites, such as the appearance or disappearance of a neurite pixel. However, the appearance or disappearance of neurite pixels do not occur randomly, but, rather, are tied to a biological process. Thus, I will use the terms neurite extension and retraction as interchangeable with neurite-pixel appearance and disappearance.

With longitudinal neurite measures, we can gain insight into the amount of neurite remodeling that occurs within a well. A significant advantage of being able to directly evaluate the per-pixel dynamics of neurites is that we can assess the respective contributions of neurite extension and retraction to the net change in neurite area. The previous methods that only measured the net change in neurite area could only detect an imbalance between extension and retraction, but that imbalance could result from very different magnitudes for extension and retraction. One extreme possibility would be if the amounts of extension and retraction were perfectly balanced. While the reality could be that every existing neurite retracted and was replaced (in area) somewhere else on the image by a newly extended neurite, there would be no net change in neurite area, which would be naively interpreted as stasis. Below, we discuss two methods for determining the contribution of extension and retraction to neurite remodeling with our neurite-probability images.

Per-pixel Arithmetic

A straightforward method for determining the amount of change in neurites would be to subtract the probability image of timepoint *T+1* from the that of timepoint *T*. Each probability image normally contains only positive-valued intensities, so the difference image would need to be capable of storing negative values. In the difference image, the values greater than zero

would indicate that the probability value of timeoint *T+1* at that location was greater than that of timepoint *T*, and the values less than zero would indicate that the probability value of timepoint *T* at that location was greater than that of timepoint *T+1*. If the amount of neurite remodeling needed to be summarized by a single number, one could take the mean of the difference image.

However, directly computing the difference between the probability images of the two timepoints can lead to misleading measurements. We do not want to count a neurite pixel changing from 80% to 90% confidence as extension or a neurite pixel changing from 20% to 10% confidence as retraction. If the goal is to measure retraction or extension, only large changes in probability should be considered, because they most likely indicate a new neurite forming or an existing neurite disappearing. Thus, we need a different method that does not conflate many small changes in neurite probability with a few large shifts in neurite probability.

To reduce the impact of small changes in neurite probability, we can use a confidence threshold on the probability image. When we set a confidence threshold, only neurite pixels that have a probability above the confidence threshold are counted as neurites. For comparing across the two timepoints, we use the same confidence threshold, so a pixel must have a probability above the confidence threshold in both timepoints to count as a persistent neurite. To ensure that our conclusions do not rest on a particular confidence threshold, we calculate neurite measurements with a range of confidence thresholds from 20% to 95% in 5% increments, which captures the "trajectory" of neurite dynamics based on the confidence threshold. This method produces two measurements: the amount of area occupied by new neurites and the amount where a neurite retracted.

When summarizing neurite remodeling at each timepoint with confidence thresholds, we must consider that the magnitude of neurite extension and retraction is related to the total amount of neurite area. The more neurites in an image, the more opportunities there are for a neurite to extend or retract. To control for this effect, we divide the measurements for extension and retraction to create a neurite-dynamics ratio that signifies whether we are more likely to get

an extended neurite when the ratio is greater than 1 or a retracted neurite when the ratio is

below 1 (Figure 12B). We use this method for all of our current analyses.

This method does have some limitations. For example, the neurite-dynamics ratio is not

defined when there is no retraction. In practice, this result only affects the measurements of

images that are almost empty. The total area of neurites represents the upper bound on the

magnitude of neurite retraction, so the amount of neurite retraction goes towards zero as the

total area goes towards zero. Fortunately, our image-analysis pipeline can automatically identify

anomalous images, such as those that are empty, so we do not get undefined ratios from zero

retraction. However, for conditions with substantial neurite retraction over the timecourse of the

experiment, the neurite-dynamics ratio can paradoxically increase at later timepoints, when the

images are close to empty.

2D Histogram

An advantage of the neurite-dynamics ratio is that it summarizes neurite remodeling into

a single number. This result helps us assess whether the differences in neurite dynamics

between healthy and diseased conditions are statistically significant. However, condensing

differences in the millions of pixels in two neurite-probability images into a single number will

inevitably oversimplify the full range of neurite changes that occur. For more open-ended tasks,

such as clustering drugs or genes in a screen of disease modifiers, we want a method that can

more fully capture the scope of neurite remodeling.

By creating a 2D histogram of the probability images of adjacent timepoints, we can

visually represent all of the changes in neurite probabilities between the two timepoints (Figure

12E). The x-and y-axis of the 2D histogram represent the probability values of the earlier and

later timepoints, respectively, and the value of the 2D histogram at a particular x,y location is

incremented by 1. The final pixel value in the 2D histogram indicates how many locations in the

two images had that starting and ending neurite probability. To complete the 2D histogram, we

iterate through each position of the probability image (the two probability images have been aligned and cropped so that each pixel position corresponds to the same physical location at each timepoint). In the resulting 2D histogram, pixels near the diagonal from the bottom-left corner to the top-right corner represent neurite probabilities that stay constant between the two timepoints. Pixels along the left edge represent neurite probabilities that have moved from low to high, consistent with neurite extension. Pixels along the bottom edge represent neurite probabilities that have moved from high to low between the two timepoints, consistent with neurite retraction. A lookup-table is used to apply a color scheme to the range of pixel intensities in the 2D histogram.

The 2D histogram embeds neurite changes into its pixels' spatial location and intensity, and we can use the resulting pattern to interpret and compare neurite phenotypes. For instance, we can use an image-distance algorithm that computes the absolute difference in pixel intensities between two images to assess the similarity in the 2D histograms between disease conditions under different drug treatments. We can also use it to compute the average 2D histogram from multiple wells of the same condition to identify the most consistent changes in neurite dynamics. If desired, we can control for differences in the overall neurite area in two probability images by normalizing the 2D histogram so that the sum of the pixel values is the same in the two 2D histograms.

The 2D histogram can also be used in the future as an alternative way to quantify neurite dynamics. For instance, rather than single thresholds, one could use bounding boxes over regions. These bounding boxes can capture neurite changes that may be spread out across a range of neurite-probability values, while still being more selective. One could create a diagonal bounding box to capture the tendency for stasis in neurites between two timepoints and a bounding box on the x- and y-axis to capture the retraction and extension. We will use this method of quantification in future analyses.

The disadvantage of any per-pixel method of measuring neurite dynamics is that it is subject to error if a neurite moves: the movement will appear as an addition to both neurite extension and retraction. Neurite-probability images help us overcome this error by considering that the intensity units represent our confidence in a pixel's neurite classification. By applying an image filter to a probability image, we can blur the precise location of the probable neurite pixels, which we can use to reduce the potential impact that neurite movement can contribute to the results.

To correct neurite extension for movement, we can perform a max filter on the probability image from the earlier timepoint. This approach will enlarge the neurites at the earlier timepoint. Because neurite extension is defined as a pixel with a low-neurite probability at the earlier timepoint and a high-neurite probability at the later timepoint, neurite movements that are smaller than the radius of the max filter will no longer contribute to measures of neurite extension. With a similar procedure, we can control for the effect of neurite movement on the neurite-retraction measure by using a max filter at the later timepoint. Similar to using an array of confidence thresholds to ensure that a result is not due to a particular threshold, we can use the max filters applied to the earlier or later timepoint to reduce the possibility that the result is due to neurite movement. We can use this same method for both confidence thresholds and 2D histograms.

Armed with these novel measures of neurite dynamics, we investigated the differences in neurite dynamics between healthy and diseased neurons.

Measures of neurite dynamics reveal that TDP43-M337V increases neurite retraction

Previous researchers have shown that overexpressing both wild-type TDP43 and mutant TDP43-M337V can decrease neurite area *in vivo* and *in vitro* (Tripathi et al., 2014). Interestingly,

knocking down TDP43 with siRNA also impairs neurite area (Fiesel et al., 2011). This result suggests that TDP43 may have a multi-faceted effect in disease by regulating both the extension and retraction of neurites. By elucidating the role of TDP43-M337V in altering neurite dynamics, we can better understand the cellular processes affected by the underlying disease mechanism. We used the assay optimization plates and primary screen plates created by Dr. Eva LaDow to screen FDA-approved drugs funded by ALS-TDI.

Detrimental measures of neurite dynamics increase in severity with increasing doses of TDP43-M337V

The first step in assessing our new neurite measures of disease is to ensure that they change appropriately when we ameliorate the disease. However, there is no intervention that we can use as a positive control for ameliorating the disease as there are no truly effective treatments for ALS. However, previous work in single cells has shown that increasing the amount of the TDP43 increases the risk of death for the cell, and the amount of TDP43 expressed in neurons is correlated with the gene dosage of TDP43 (Barmada et al., 2010). Because our ALS disease model is based on overexpression from a transfected plasmid, we can modify the gene dose of TDP43-M337V by controlling the amount of transfected plasmid DNA. Thus, we used a dose-response of the mutant gene as a positive control for disease-specific survival and neurite phenotypes at different levels of disease severity.

Two 96-well plates of primary cortical neurons from rats were transfected with mApple (100 ng per well), as the morphology marker, and three doses (10 ng, 50 ng, and 100 ng of DNA per well) of either TDP43-M337V as the ALS disease model or EGFP as the innocuous protein as the healthy control. Some conditions were also treated with drugs, but those were excluded from this analysis. This approach provided a total of 8 wells per transfection condition for our analysis. Plates were transfected after 4 days *in vitro* (DIV) and then placed into the automated incubator of Roboticscope III, where they were imaged daily for seven days.

Our longitudinal measurements are a superset over traditional methods, so we can replicate previous results before investigating with our novel measures. With this experiment, we first verified that TDP43-M337V produced the known disease phenotypes of increased risk of death and decreased neurite area. We assess neuron survival using the automated survival algorithm in Enlighten with human curation. We found that increasing the dose of TDP43-M337V resulted in a parallel increases in risk of death, as shown by the higher slope of the cumulative-hazard curves with increasing dose of TDP43 (Figure 13A). We also determined the Cox proportional hazard ratios (HR) using the 10 ng TDP43-M337V as the baseline hazard (Figure 13B), which revealed a significant dose-dependent increase in the hazard ratio for 50 ng (HR: 1.85, $p < 1e-11$) and 100 ng of TDP43-M337V (HR: 2.34, $p < 1e-11$). We analyzed the risk from the EGFP dose to ensure that the increased risk of the higher TDP43-M337V doses is not an experimental artifact of increased amounts of transfected plasmid DNA. The EGFP dose did not increase the risk of death, as evident by the overlapping slopes of the cumulative hazard curves (Figure 13C). Using 10 ng EGFP as the baseline hazard (Figure 13D), we also showed no significant difference in hazard ratios for 50 ng (HR: 0.93, $p=0.32$) or 100 ng of EGFP (HR: 0.94, $p=0.37$). Thus, we reproduced the disease-specific and dose-dependent increase in risk of death with TDP43-M337V.

We assessed neurite area using the neurite-probability image created by Enlighten with a threshold of 50% confidence. In contrast to previous work, which evaluated neurite area via a single timepoint, we assessed neurite area over the 7-day timecourse of the experiment. In the first 2–3 days, neurite area remained the same between the three doses of TDP43-M337V; however, at later days, it decrease with increasing doses of TDP43-M337V (Figure 14A). In contrast, neurite area for all three doses of EGFP was similar at all timepoints (Figure 14B).

To quantify differences in our neurite measures over time, we used a longitudinal mixed-effect model adapted from (Crainiceanu et al., 2005), which we will refer to as LNPB (Longitudinal Non-Parametric Bayesian) analysis. This approach allows us to structure our data

in a hierarchy, such that the neurite measure for a well at a timepoint is a combination of the

effect of the condition on the neurite measure and the possible individual deviations for that well.

Thus, we can share information between wells with the same condition to more accurately infer

the condition's effect on the neurite measure without the influence of well variation. We perform

LNPB analysis using markov chain monte carlo (MCMC) sampling, which we run with JAGS

(Just Another Gibbs Sampler) (Plummer, 2003) using a burn-in of 70,000 iterations and 100,000

sampling iterations with a thinning rate of 100. Because we want to determine whether two

conditions are statistically significantly different, we use LNPB analysis to compute the

difference between group means. If the 95% credible interval of the difference between group

means does not overlap with zero, we define the result as significant in analogy with a p-value

of 0.05. However, the mean of the raw neurite measures is not ideal, because the neurite

measures cannot be negative, so the distributions show a positive skew. We can transform the

distribution to more closely resemble the normal distribution by using the natural log of the

neurite measures. This means that the difference between group means is now the fold-

difference between group means, because subtracting logarithms equals the logarithm of the

division of their arguments. Statistical significance for the fold-difference is determined by no

overlap of the 95% credible interval with one (rather than zero), because the natural log of zero

is one. We use LNPB analysis to assess the statistical significance of all of our longitudinal

neurite measures.

With LNPB analysis, we assessed the dose-dependent effect of TDP43-M337V on

neurite area over time, as shown in Figure 14C. The increase in gene dose of TDP43-M337V

from 10 ng to 50 ng and 50 ng to 100 ng both have fold-differences significantly under one for

the later timepoints, indicating that higher doses cause lower neurite areas. With our longitudinal

dataset, we also assessed the earliest timepoint that the fold-difference was significant. The 50

ng dose took four days to be significantly different from 10 ng, while the 100 ng took only three

days to be significantly different from 50ng. Thus, we detected increased disease severity with

higher doses of TDP43-M337V via two measures: lower fold-difference in neurite area and earlier divergence in the fold-difference. Although our analysis is capable of more sophisticated measurements of neurite dynamics, we reproduced the TDP43-M337V-dependent decrease in neurite area that previous studies found when assessing neurite length at single timepoints 3–4 days after expression of the mutant protein (Tripathi et al., 2014).

We have validated that our dataset containing a TDP43-M337V dose-response reproduces the previously known disease-specific phenotypes: increased risk of death and decreased neurite area. Next, we assessed whether our novel neurite-dynamics ratio could detect disease-specific phenotypes in a dose-dependent manner. We measured the neurite-dynamics ratio using the neurite-probability images generated by Enlighten with a confidence threshold of 50%. The curves of the neurite-dynamics ratio for each TDP43-M337V dose are shown in Figure 15A. For all but one timepoint, the neurite-dynamics ratio is significantly lower with higher dose of TDP43-M337V, as shown by LNPB analysis in Figure 15B. The exception is the last day for the fold-difference between 100 ng and 50 ng, which may be due to the low total neurite area for the 100 ng condition that could make the estimation of neurite retraction less reliable. The lower magnitude of the neurite-dynamics ratio indicates that the balance of neurite extension and retraction shifted towards retraction. We used the EGFP dose-response to ensure that the decreased neurite-dynamics ratio is disease-specific. The curves of the neurite-dynamics ratio overlap for each EGFP dose (Figure 15C). These curves are not significantly different at any timepoint, as shown by LNPB analysis (Figure 15D). Thus, our novel neurite-dynamics ratio demonstrates detrimental effects that are dose-dependent and disease-specific as seen with previously validated measures of disease.

We compared the LNPB analysis of the neurite area to the neurite-dynamics ratio and found a striking difference in how early we can detect a dose-dependent difference. While the fold-difference in neurite area requires 3–4 days to reach significance (Figure 14C), the fold-difference in the neurite-dynamics ratio is evident from the first timepoint (Figure 15B). This

result demonstrates that significant differences in neurite dynamics occur before differences in the total neurite area. Thus, our novel measure of neurite dynamics can detect dose-dependent and disease-specific effects not possible by previous neurite measures.

Measures of neurite dynamics reveal that the loss of neurite area in our ALS model is mainly due to increased neurite retraction

Our new and dynamic neurite measures help us explain the effect of TDP43-M337V on neurite area over time through four possible hypotheses. The first is that the disease protein does not affect neurite outgrowth, but rather solely increases the rate of retraction. The second is that it doesn't affect the rate of retraction, but rather eliminates the extension of new neurites. The third is that both extension and retraction are increased, but that retraction has a larger increase that decreases neurite area. The fourth is that both extension and retraction are decreased, but that retraction has a smaller decrease. While we have focused on the neurite-dynamics ratio as a summary statistic for determining neurite dynamics, the power of our per-pixel measures of neurites is that we can directly measure both neurite extension and retraction. These direct measures allow us to investigate the neurite processes that are responsible for the decrease in neurite area in the TDP43-M337V model.

For our individual measure of neurite extension or retraction, we correct for the dependence of each measure on the amount of neurite area in the image. For neurite retraction, the amount of neurite area that could retract is limited by the amount of neurite area available. For neurite extension, the number of neurites that can extend is proportional to the amount of neurite area. Thus, we can normalize our individual measure of neurite extension and retraction by the total neurite area of the earlier image. This approach converts each measure into a percentage where 1.0 is 100%. If neurite extension is one, then an amount of neurite area equal to the previous neurite area has been added. If neurite retraction is one, then all of the previous area of the image covered by neurites is now empty.

The LNPB analysis shows that both neurite extension (Figure 16A) and retraction (Figure 16B) change in a dose-dependent manner; but the duration and magnitude of neurite retraction is longer and larger. Neurite extension significantly decreases for 3–4 of the earlier timepoints. In contrast, neurite retraction significantly increases for all but one timepoint for one condition. The exception, as before with the neurite-dynamics ratio, is the 100 ng compared to 50 ng at the last timepoint that may be due to low neurite area that makes measuring neurite retraction unreliable. The fold-difference is also higher for neurite retraction. While the mean fold-difference for neurite extension does not drop below a 20% decrease, the mean fold-difference for neurite retraction for 50 ng compared to 10 ng is statistically greater than 20% for all but the first timepoint. The 100 ng compared to 50 ng also shows an additional increase in neurite retraction; although, it is not as dramatic, potentially because the change in gene dose is small (2x versus 5x) or the disease process that regulates neurite retraction is saturable. These results support the conclusion that increased neurite retraction is the main cause of lower neurite area due to TDP43-M337V, with decreased neurite extension as a contributing factor.

To improve the resolution with which we can resolve the differences between neurite extension and retraction in diseased and healthy wells, we shifted our analysis to a larger dataset of the TDP43-M337V and EGFP wells used as positive and negative controls, respectively, in our primary screen of ALS-TDI drugs. As shown in Figure 17A, this screen is composed of eight drug plates that represent a unique pattern of drugs that are put into four replicate plates for a total of 32 96-well plates. The plate layout (Figure 17B) has 8 wells of positive and negative controls for a total of 256 wells (32 plates * 8 wells) for each condition. Each 96-well plate contained primary cortical neurons from rats that were transfected at four DIV with mApple (100 ng per well) to highlight the morphology, and either TDP43-M337V or EGFP at 100 ng per well. After transfection, the plates were stored in the automated incubator of Roboticscope III and imaged daily for 7 days as before. We ignored all drug-treated wells for this analysis.

The results of this larger dataset also support that neurite retraction mainly affects the deficit in neurite area with TDP43-M337V expression. Neurite retraction was increased for all timepoints (Figure 18C,D), while neurite extension was only decreased for three timepoints (Figure 18A,B). The fold-difference in neurite retraction was above a 100% increase for all but the first timepoint; however, the fold-difference in neurite extension did not decrease more than 20%.This replicated the sustained and larger magnitude changes in neurite retraction seen in the dose-response experiment; although, all of the differences in neurite extension and retraction are significant by LNPB analysis due to the increased sample size and the comparison of a high concentration of TDP43-M337V with EGFP.

We have shown that our novel neurite-dynamics ratio is both disease-specific and dose-dependent, with a sensitivity that is not possible to achieve by evaluating total neurite area over time. In particular, our ability to measure both neurite extension and retraction enables a higher-resolution view of the changes in neurite dynamics in disease. We used this higher resolution to determine that neurite retraction mainly affects the decreased neurite area observed with TDP43-M337V expression.

We next investigated if the insights from neurite dynamics could more sensitively detect and accurately classify neurons expressing a disease-associated protein than our traditional approach of survival analysis.

Measures of neurite dynamics sensitively discriminate between healthy and diseased wells

Cell survival and neurite measures are dramatically different in their potential to detect early stages of disease in ALS-mouse models *in vivo*. Neurite alterations are visible earlier in disease progression than cell survival (Fischer et al., 2004), and importantly for therapeutic development, rescuing neurites is well correlated with improving disease outcome (Gould et al., 2006) (Li et al., 2007) (Suzuki et al., 2007) (Suzuki et al., 2008). In our *in vitro* experiments, we

61

found that the early visible differences in measures of neurite dynamics in the TD43-M337V

dose-response experiments, and the dramatic increase in neurite retraction at all timepoints for

TDP43-M337, both indicate that changes in neurites are sensitive measures of disease

progression. The larger area occupied by neurites compared to neuron-cell bodies and the

binary state of cell survival also give neurite measures a larger possible range and finer possible

gradations. We wanted to test for a difference in the discriminatory power between our

traditional longitudinal measure of cell survival and our new longitudinal measures of neurite

dynamics.

Our goal was not necessarily to identify the superior measure. Instead, we wanted to

explore the richer parameter space of the measures in combination and how they might provide

more nuance into how the cell state changes during disease, and, conversely, detect the rescue

of those parameters with therapeutic interventions. One caveat is that neurites are always tied

to survival to some degree. If a cell dies, then its neurites must also disappear, although the

timing may vary. This result means that the clearest sign of unique insights from our neurite

measures occurs when you lose neurites without losing the cell body.

In addition to insights into disease progression, the discriminatory power of a measure

can practically benefit cellular assays that are often plagued by higher variation than simpler *in*

vitro experiments. By increasing the separation between the measurements of healthy and

diseased conditions, we can reduce the chance of false positives in a large screen and improve

our ability to identify perturbations that only partially rescue the disease phenotype.

Measurements of neurite dynamics are more predictive than survival in "diagnosing" a well

We wanted to compare the discriminatory power of our longitudinal measures for

determining the disease status of the well, i.e., diagnosing a well. We approached this

comparison with two methods: visualization of the separation between conditions and

classification accuracy using a machine learning model. For our examples of health and disease, we used the TDP43-M337V and EGFP dataset of only the positive and negative controls from the ALS-TDI primary screen.

Before quantifying their relative discriminatory power, we visualized the ability of our per-well measures of neurites and survival to distinguish wells with healthy cells from wells with and diseased cells. Because we observed a high neurite retraction phenotype previously, we plotted the percent of surviving cells versus the normalized neurite retraction between two timepoints (Figure 19A). The red dots indicate TDP43-M337V wells, while the green dots indicate EGFP wells. Each panel represents the timepoint labeled by the gray header. In order to make the units on each axis comparable, we divided the measurements of survival and neurite retraction for each timepoint by their respective standard deviation. The resulting measurements are unitless because the division cancels the units.

We assessed the discriminatory power of each measure by the degree of overlap for that axis between the two clusters of conditions. We saw evidence that the neurite retraction has more discriminatory power than survival at the first and second timepoints. The healthy and disease clusters are well separated on the neurite retraction axis but show overlap on the survival probability axis. We also saw that combining information from both cell survival and neurite dynamics can increase discriminatory power. The cleanest separation between healthy and diseased clusters is not wholly along the axes of survival probability or neurite retraction; rather, it lies at an angle to these two axes. This separation indicates that both measures in combination more accurately classify conditions than either measure alone. We can calculate the power of these parameters to distinguish the two groups as the separation distance between the centers of the two clusters by the Euclidean distance of the hypotenuse:

$$\text{Equation 9: } \quad \text{Separation distance} = \sqrt{(\text{survival probability})^2 + (\text{neurite retraction})^2}$$

63

From basic geometry, we know that the distance of the hypotenuse will be larger than either of the single-axis lengths. However, we also need to quantify how well separated the healthy and disease conditions are on each axis to assess the relative discriminatory power.

We assessed the importance of each parameter to accurately classify conditions by training a machine-learning model to predict whether a well is the disease or healthy condition based on different subsets of longitudinal measures. We chose to use the random forest (RF) method, because it can easily capture non-linear relationships and calculate the contribution of each parameter to overall prediction accuracy (Statistics and Breiman, 2001). An RF model is composed of many decision trees: in our case, 500 trees. Each decision tree is trained on a unique dataset that is ~⅔ the size of the original dataset and built by randomly sampling, with replacement, from the original dataset. At each decision branch of a tree, a subset of parameters, totalling $\sqrt{\text{Number of Parameters}}$, are randomly chosen. The parameter in this subset that best splits the data is chosen for this decision branch. An RF model predicts classification by counting the number of votes for each class from all of the decision trees and going with the majority. RF provides an accuracy measure that is a normalized average of the difference in prediction accuracy when that parameter is left untouched and when it is randomly permuted across examples.

We performed five-fold cross validation, in which the data were split into five even groupings by an identification (ID) that included both the experimental plate and the well position, such that each well in the experiment had a unique ID. For each round of training, four of the groups were combined for training the RF model, and the last group was used to test the classification on data that the model had not yet seen. The test group is different for each round, which allows us to control for overfitting while still using every datapoint to test the model. Measurements from all timepoints were included in the dataset, but all timepoints from a well were kept within the same group. The parameters included were the timepoint, the total neurite

area, the normalized neurite extension, the normalized neurite retraction, the neurite-dynamics ratio, the survival probability, and a column filled with randomly generated numbers to show the baseline accuracy provided by a non-informative parameter.

The important measures for each parameter from each fold of the cross validation are shown in Figure 19B. The main role of neurite retraction in TDP43-M337V neurite dynamics that we described previously is reflected in the dramatic ~2x increase in classification accuracy compared to survival probability. In contrast, neurite extension contributed less to accuracy than survival probability. The importance of neurite retraction also likely explains why the neurite-dynamics ratio is also ranked higher than survival probability. All of the parameters were more informative than the random variable.

The RF model trained with the parameters listed above has an average misclassification error for either confusing healthy for disease or vice versa of ~1%. Another way to assess the accuracy of a machine-learning model is to plot the Receiver Operating Characteristic (ROC) curve for each single-parameter model. The ROC curve shows the trade-off between sensitivity and selectivity for a binary classification depending on the threshold. The diagonal represents the performance of random chance, and the ideal model curve is a flat line with sensitivity of 1.0 that represents no trade-off. The RF model trained above approximates the ideal model curve, as shown in Figure 19C.

To further discriminate the power between survival and measures of neurite dynamics, we trained an RF model using all of the wells and timepoints as examples, but with only the survival probability or normalized neurite retraction as a model parameter. The model is trained with an 80/20 split between training and test data, using an ID that combines both the experimental plate and the well position. Even though there is no timepoint parameter in either model, the ROC curve of the neurite-retraction-only model still approximates the curve of the ideal model (Figure 19D). In contrast, the performance of the survival probability model performance was obviously better than chance, but there is a clear trade-off between keeping

sensitivity high and increasing the selectivity. This idea demonstrates that our measures of neurite dynamics can be more informative than cell survival in diagnosing a TDP43-M337V model of disease.

Measures of neurite dynamics can separate healthy and diseased wells under synthetic conditions with no differences in survival

One implication of the difference in the ROC curves of the single-parameter models is that there are examples that are indistinguishable by analyzing survival but distinguishable using measures of neurite dynamics. Using the same dataset from our ALS-TDI primary screen, we can isolate these examples by restricting our dataset: we use just the first two timepoints when there is a significant overlap in survival probability between the two conditions, and we restrict the survival probably between 0.74 ad 0.785 to collect an equal number of both conditions (Figure 20A). This approach captures 32 EGFP wells and 33 TDP43-M337V wells, the subset of EGFP wells that are more toxic than normal, and the TDP43-M337V wells that are less toxic than normal. Survival analysis of this subset of wells by Cox proportional hazards analysis shows a hazard ratio that is not significantly different from baseline (HR: 1.06, p=0.37) (Figure 20B).

We use a Bayesian equivalent to the T-test to assess the difference between group means for survival probability and neurite retraction. The BEST (Bayesian Estimation Supersedes the T-Test) method computes the difference between group means along with its credible interval (Kruschke, 2013). Similar to our LNPB analysis, we call the result statistically significant if the 95% credibility does not include zero.

BEST analysis of the survival probability confirms that the difference between the two groups is not statistically different (Figure 20C). In contrast, the BEST analysis for neurite retraction does show a statistically significant difference in the group mean (Figure 20D). Thus,

we can use our measures of neurite dynamics to infer the disease condition that would be unreachable by looking at survival divergence alone.

Measures of neurite dynamics can detect detrimental changes in the absence of cell loss

We chose the experimental conditions for the primary ALS-TDI screen based on the robust survival difference between EGFP and TDP43-M337V. We can return to the dose-response plates to investigate whether we can discern a difference between healthy and disease conditions when the disease model is not severe enough to cause a difference in cell survival.

The lowest dose of TDP43-M337V in the dose-response plate did not show a significant difference in survival (HR: 1.04, p=0.58), as shown in Figure 21A,B. However, the neurite-dynamics ratio showed a divergence in the mean neurite-dynamics ratio at later timepoints (Figure 21C). With LNPB analysis, we confirmed that the fold-difference in the neurite-dynamics ratio is statistically significant at days 4 and 5 (Figure 21D). The later divergence may be due to the mutant protein slowly accumulating in a low-dose overexpression model of disease. This result demonstrates that our measures of cell survival and neurite dynamics replicates the difference in sensitivity seen in ALS models *in vivo* and in ALS patients (Fischer et al., 2004). *We wondered if the neurite dynamics were consistent across disease models or if potentially unique fingerprints appeared.*

Signatures of neurite dynamics differentiate diagnoses between disease models

We have seen that we can use measures of neurite dynamics to accurately diagnose whether a well is the healthy condition or the TDP43-M337V model of ALS. In part, this accuracy is derived from the early detectable changes in neurites, even before cell death. Given

the importance of neurites in the progression of symptoms in neurodegenerative disease, we can further explore the predictive power of neurite dynamics to investigate if they also contain clues about which disease model is present in the well. This approach is more difficult, because there is more overlap in the survival variation between disease models than with healthy controls. However, if possible, it would indicate that neurite retraction is not simply a sensitive but non-specific indicator of reduced cell viability but rather a disease-specific biomarker, which would further support our search for therapies using *in vitro* models.

We used the same longitudinal fingerprint of neurites defined above to see if we could distinguish disease-causing proteins based on the multi-dimensional study of relevant biological phenotypes. This step is the first towards understanding the mechanisms by which these proteins cause neurodegeneration and move cells from the healthy state to the disease state (or vice versa) and demonstrating the value that these measures add as an analytical tool and endpoint assay beyond survival.

Separating models of different diseases

The first step in assessing our ability to distinguish between models of neurodegenerative disease is to visualize changes in the balance of neurite extension and retraction in each model. With our ALS model, a LRRK2-PD model, and an EGFP-healthy control, we noticed that disease conditions have a bias towards neurite retraction, especially at later timepoints (Figure 22A). However, the longitudinal pattern of neurite changes do not entirely overlap over the timecourse of the experiments, indicating that there is a chance of differentiating them.

The ALS dataset is from the ALS-TDI primary screen as described before. The PD dataset is from the positive and negative conditions of a screen that is in progress by Dr. Gaia Skibinski and Ms. Alica Lee in our laboratory. The PD screen has 16 positive and negative

controls per plate, and we have used the first 9 plates that have been run. Both screens used primary cortical neurons from rats and were stored and imaged daily on Roboticscope III.

To quantify our ability to distinguish the identity of the disease model in each well, we trained an RF model to classify all three conditions. We performed five-fold cross validation as described previously. The parameter importance in our RF model with two disease conditions was quite different from our model with just EGFP and TDP43-M337V (Figure 22B). Neurite extension and retraction have reversed places: neurite extension is now the most important parameter, while neurite retraction is second lowest. However, the top and bottom parameters do not have the large ~2x spread in accuracy we saw previously, so the difference between best and worst does not indicate the same range in utility. This result indicates that there isn't a single-most-useful measure of neurite dynamics. Rather, the disease context will determine which measure has the most information.

As we saw before, the EGFP-healthy condition and the ALS model are almost always correctly classified (Figure 22C). The PD model does not fare as well: ~54% of its wells were misclassified. Surprisingly, the most common misclassification for the PD model is the EGFP condition rather than ALS model. This finding indicates that the lower classification rate is more a function of being closer to health rather than a similarity between all types of neurodegenerative diseases.

Chapter 5: Conclusions and Future Directions

In the preceding work, we have developed a fully-automated system for high-throughput robotic microscopy and an automated image analysis pipeline for our longitudinal datasets that incorporates supervised machine learning. Our third generation robotic microscope enables a three-fold higher-throughput than the previous generation, which is critical for the increasing number of screening projects that require imaging many wells in the hope of finding an effective treatment and iPSC experiments that require imaging of a large area due to sparsity of the cell type of interest. In addition to faster imaging, it is practical to run more experiments because they can be scheduled to start at any time during the day due to the integration of a robotic arm and an automated incubator with the robotic microscope and the development of a Micro-Manager plugin to perform the tasks for robotic microscopy that normally required a human operator. The ability to run fully-automated around the clock has also made experiments feasible that were previously challenging such as studies of protein degradation that require hourly imaging.

The high number of images from our latest robotic microscope made an automated solution to image analysis a necessity. Our datasets from longitudinal imaging of neurons are especially challenging to analyze due to the high dynamic range of the bright cell bodies and dim neurites and the increase in fluorescence intensity as the fluorescent proteins accumulate over time. We developed our new pipeline Enlighten along with software to capture image annotation by scientists to use supervised machine learning to classify each pixel in our images. The output of our pipeline is a series of probability images where each pixel value represents our confidence in the pixel's classification. We demonstrate that these probability images can sensitively and selectively detect even dim regions such as neurites. These probability images can be combined in order to identify more complex biological features.

The primary goal for the new analysis pipeline was to extend our measurements to include the quantification of dynamics neurites. The positional accuracy of our robotic microscopy combined with probability images enables us to measure neurite extension and retraction over time. We have validated that these measures are disease-specific and dose-dependent in an ALS model. We have shown that the decrease in neurite area in this model is dominated by an increase in neurite retraction. We compared our dynamic neurite measures to cell survival and demonstrated that dynamic neurite measures are superior at separating the healthy and disease conditions, and they can quantify differences between the conditions that would not be detectable using survival analysis.

We investigated the ability of dynamic neurite measures to distinguish between an ALS and PD disease model. While accuracy of the PD model was lower than the ALS model, the parameters that contributed to the accuracy were very differently ranked, indicating there is not a single useful neurite measure for differentiating diseases.

In the future, we would like to continue to improve the reliability and sophistication of Enlighten by growing the number of training examples. In particular, we would like to shift to neurons derived from iPSCs that have more heterogeneity than our rodent neuron cultures, and we will need to improve our ability to handle the objects in those images such as clusters of cells that we don't want to segment. We'd like to extend our pipeline to include subcellular measures such as mitochondrial morphology for those experiments that include a mitochondrial biosensor. An ideal experimental addition would be to adopt a protocol that gives each neuron a different color to highlight the neuronal morphology. This would enable us to investigate the individual arbors of each neuron rather than just a well-level measure. To increase our understanding of how the patterns of neurite changes may reflect different disease mechanisms, we'd like to perform our neurite dynamic measures on more neurodegenerative disease models. This comparison would be further aided by incorporating the 2D histogram of dynamic neurite measures more fully into our image analysis. Finally, we would like to further

train networks to resolve neurites into axons and dendrites, since they are often differentially affected in disease.

Using the combination of robotic microscopy and supervised machine learning, we can delve more deeply into our existing experiments and enable investigations that would otherwise be impractical. The large-scale screens now possible with high-throughput imaging and the ability to measure several biologically relevant measures in each well enable a longer term investigation of the measures that best translate from the dish into humans. Thus moving this technology forward will help us in our search to fill the unmet need for therapies for neurodegenerative disease.
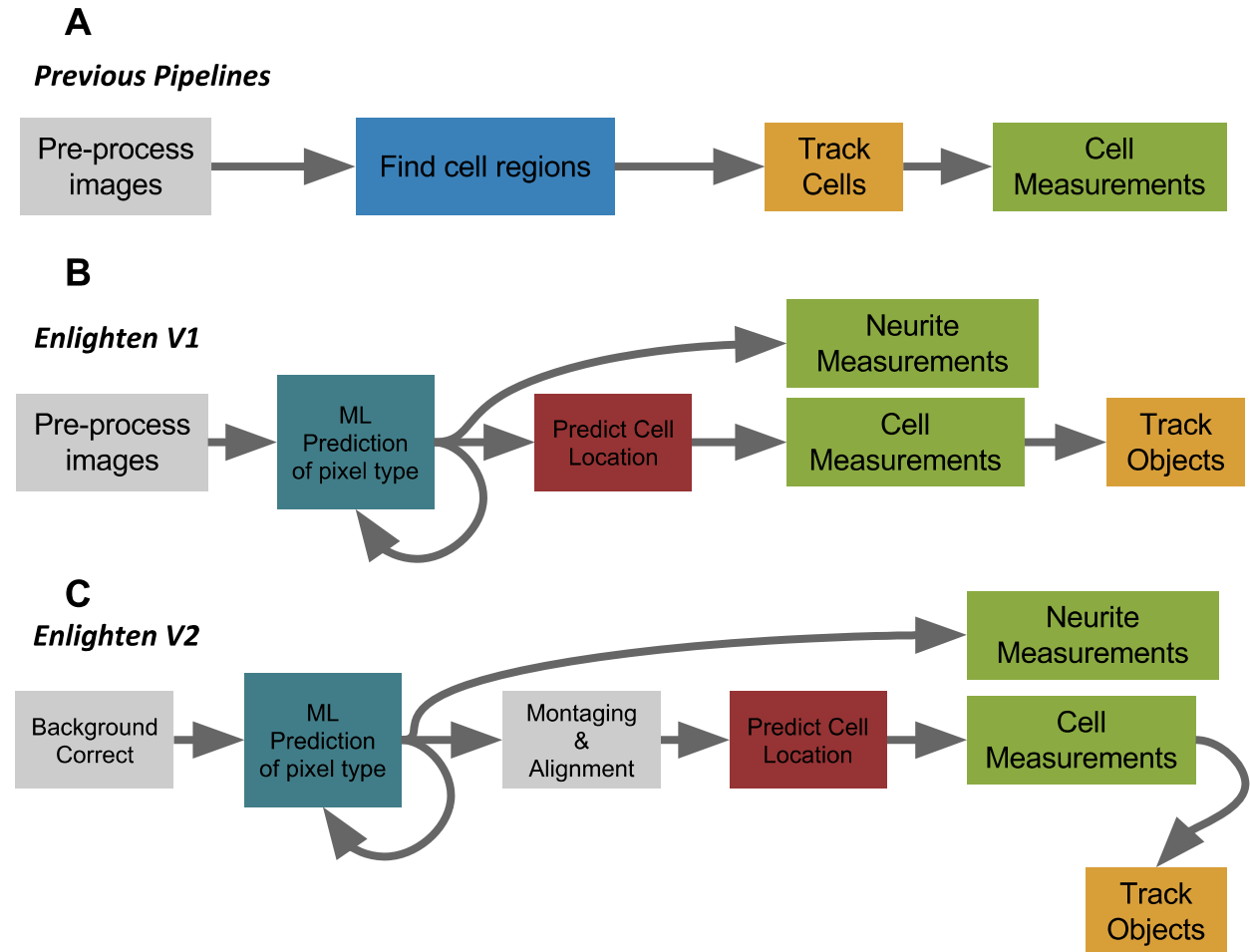
Figures

Figure 1.

Figure 1. Comparison of Enlighten and the previous image-analysis pipeline.
A. Schematic of the stages of the previous pipeline. First, images are pre-processed (grey), including background correction, montaging, and alignment. Second, neuron regions are identified (blue), using histogram adjustment, image filters, and thresholding. Third, neuron regions are tracked over time (orange), providing a unique ID for each neuron. Last, neuron regions are measured for shape and intensity (green). **B**. Schematic of the stages of the first version of Enlighten. After pre-processing, probability images are calculated using machine-learning models (dark blue). Probability images can predict the location of the neuron cell body (red). Measurements are expanded to cover both neuron cell bodies and neurites (green). The neurite measures include dynamic measures. The cell measurements tracked objects using similarity measures over time (orange). **C**. The second version of Enlighten split the pre-processing steps to avoid errors in measuring neurite dynamics from montaging errors.
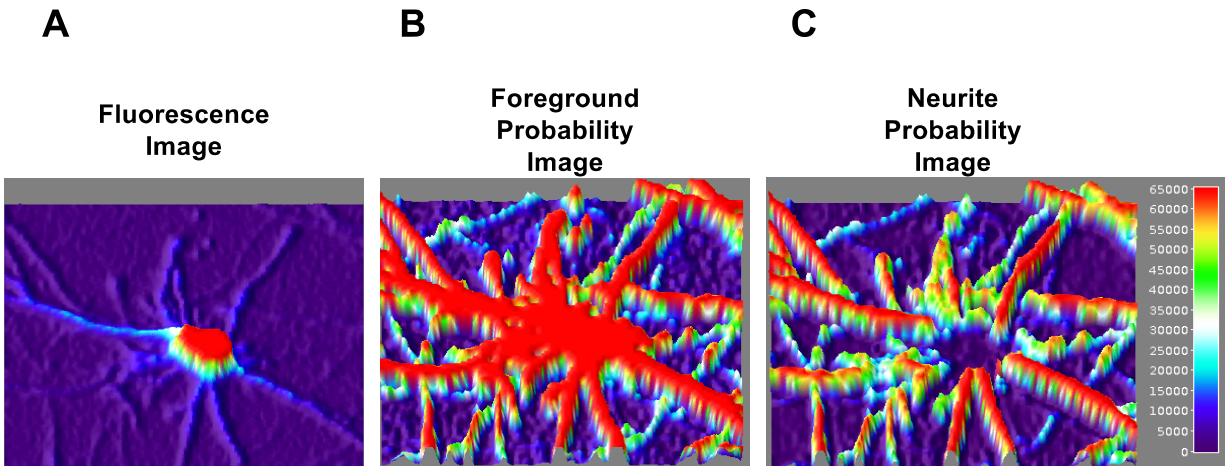
Figure 2.

A   **B**   **C**

Figure 2. **Fluorescence versus probability images. A**. Fluorescence image from mApple, the freely diffusible fluorescent protein that highlights the morphology of the neuron. **B**. Foreground-probability image of the same region highlighting that previously dim neurites have a much higher signal while the background signal remains low. **C**. Neurite-probability image demonstrates exclusion of non-neurite pixels, such as those belonging to the neuron cell body.
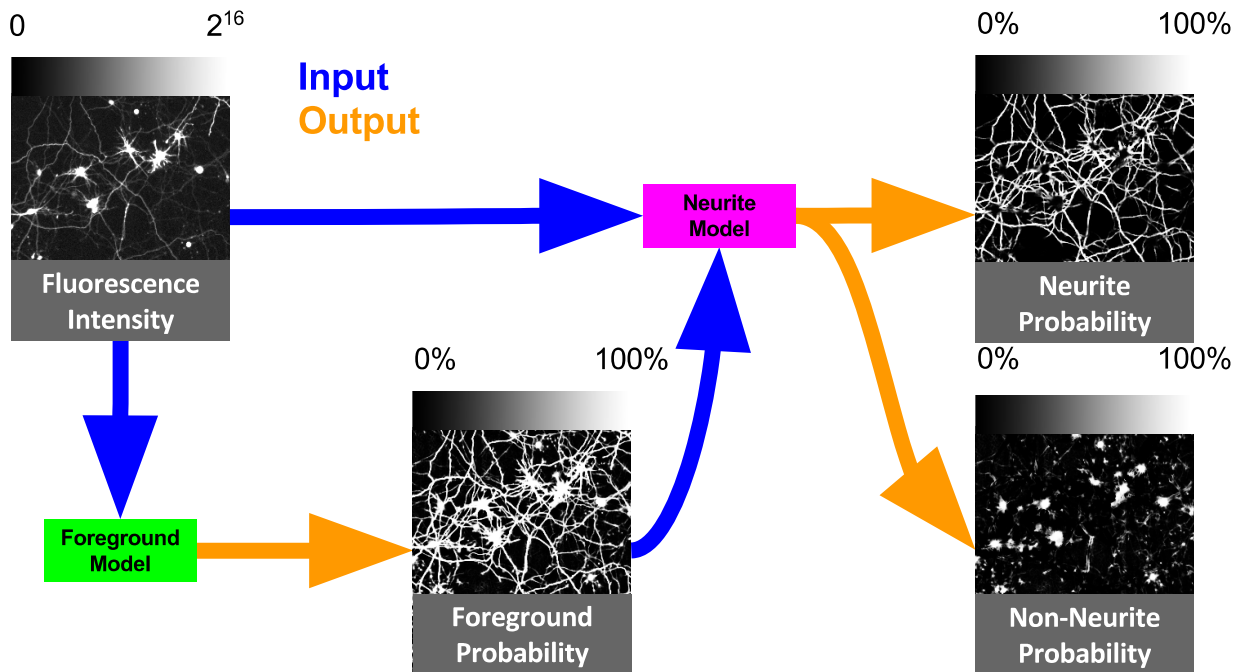
Figure 3.

Figure 3. **Probability images enable a hierarchical approach to segmentation.** For the foreground model, both the input and output images are continuous-valued images, meaning the exact same filters can be applied to the foreground-probability image as the fluorescence image. Thus, both the foreground-probability image and the fluorescence image can be used as input to the neurite model.
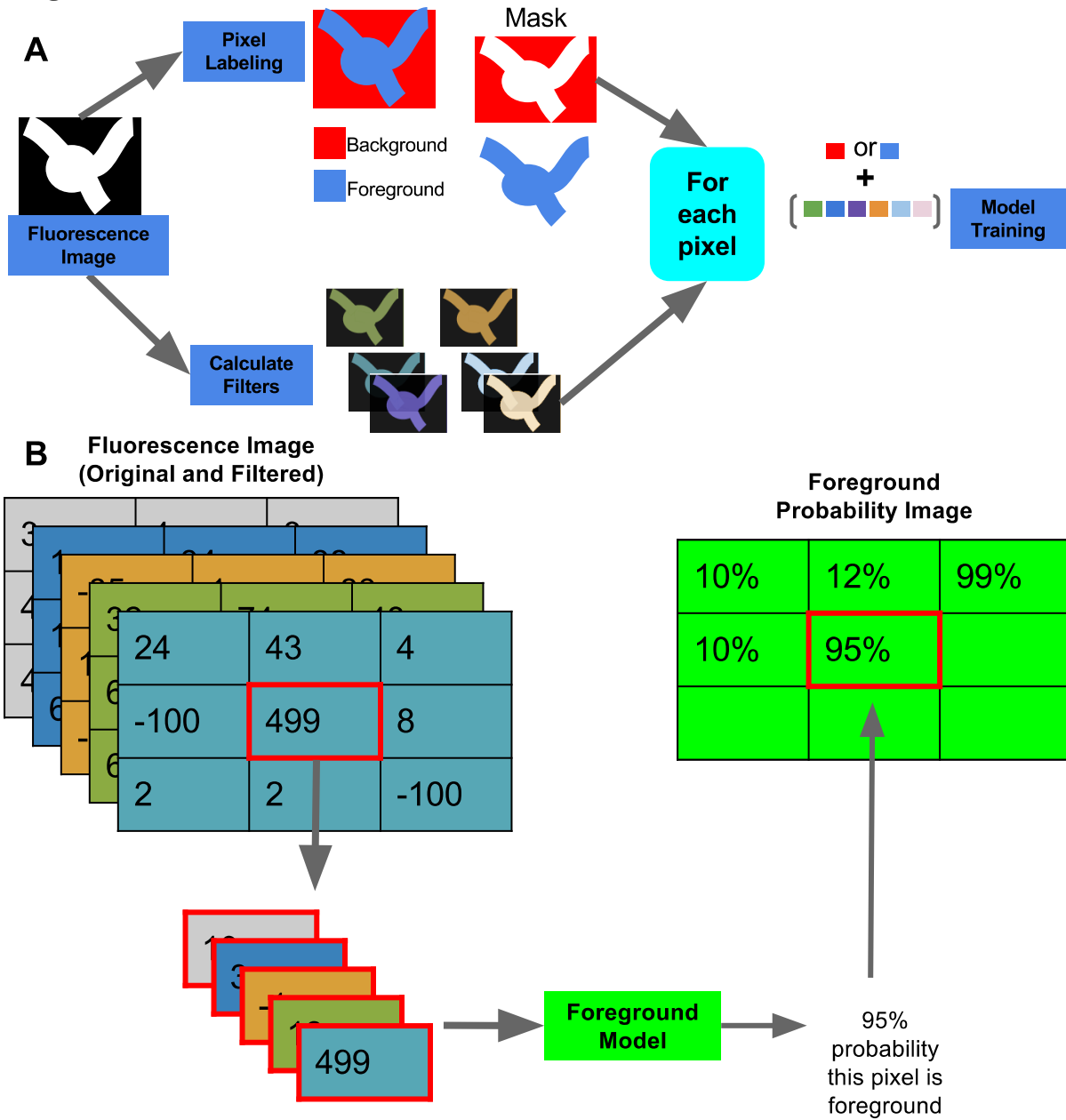
Figure 4.

Figure 4. **Training and evaluating machine-learning models. A**. Workflow of training a supervised machine–learning model. **B**. Evaluation of the foreground-probability model to generate a foreground-probability image.

Figure 5.
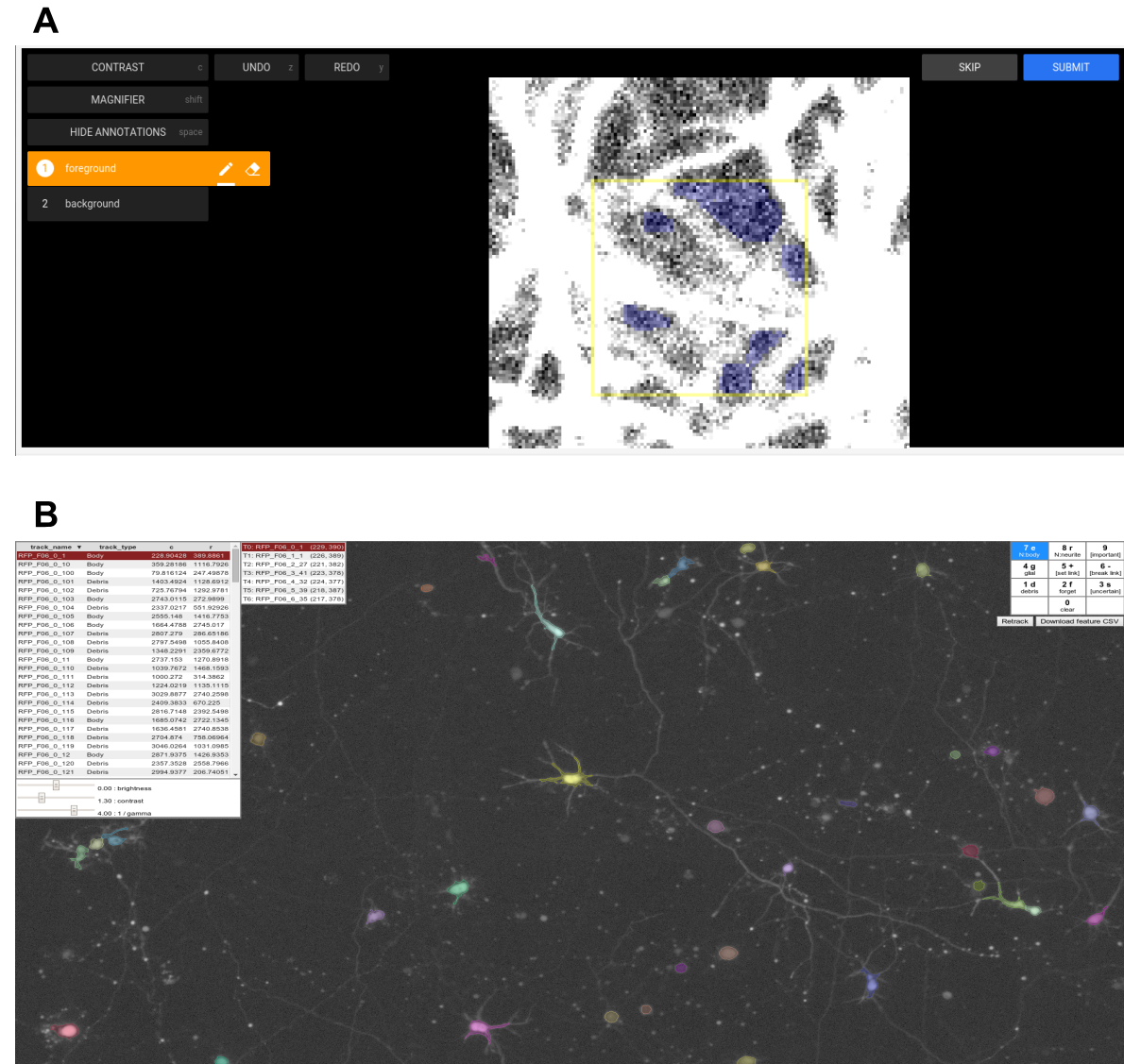
A

B

Figure 5. **Annotation software for Enlighten. A**. The Pixel Labeler is used to paint an image patch with class labels, in this case, foreground and background. The region outside the yellow box is provided for context and is not annotated. **B**. The Cell Labeler is used to annotate and correct the cell type and track from survival experiments analyzed by Enlighten.
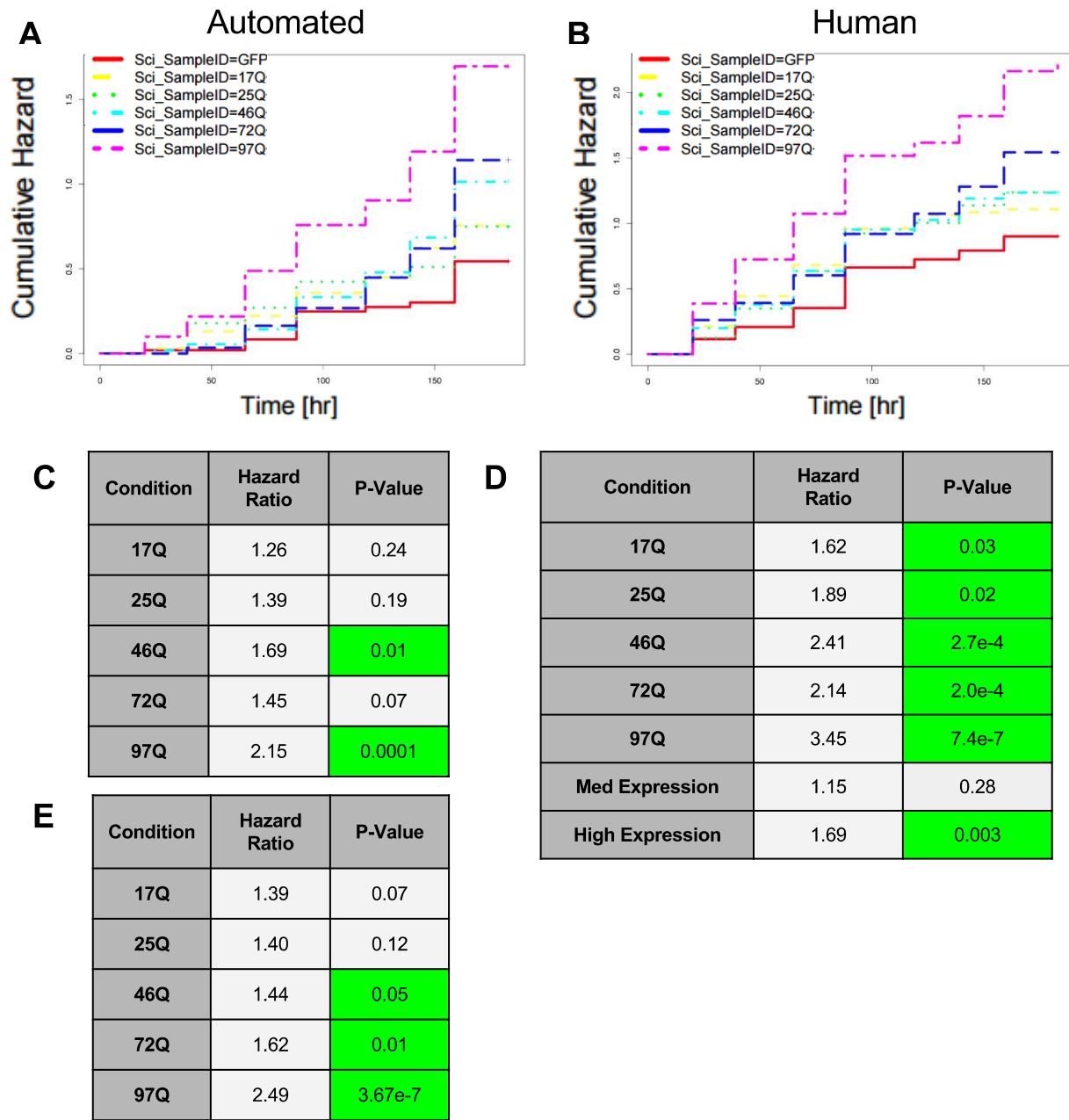
Figure 6.

A — Automated

Cumulative Hazard

- Sci_SampleID=GFP
- Sci_SampleID=17Q
- Sci_SampleID=25Q
- Sci_SampleID=46Q
- Sci_SampleID=72Q
- Sci_SampleID=97Q

Time [hr]

B — Human

Cumulative Hazard

- Sci_SampleID=GFP
- Sci_SampleID=17Q
- Sci_SampleID=25Q
- Sci_SampleID=46Q
- Sci_SampleID=72Q
- Sci_SampleID=97Q

Time [hr]

C

Condition	Hazard Ratio	P-Value
17Q	1.26	0.24
25Q	1.39	0.19
46Q	1.69	0.01
72Q	1.45	0.07
97Q	2.15	0.0001

D

Condition	Hazard Ratio	P-Value
17Q	1.62	0.03
25Q	1.89	0.02
46Q	2.41	2.7e-4
72Q	2.14	2.0e-4
97Q	3.45	7.4e-7
Med Expression	1.15	0.28
High Expression	1.69	0.003

E

Condition	Hazard Ratio	P-Value
17Q	1.39	0.07
25Q	1.40	0.12
46Q	1.44	0.05
72Q	1.62	0.01
97Q	2.49	3.67e-7

Figure 6. **Validation of automated survival analysis by comparison with human-annotated results of condition with EGFP, wild-type Htt, and mutant Htt. A**. Cumulative-hazard curve generated by Enlighten's automated algorithm. **B**. Cumulative-hazard curve generated by human annotation. **C**. Hazard and significance of results by the automated algorithm, including only the poly-glutamine length as a covariate. **D**. Hazard and significance of results by the automated algorithm, including both poly-glutamine length and gene dosage as covariates. **E**. Hazard and significance of results by human annotation, including only poly-glutamine length as a covariate.
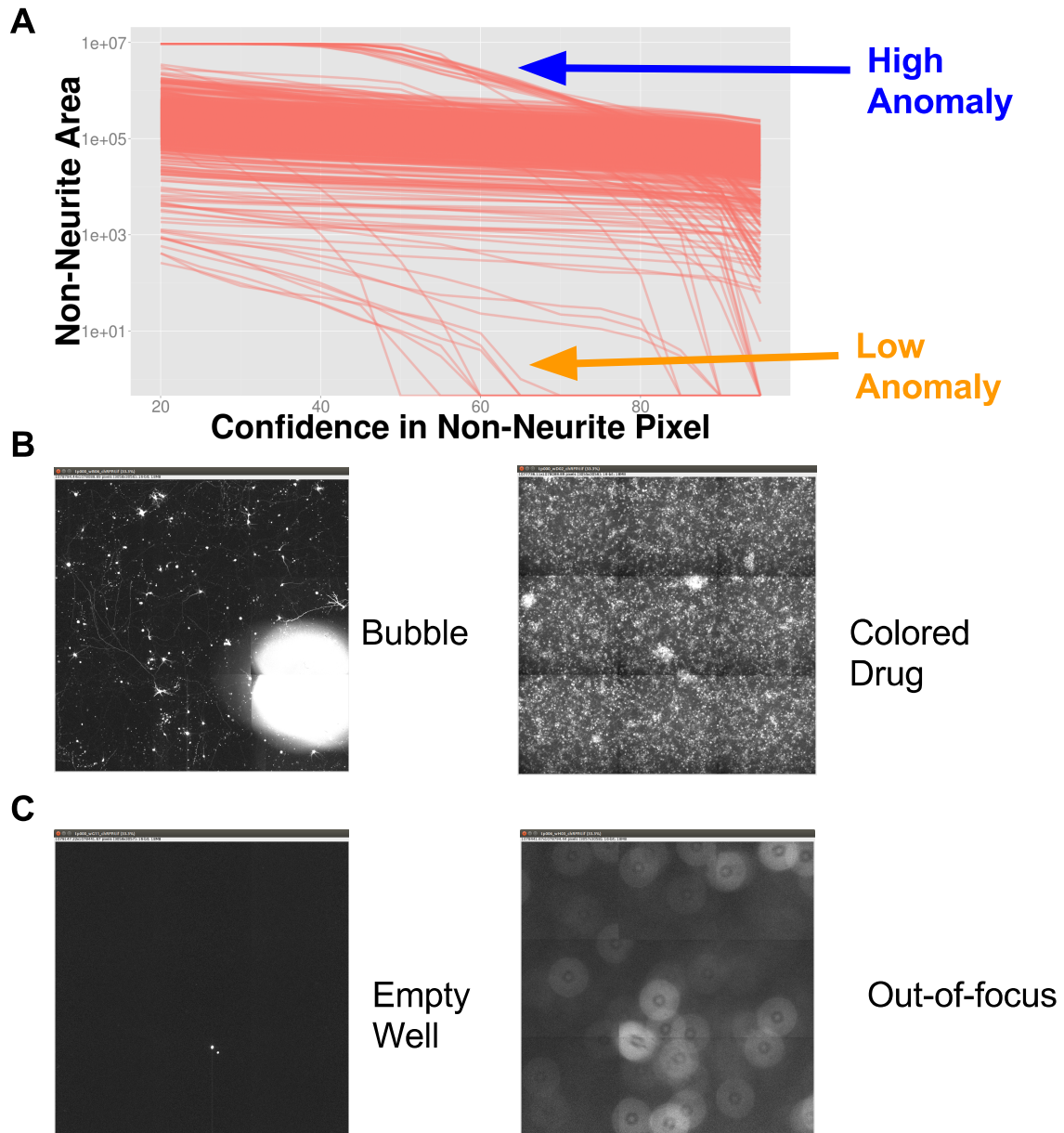
Figure 7.

A

B

C

Figure 7. **Identification of anomalous images using non-neurite-probability images.**
A. Rate of change in the non-neurite area with decreases in the confidence threshold on the non-neurite-probability image can identify anomalous images. The anomalies with high slope and starting area are highlighted by the blue arrow. The anomalies with low starting area are highlighted by the orange arrow. **B**. Anomalies with high slope and starting area include bubbles and wells that have colored drugs. **C**. Anomalies with low-starting area include empty wells and out-of-focus images.
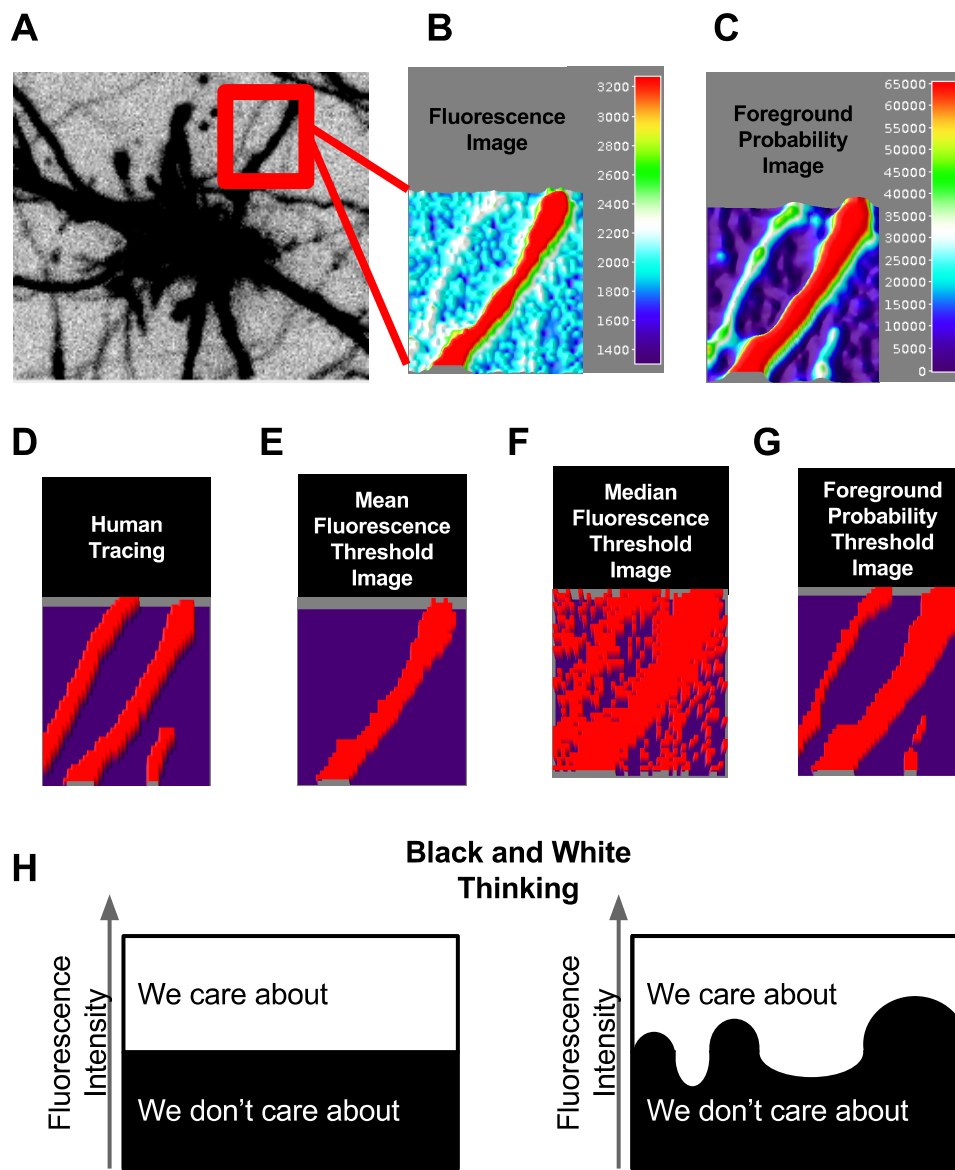
Figure 8.

Figure 8. **Improved foreground sensitivity and selectivity in probability images compared to fluorescence images. A**. Inverted fluorescence image shows both dim and bright neurites, highlighted by the red square. **B**. 3D-color heatmap of the fluorescence image contained within the red square in A. Note that the pixel intensities of the dim neurite are very similar to those of nearby background regions. **C**. 3D-color heatmap of the foreground-probability image contained within the red square in A. Note that the pixel intensities of the dim neurite are now distinct from the nearby background regions. **D**. Human segmentation of the red square in A. **E**. Using the mean of the fluorescence image as a threshold can capture the brighter neurite but not the dim neurite. **F**. Using the median of the fluorescence image as a threshold can capture pixels from

both the dim and bright neurite, but it also includes background pixels. **G**. Using the mean of the foreground-probability image can capture both the dim and bright neurite while still excluding background pixels. **H**. Thresholding is black-and-white thinking. It can only work well when the two conditions we care about do not have overlap.
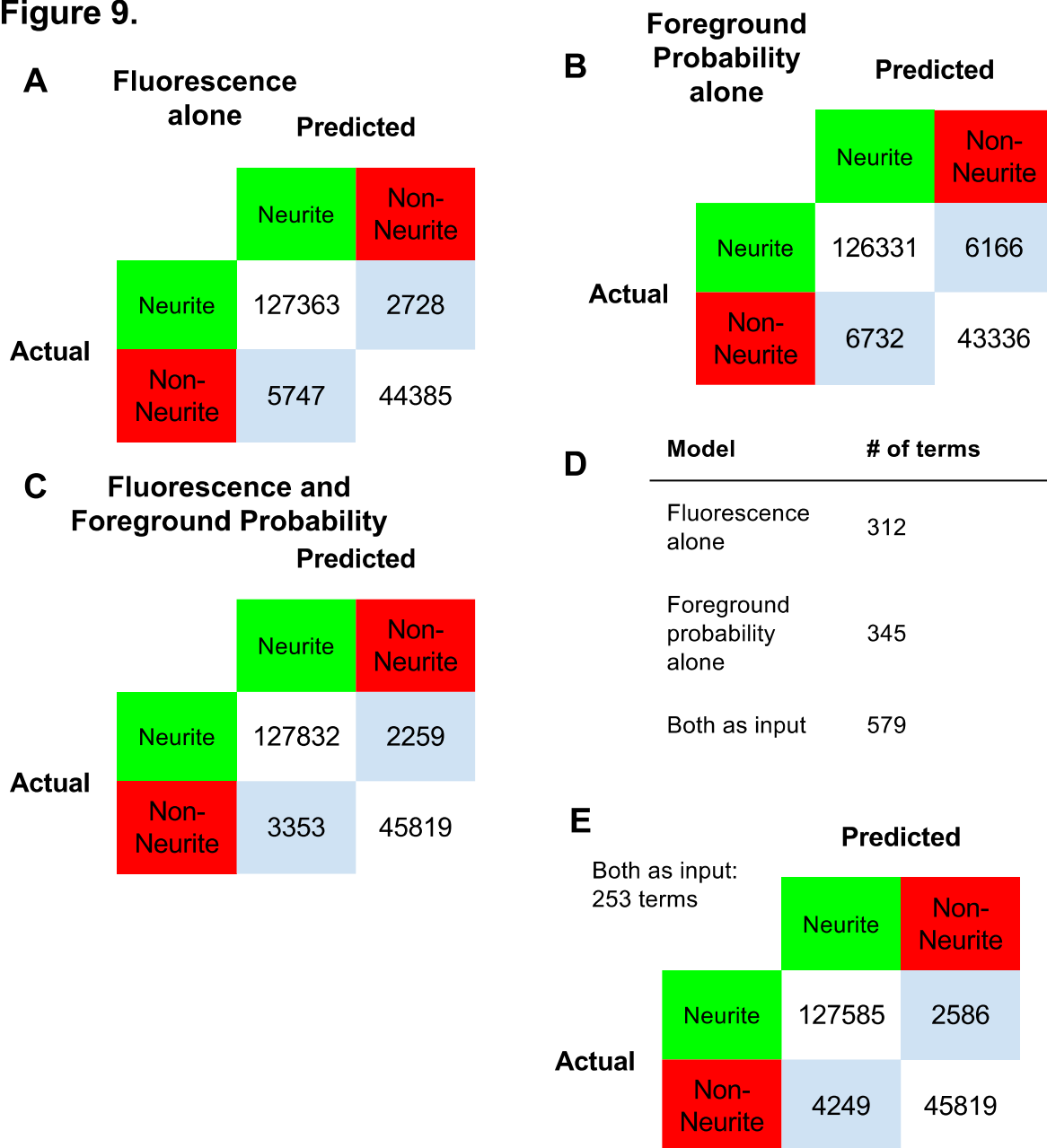
Figure 9.

Figure 9. **Increase in the accuracy of predicting neurites when using both the fluorescence image and foreground-probability image as inputs**. **A**. The accuracy of predicting neurites per pixel with glmnet model using only the fluorescence image as input. **B**. The accuracy of predicting neurites per pixel with glmnet model using only the foreground probability image as input. This model performed worse than the model trained in A for both false positives and false negatives. **C**. The accuracy of predicting neurites per pixel with glmnet model using both the fluorescence image and foreground-probability image. Compared to the models trained in A and B, this model has the lowest number of false positives and false negatives. **D**. Number of terms in each glmnet model. The glmnet model trained with both input

images also had more terms. **E**. Increasing the glmnet penalty for the model trained with both input images can decrease the number of terms in the model to below the number for any model trained in A or B. Although this model performs worse than the full-term model trained in C, it still outperformed either of the models trained in A or B.
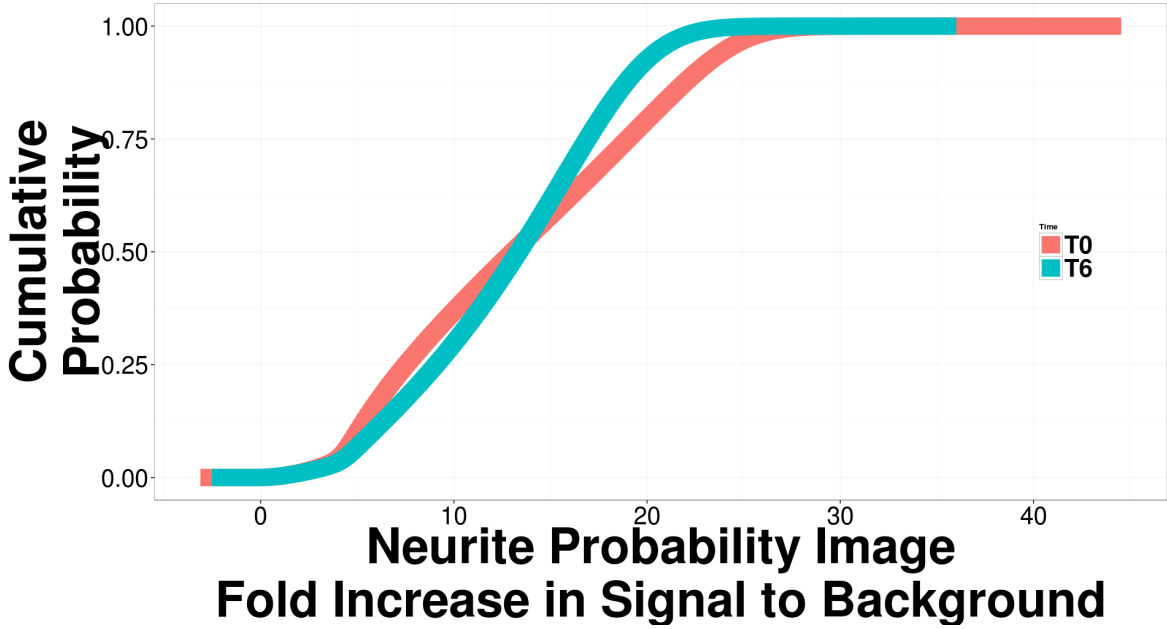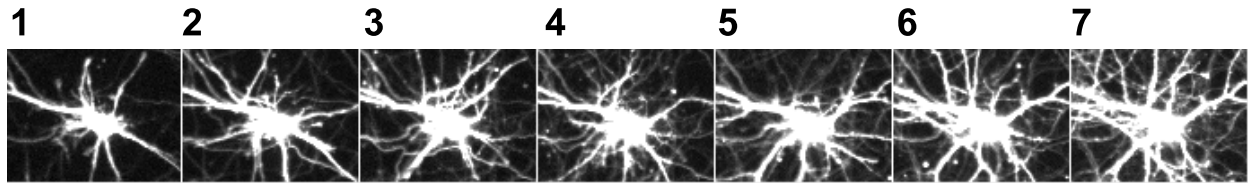
Figure 10.

Figure 10. Increase in signal-to-background in the neurite-probability image compared to the fluorescence image. The cumulative histogram of the per-pixel increase in signal-to-background for the neurite-probability image compared to the fluorescence image, which demonstrates that most of the neurite pixels are being boosted greater than 10-fold. This effect occurs at both early and late timepoints.
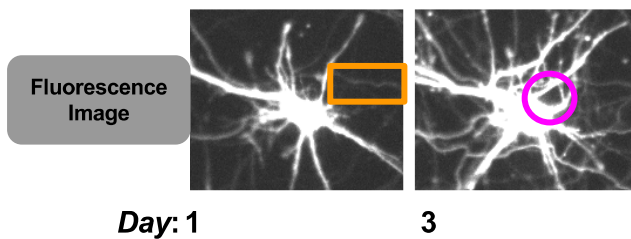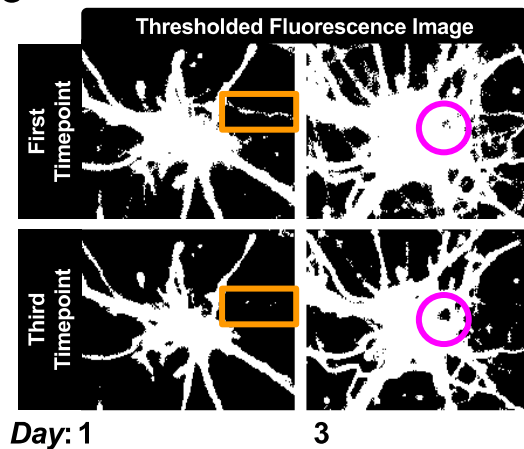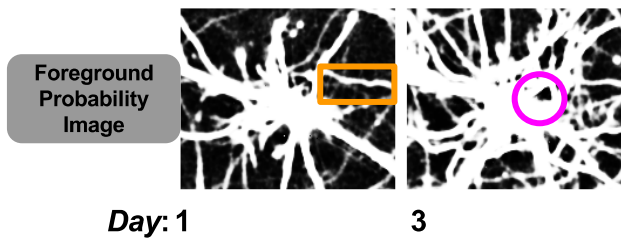
Figure 11.

A

Day:

1   **2**   **3**   **4**   **5**   **6**   **7**

B

Fluorescence Image

Day: 1   3

C

Thresholded Fluorescence Image

First Timepoint

Third Timepoint

Day: 1   3

D

Foreground Probability Image

Day: 1   3

E

Thresholded Probability Image

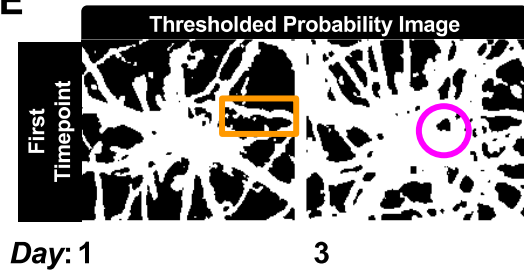First Timepoint

Day: 1   3

Figure 11. Longitudinal foreground sensitivity and selectivity in fluorescence images compared to probability images. **A**. Fluorescence intensity of objects increases over time. **B**. A dim neurite (orange rectangle) and background patch (purple circle) near a bright neuron in a fluorescence image. **C**. Thresholding the fluorescence images, using the mean intensity of the first timepoint, preserves the dim neurite at the first timepoint (orange rectangle) but cannot capture the background (purple circle) near the bright neuron. Thresholding the fluorescence images using the mean intensity on the third timepoint loses the dim neurite at the first timepoint (orange rectangle), but preserves the background (purple circle) near the bright neuron. **D**. A dim neurite (orange rectangle) and background patch (purple circle) near a bright neuron in a foreground-probability image. **E**. Thresholding the foreground-probability image using the mean

intensity at the first timepoint can capture both the dim neurite at the first timepoint (orange rectangle) and the background near the bright neuron at the third timepoint (purple circle).
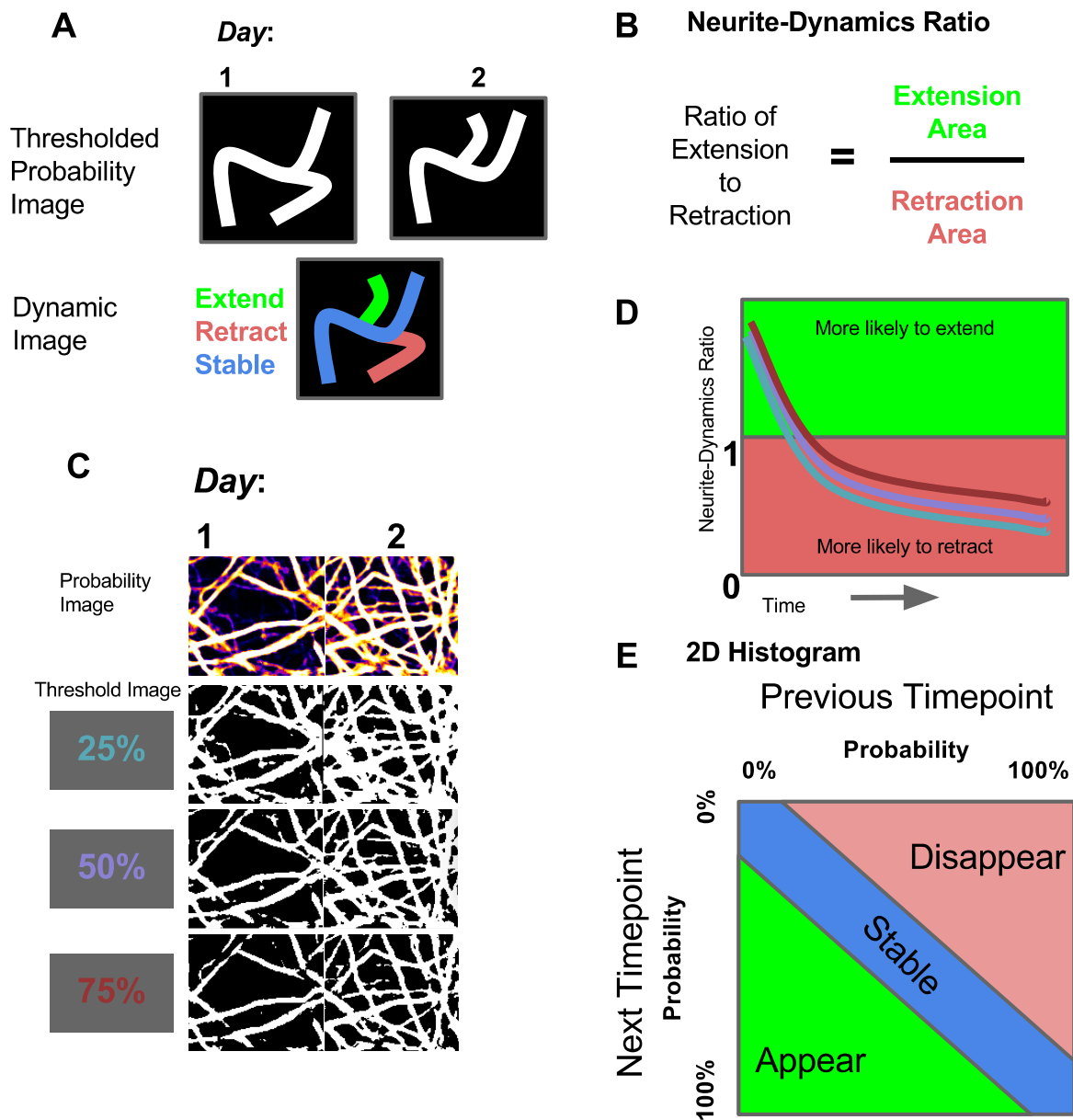
Figure 12.

A

Day:

1   **2**

Thresholded Probability Image

Dynamic Image

Extend
Retract
Stable

B   **Neurite-Dynamics Ratio**

$$\text{Ratio of Extension to Retraction} = \frac{\textcolor{green}{\text{Extension Area}}}{\textcolor{red}{\text{Retraction Area}}}$$

D

Neurite-Dynamics Ratio

More likely to extend

1

More likely to retract

0   Time

C   **Day**:

1   **2**

Probability Image

Threshold Image

25%

50%

75%

E   **2D Histogram**

Previous Timepoint

Probability

0%   100%

0%

Next Timepoint Probability

100%

Disappear

Stable

Appear

Figure 12. **Neurite-dynamics ratio as a summary of neurite extension and retraction**. **A**. Identification of neurites that extend and retract with positional correspondence and per-pixel changes. **B**. Definition of the neurite-dynamics ratio. **C**. Changing the threshold of the neurite-probability image can cause small shifts in the measured area. **D**. Example graph of the neurite-dynamics ratio over time using the three thresholds shown in C. **E**. Neurite extension, retraction, and stasis can be captured using a 2D histogram of the neurite-probability images from the earlier timepoint and the later timepoint.
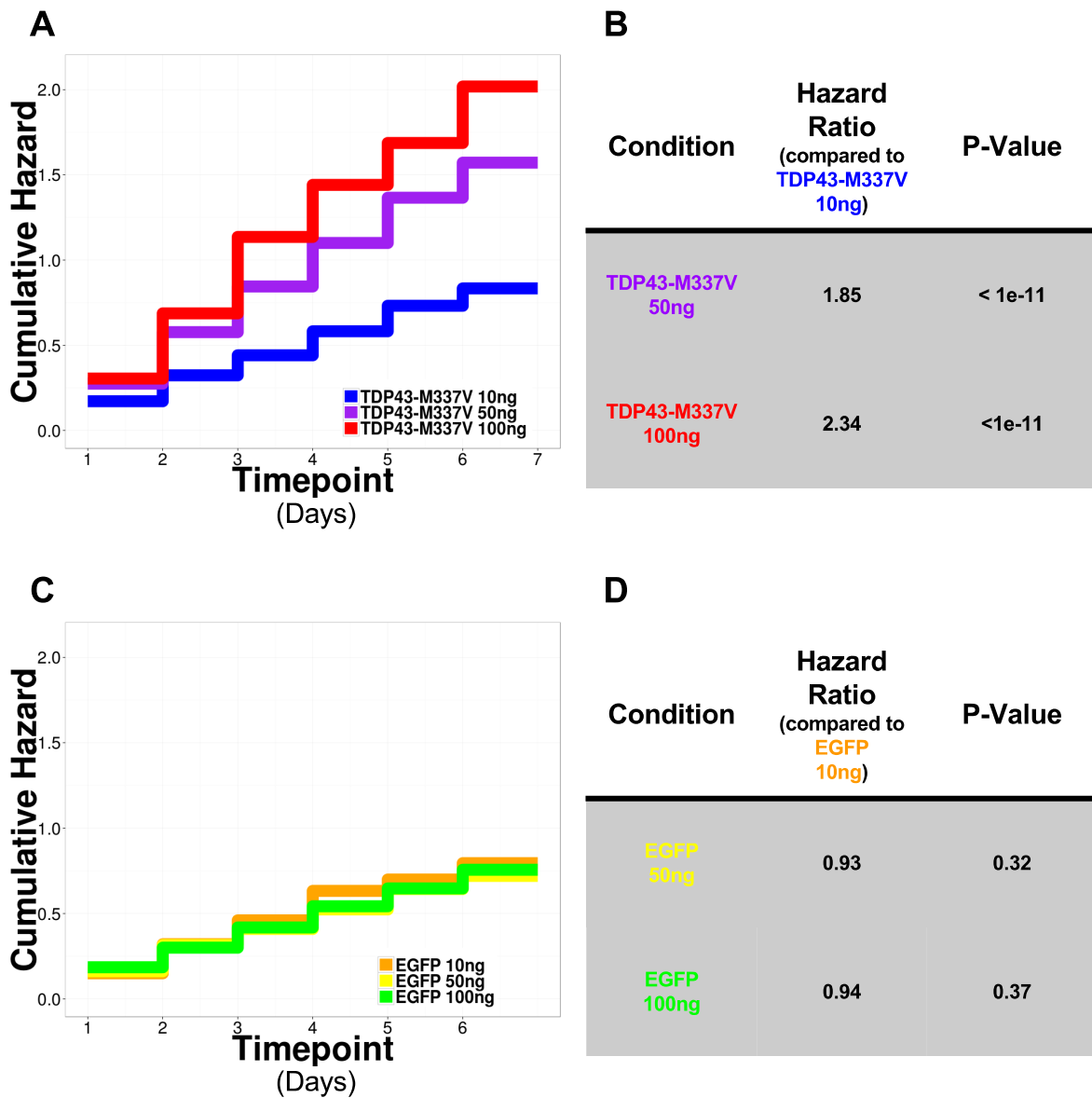
Figure 13.

A

B

Condition	Hazard Ratio (compared to TDP43-M337V 10ng)	P-Value
TDP43-M337V 50ng	1.85	< 1e-11
TDP43-M337V 100ng	2.34	<1e-11

C

D

Condition	Hazard Ratio (compared to EGFP 10ng)	P-Value
EGFP 50ng	0.93	0.32
EGFP 100ng	0.94	0.37

Figure 13. Dose-dependent increase in the risk of cell death for TDP43-M337V but not EGFP. **A**. Cumulative-hazard curve of three doses of TDP43-M337V show increasing risk with gene dose. **B**. Significant increase in the hazard ratio for TDP43-M337V in a dose-dependent fashion. **C**. Cumulative-hazard curve of three doses of EGFP show no dose-dependent change. **D**. No significant increase in the hazard ratio for EGFP with increasing dose.
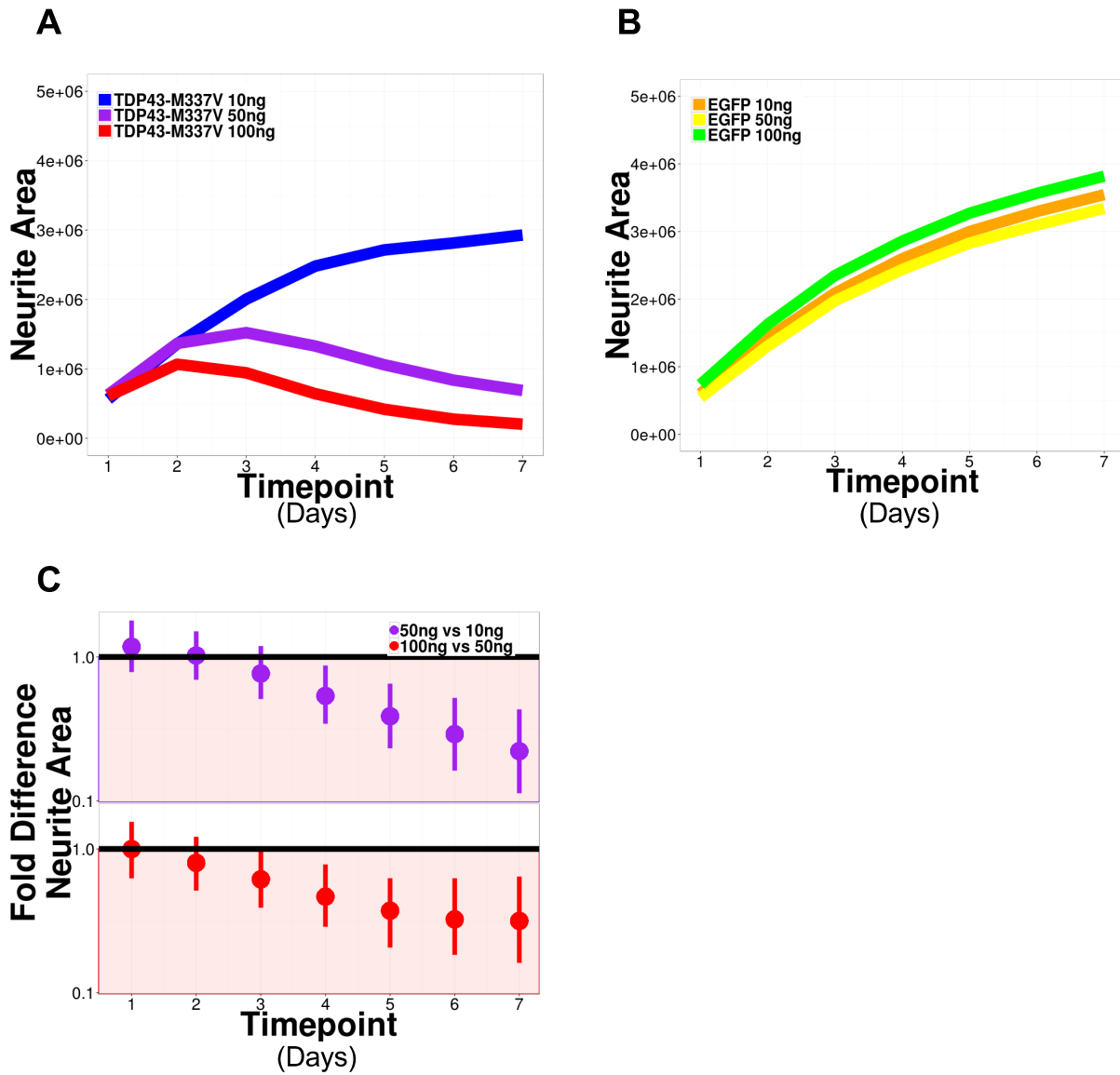
Figure 14.

A

B

C

Figure 14. **Dose-dependent decrease in neurite area for TDP43-M337V but not EGFP**.
A. Mean neurite area for three doses of TDP43-M337V. **B**. Neurite area for three doses of
EGFP. **C**. Fold difference in neurite area for increasing doses of TDP43-M337V. The dot
indicates the mean fold-difference between the conditions, and the line indicates the 95%
credible interval.

Figure 15.

Figure 15. **Neurite-dynamics ratio shows a dose-dependent bias towards neurite retraction for TDP43-M337V but not EGFP**. **A**. Neurite-dynamics ratio for three doses of TDP43-M337V. **B**. Fold-difference in neurite-dynamics ratio for increasing doses of TDP43-M337V. The dot indicates the mean fold-difference between the con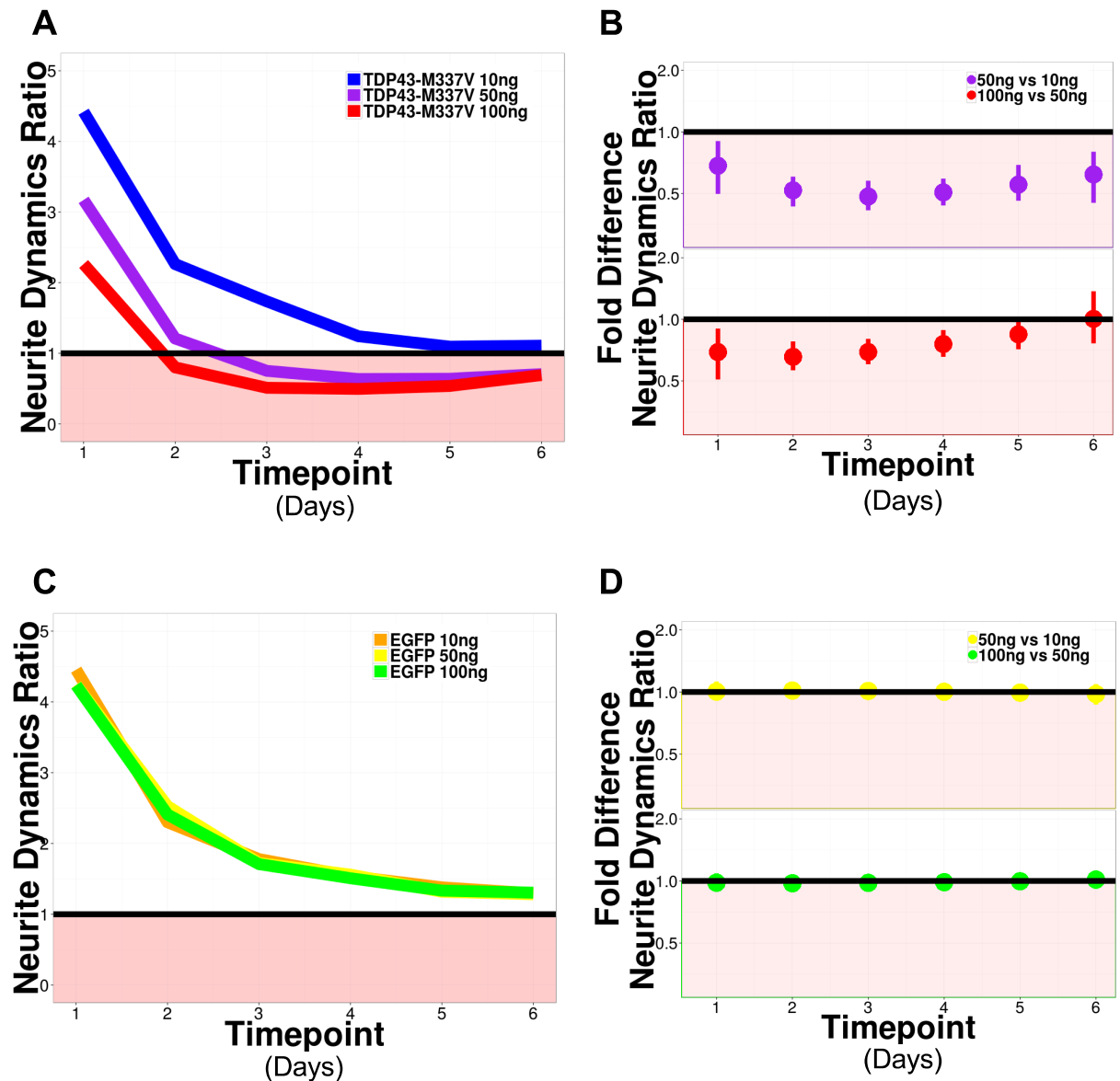ditions, and the line indicates the 95% credible interval. **C**. Neurite-dynamics ratio for three doses of EGFP. **D**. Fold-difference in neurite-dynamics ratio for increasing doses of EGFP. The dot indicates the mean fold-difference between the conditions, and the line indicates the 95% credible interval.
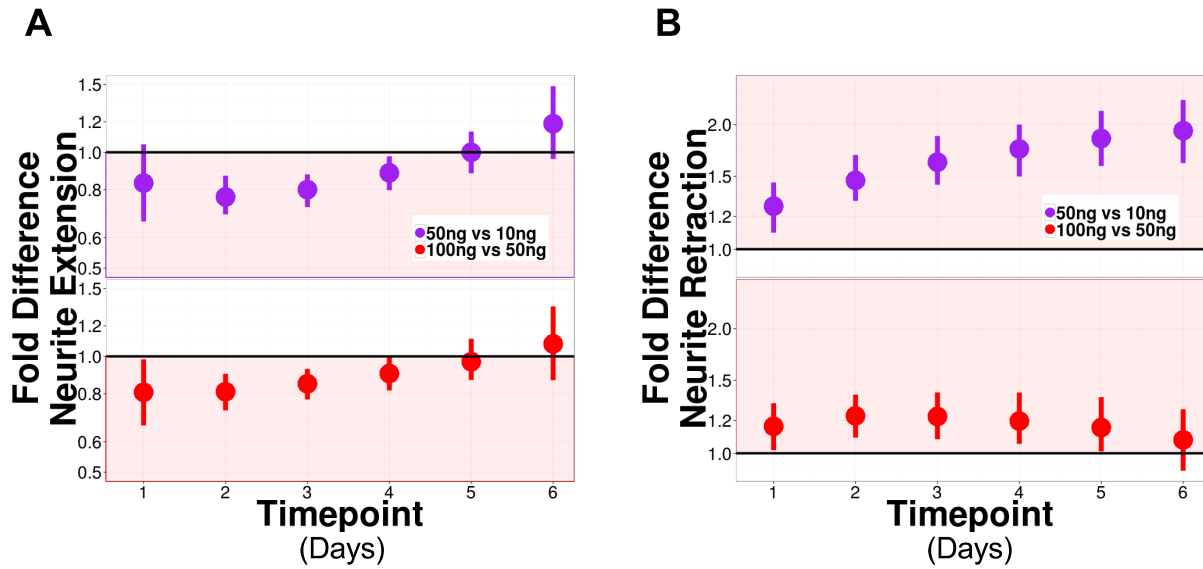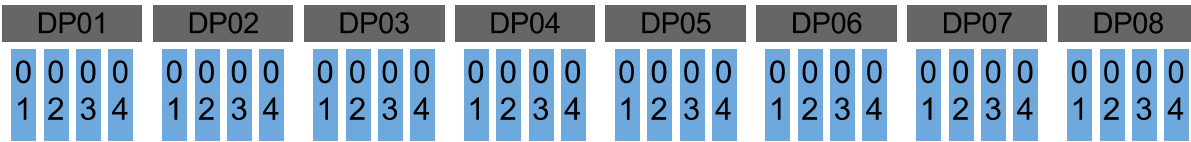
Figure 16.

A

B

Figure 16. **Comparison of the dose-dependent effect of TDP43-M337V on neurite extension and retraction**. **A**. Neurite extension is decreased at earlier timepoints. The dot indicates the mean fold-difference between the conditions, and the line indicates the 95% credible interval. **B**. Neurite retraction shows a prolonged increase. The dot indicates the mean fold-difference between the conditions, and the line indicates the 95% credible interval.
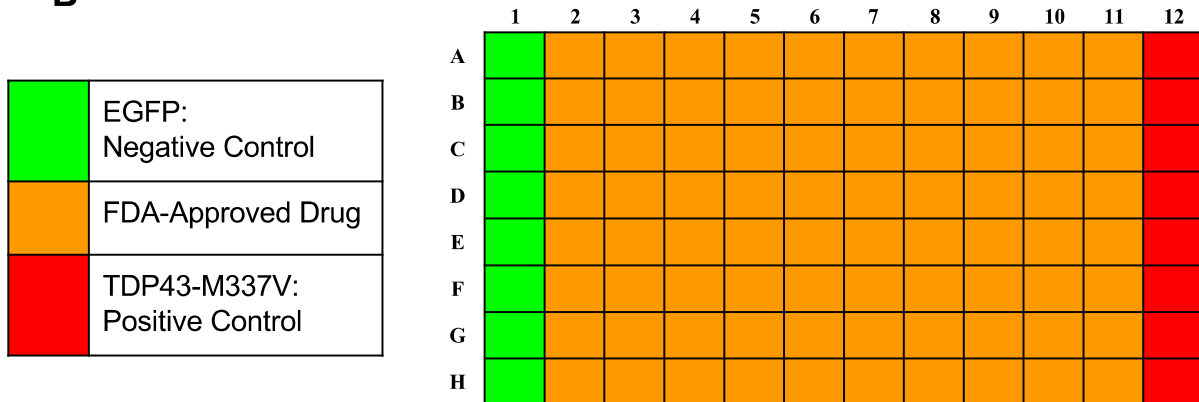
Figure 17.

A

DP01	DP02	DP03	DP04	DP05	DP06	DP07	DP08

01 02 03 04 01 02 03 04 01 02 03 04 01 02 03 04 01 02 03 04 01 02 03 04 01 02 03 04 01 02 03 04

B

	EGFP: Negative Control
	FDA-Approved Drug
	TDP43-M337V: Positive Control

Figure 17. **ALS-TDI primary screen**. **A**. Schematic of experimental organization. **B**. Plate layout highlighting control and test wells.
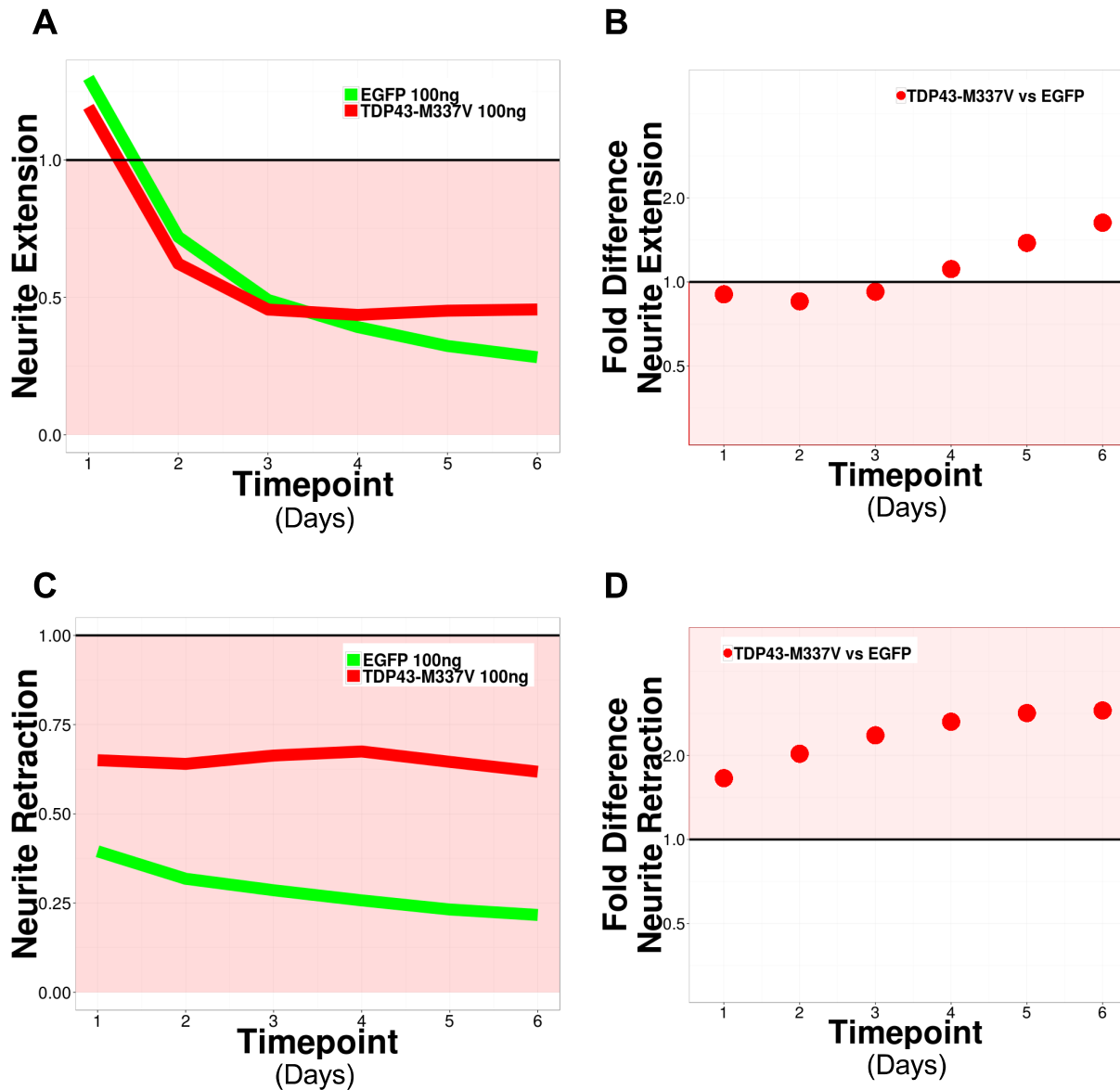
Figure 18.

Figure 18. **Comparison of neurite extension and retraction between TDP43-M337V and EGFP**. **A**. Neurite extension is decreased in TDP43-M337V at early timepoints but higher at later timepoints. **B**. Quantification of fold-difference in neurite extension. The dot indicates the mean fold-difference between the conditions, and the line indicates the 95% credible interval. **C**. Neurite retraction is consistently increased in TDP43-M337V. **D**. Quantification of fold-difference in neurite retraction. The dot indicates the mean fold-difference between the conditions, and the line indicates the 95% credible interval.
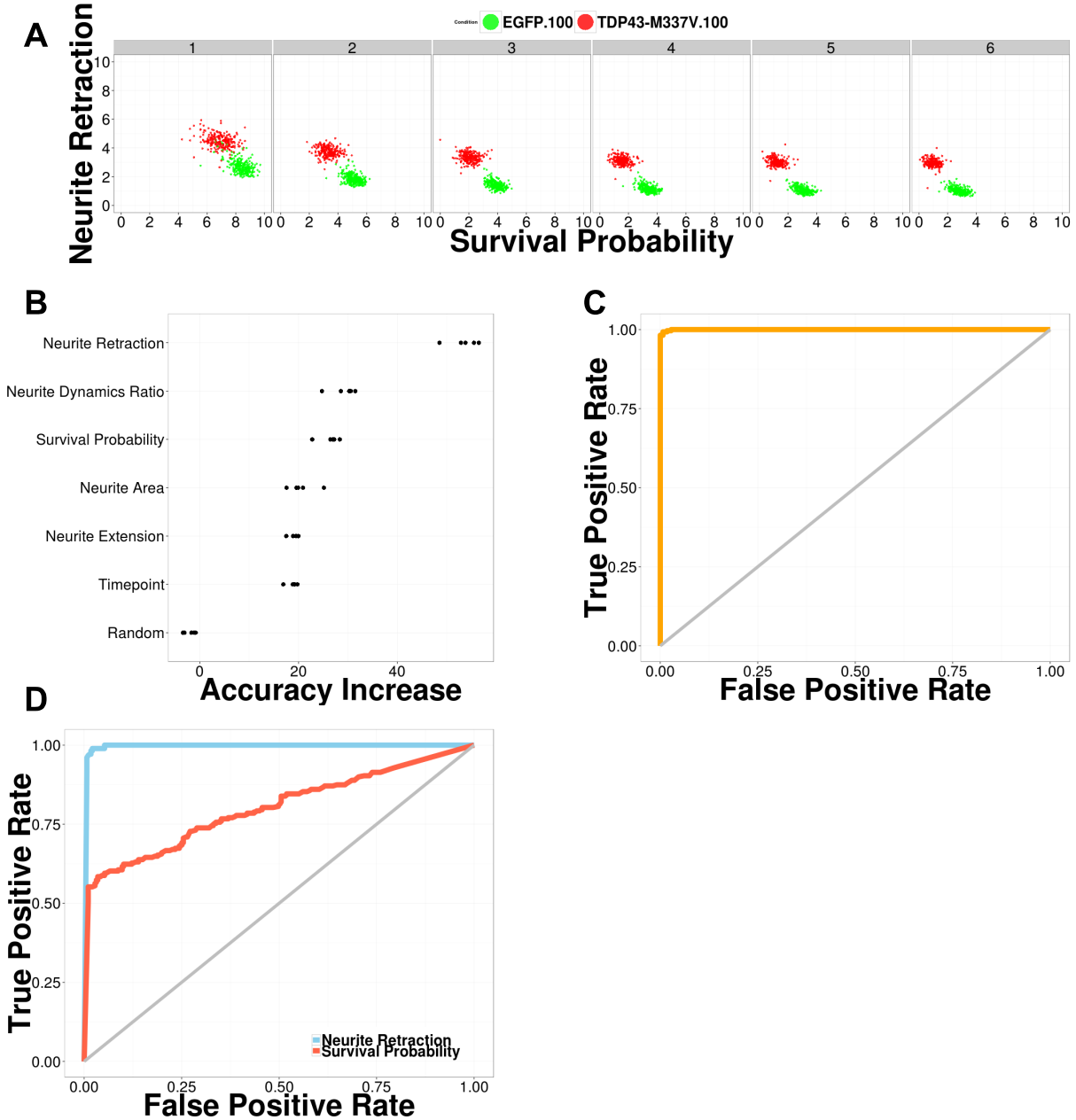
Figure 19.

Figure 19. **Measurements of neurite dynamics more accurately identifies TDP43-M337V and EGFP well conditions than neuron survival**. **A**. Plot of the neurite retraction and survival probability for each well over time. Each panel represents the single timepoint labeled by the gray header. **B**. Contribution of each variable to accuracy of RF model. **C**. ROC curve for RF model with all terms. **D**. ROC curve for a single-term model of just neurite retraction or survival probability. RF, random forest; ROC, Receiver operating curve.
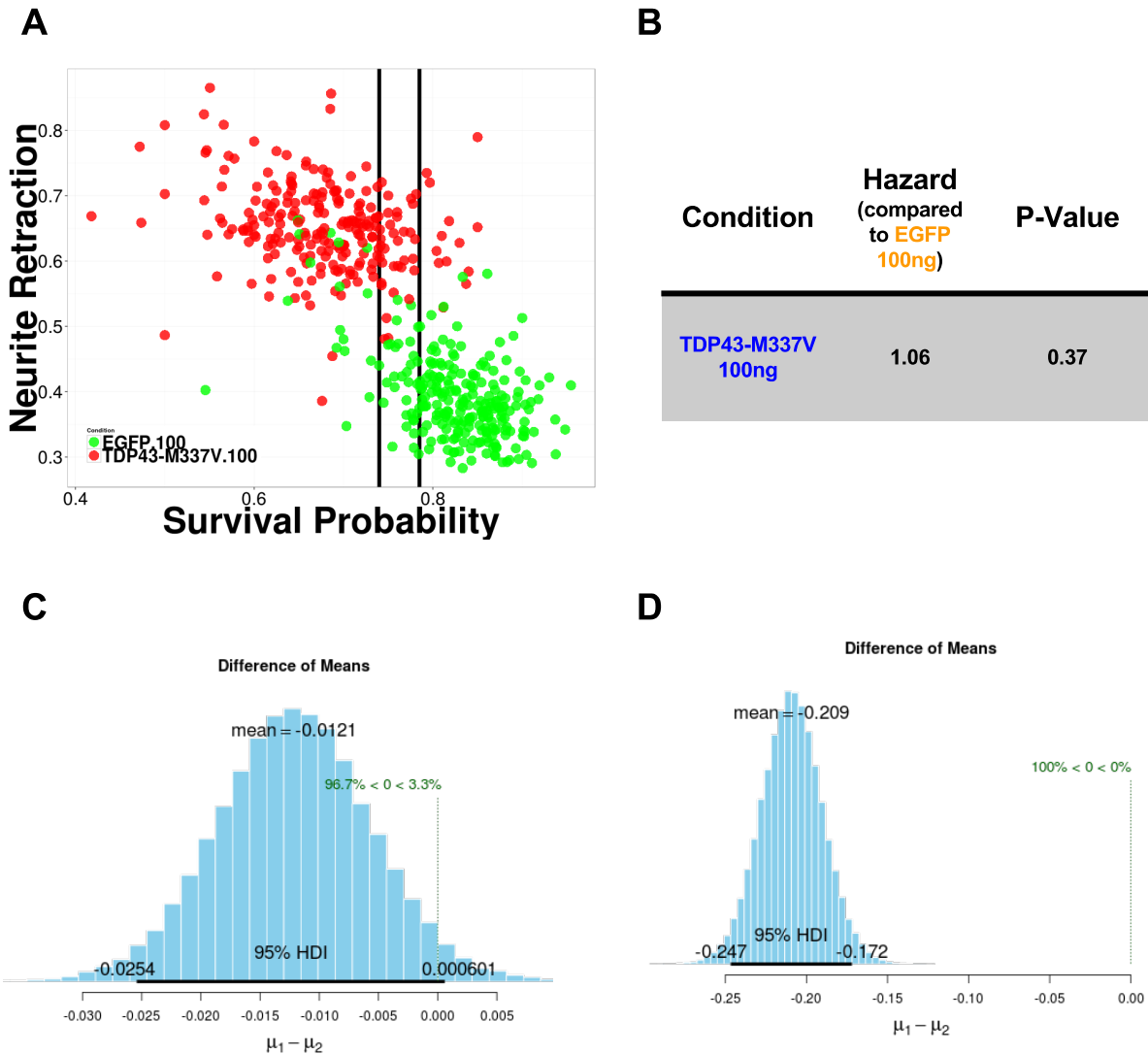
Figure 20.

A

B

Condition	Hazard (compared to EGFP 100ng)	P-Value
TDP43-M337V 100ng	1.06	0.37

C

Difference of Means

mean = -0.0121

96.7% < 0 < 3.3%

95% HDI

-0.0254   0.000601

$\mu_1 - \mu_2$

D

Difference of Means

mean = -0.209

100% < 0 < 0%

95% HDI

-0.247   -0.172

$\mu_1 - \mu_2$

Figure 20. **Neurite retraction can differentiate between TDP43-M337V and EGFP under synthetic conditions with no survival difference. A**. 2D plot of neurite retraction and survival probability without normalization, with lines showing the boundaries of survival probability. (EGFP: 32 wells, TDP43-M337V: 33 wells) **B**. Hazard ratio for survival probability. **C**. Mean difference in survival probabilities using BEST technique. **D**. Mean difference in neurite retraction using BEST technique. BEST, Bayesian estimation supercedes the t-test.
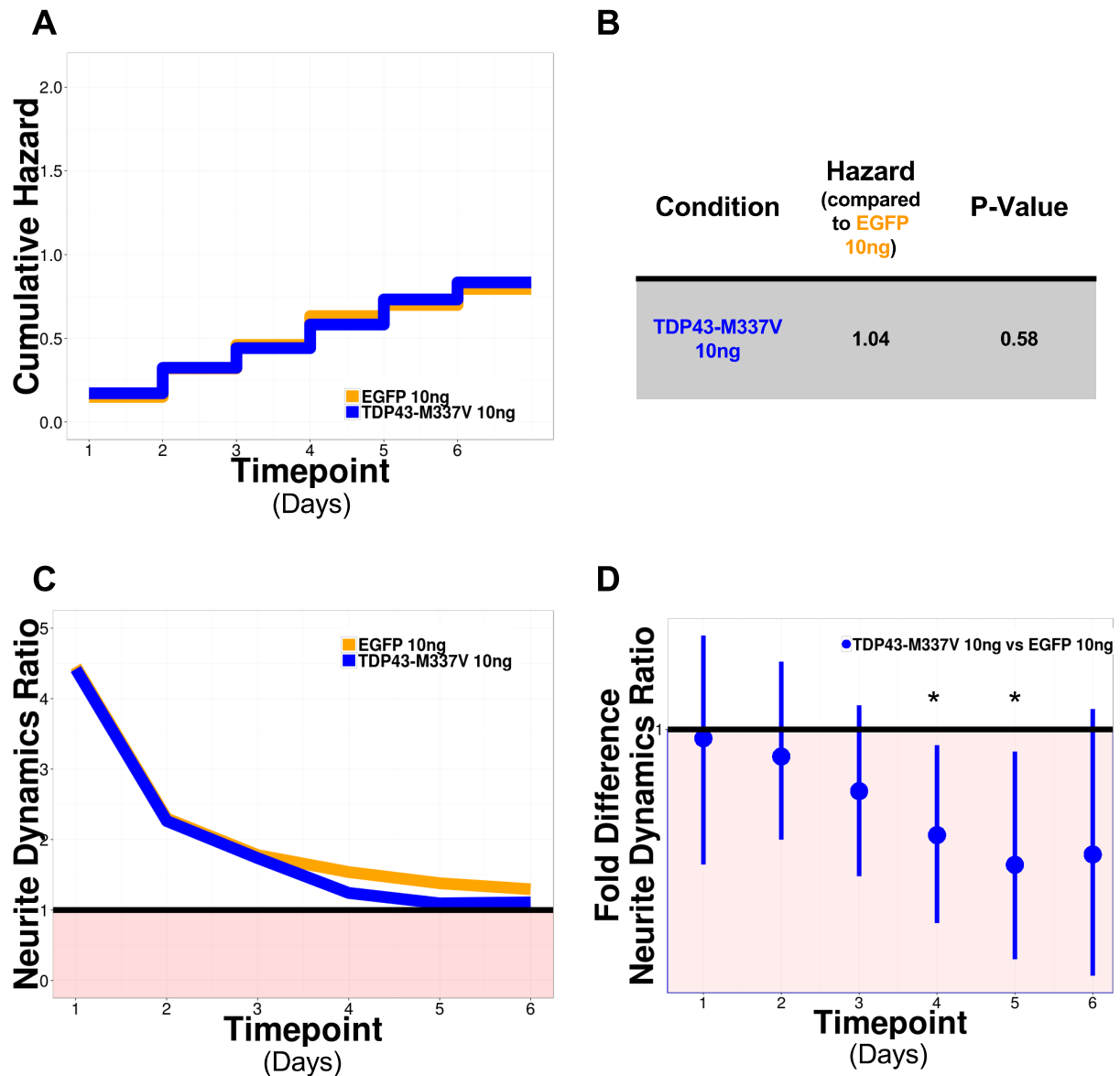
Figure 21.

A

B

Condition	Hazard (compared to EGFP 10ng)	P-Value
TDP43-M337V 10ng	1.04	0.58

C

D

Figure 21. **Neurite-dynamics ratio can differentiate between low concentrations of TDP43-M337V and EGFP when no survival difference is present.**
A. Cumulative hazard of low doses for both TDP43-M337V and EGFP are similar over time. **B**. No significant difference in hazard ratio. **C**. Neurite-dynamics ratio for TDP43-M337V and EGFP over time. D. Fold-difference in neurite dynamics ratio shows significance for two later timepoints.
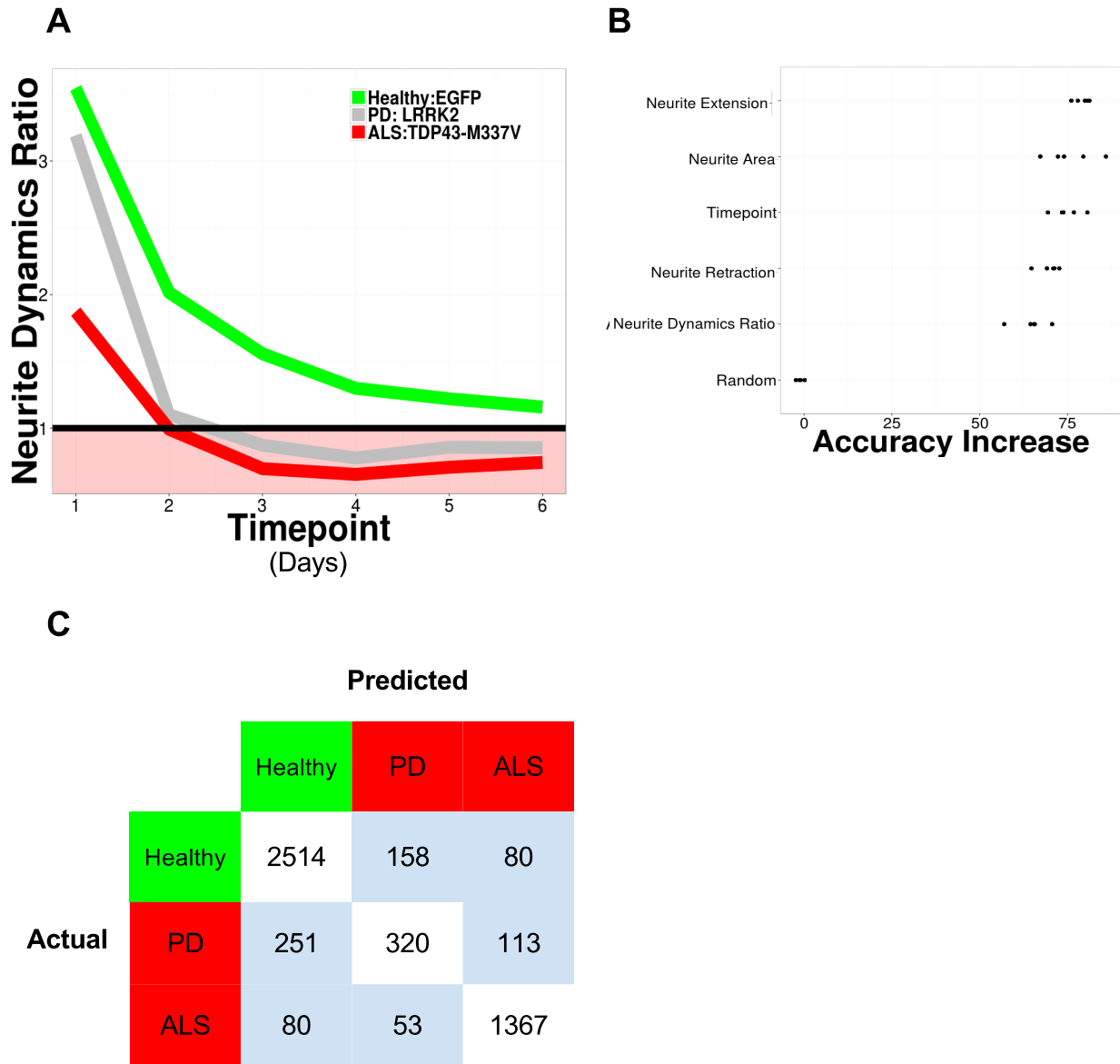
Figure 22.

A

B

C

Figure 22. **Measurements of neurite dynamics enable differentiation of PD and ALS models**. **A**. Neurite-dynamics ratio for a healthy condition and two disease models over time. **B**. Variables required for RF differ between models. **C**. Confusion table of PD versus ALS models. Each cell in the table represents the number of images with that actual and predicted classification. RF, random forest.

Bibliography

Albalate, A., and Minker, W. (2011). Semi-supervised and unsupervised machine learning: novel strategies (London : Hoboken, NJ: ISTE ; Wiley).

Arai, T., Hasegawa, M., Akiyama, H., Ikeda, K., Nonaka, T., Mori, H., Mann, D., Tsuchiya, K., Yoshida, M., Hashizume, Y., et al. (2006). TDP-43 is a component of ubiquitin-positive tau-negative inclusions in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. Biochem. Biophys. Res. Commun. *351*, 602–611.

Arrasate, M., and Finkbeiner, S. (2005). Automated microscope system for determining factors that predict neuronal fate. Proc. Natl. Acad. Sci. U. S. A. *102*, 3840–3845.

Arrasate, M., Mitra, S., Schweitzer, E.S., Segal, M.R., and Finkbeiner, S. (2004). Inclusion body formation reduces levels of mutant huntingtin and the risk of neuronal death. Nature *431*, 805–810.

Barmada, S.J., Skibinski, G., Korb, E., Rao, E.J., Wu, J.Y., and Finkbeiner, S. (2010). Cytoplasmic Mislocalization of TDP-43 Is Toxic to Neurons and Enhanced by a Mutation Associated with Familial Amyotrophic Lateral Sclerosis. J. Neurosci. *30*, 639–649.

Billinton, N., and Knight, A.W. (2001). Seeing the Wood through the Trees: A Review of Techniques for Distinguishing Green Fluorescent Protein from Endogenous Autofluorescence. Anal. Biochem. *291*, 175–197.

Bradski, G.R., and Kaehler, A. (2008). Learning OpenCV: computer vision with the OpenCV library (Sebastopol, Calif: O'Reilly).

Brown, K.M., Barrionuevo, G., Canty, A.J., Paola, V.D., Hirsch, J.A., Jefferis, G.S.X.E., Lu, J., Snippe, M., Sugihara, I., and Ascoli, G.A. (2011). The DIADEM Data Sets: Representative Light Microscopy Images of Neuronal Morphology to Advance Automation of Digital Reconstructions. Neuroinformatics *9*, 143–157.

Charoenkwan, P., Hwang, E., Cutler, R.W., Lee, H.-C., Ko, L.-W., Huang, H.-L., and Ho, S.-Y. (2013). HCS-Neurons: identifying phenotypic changes in multi-neuron images upon drug treatments of high-content screening. BMC Bioinformatics *14*, S12.

Crainiceanu, C., Ruppert, D., and Wand, M.P. (2005). Bayesian analysis for penalized spline regression using WinBUGS.

Danuser, G. (2011). Computer Vision in Cell Biology. Cell *147*, 973–978.

Daub, A.C. (2013). An Automated Microscope System to Monitor Dynamic Stress Responses in Neurons. Ph.D. University of California, San Francisco.

Daub, A., Sharma, P., and Finkbeiner, S. (2009). High-content screening of primary neurons: ready for prime time. Curr. Opin. Neurobiol. *19*, 537–543.

Edelstein, A.D., Tsuchida, M.A., Amodaj, N., Pinkard, H., Vale, R.D., and Stuurman, N. (2014). Advanced methods of microscope control using µManager software. J. Biol. Methods *1*.

Fanti, Z., Elena Martinez-Perez, M., and De-Miguel, F.F. (2011). Neurongrowth, a software for automatic quantification of neurite and filopodial dynamics from time-lapse sequences of digital images. Dev. Neurobiol. *71*, 870–881.

Fiesel, F.C., Schurr, C., Weber, S.S., and Kahle, P.J. (2011). TDP-43 knockdown impairs neurite outgrowth dependent on its target histone deacetylase 6. Mol. Neurodegener. *6*, 64.

Finkbeiner, S. (2010). Bridging the Valley of Death of therapeutics for neurodegeneration. Nat. Med. *16*, 1227–1232.

Fischer, L.R., Culver, D.G., Tennant, P., Davis, A.A., Wang, M., Castellano-Sanchez, A., Khan, J., Polak, M.A., and Glass, J.D. (2004). Amyotrophic lateral sclerosis is a distal axonopathy: evidence in mice and man. Exp. Neurol. *185*, 232–240.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. *33*, 1–22.

Friedman, J., Hastie, T., Simon, N., and Tibshirani, R. (2015). glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models.

George, E.I., and Mcculloch, R.E. (1997). Approaches for Bayesian variable selection. Stat. Sin. 339–374.

Gould, T.W., Buss, R.R., Vinsant, S., Prevette, D., Sun, W., Knudson, C.M., Milligan, C.E., and Oppenheim, R.W. (2006). Complete Dissociation of Motor Neuron Death from Motor Dysfunction by Bax Deletion in a Mouse Model of ALS. J. Neurosci. *26*, 8774–8786.

Ho, S.-Y., Chao, C.-Y., Huang, H.-L., Chiu, T.-W., Charoenkwan, P., and Hwang, E. (2011). NeurphologyJ: An automatic neuronal morphology quantification method and its application in pharmacological discovery. BMC Bioinformatics *12*, 230.

Huang, Y., Zhou, X., Miao, B., Lipinski, M., Zhang, Y., Li, F., Degterev, A., Yuan, J., Hu, G., and Wong, S.T.C. (2010). A computational framework for studying neuron morphology from in vitro high content neuron-based screening. J. Neurosci. Methods *190*, 299–309.

Jain, S., van Kesteren, R.E., and Heutink, P. (2012). High Content Screening in Neurodegenerative Diseases. J. Vis. Exp. JoVE.

Johnson, B.S., Snead, D., Lee, J.J., McCaffery, J.M., Shorter, J., and Gitler, A.D. (2009). TDP-43 Is Intrinsically Aggregation-prone, and Amyotrophic Lateral Sclerosis-linked Mutations Accelerate Aggregation and Increase Toxicity. J. Biol. Chem. *284*, 20329–20339.

Kaech, S., Huang, C.-F., and Banker, G. (2012). Long-Term Time-Lapse Imaging of Developing Hippocampal Neurons in Culture. Cold Spring Harb. Protoc. *2012*, pdb.prot068239.

Kandaswamy, U., Rotman, Z., Watt, D., Schillebeeckx, I., Cavalli, V., and Klyachko, V.A. (2013). Automated condition-invariable neurite segmentation and synapse classification using textural analysis-based machine-learning algorithms. J. Neurosci. Methods *213*, 84–98.

Kruschke, J.K. (2013). Bayesian estimation supersedes the t test. J. Exp. Psychol. Gen. *142*, 573–603.

Lattante, S., Ciura, S., Rouleau, G.A., and Kabashi, E. (2015). Defining the genetic connection linking amyotrophic lateral sclerosis (ALS) with frontotemporal dementia (FTD). Trends Genet. *31*, 263–273.

Li, W., Brakefield, D., Pan, Y., Hunter, D., Myckatyn, T.M., and Parsadanian, A. (2007). Muscle-derived but not centrally derived transgene GDNF is neuroprotective in G93A-SOD1 mouse model of ALS. Exp. Neurol. *203*, 457–471.

Mackenzie, I.R.A., Bigio, E.H., Ince, P.G., Geser, F., Neumann, M., Cairns, N.J., Kwong, L.K., Forman, M.S., Ravits, J., Stewart, H., et al. (2007). Pathological TDP-43 distinguishes sporadic amyotrophic lateral sclerosis from amyotrophic lateral sclerosis with SOD1 mutations. Ann. Neurol. *61*, 427–434.

Meijering, E. (2012). Cell Segmentation: 50 Years Down the Road [Life Sciences]. IEEE Signal Process. Mag. *29*, 140–145.

Meijering, E., Jacob, M., Sarria, J.-C. f., Steiner, P., Hirling, H., and Unser, M. (2004). Design and validation of a tool for neurite tracing and analysis in fluorescence microscopy images. Cytometry A *58A*, 167–176.

Miller, J., Arrasate, M., Shaby, B.A., Mitra, S., Masliah, E., and Finkbeiner, S. (2010). Quantitative Relationships between Huntingtin Levels, Polyglutamine Length, Inclusion Body Formation, and Neuronal Death Provide Novel Insight into Huntington's Disease Molecular Pathogenesis. J. Neurosci. *30*, 10541–10550.

Miller, R.G., Mitchell, J.D., and Moore, D.H. (2012). Riluzole for amyotrophic lateral sclerosis (ALS)/motor neuron disease (MND). In Cochrane Database of Systematic Reviews, (John Wiley & Sons, Ltd),.

Misiak, D., Posch, S., Lederer, M., Reinke, C., Hüttelmaier, S., and Möller, B. (2014). Extraction of protein profiles from primary neurons using active contour models and wavelets. J. Neurosci. Methods *225*, 1–12.

Neumann, M., Sampathu, D.M., Kwong, L.K., Truax, A.C., Micsenyi, M.C., Chou, T.T., Bruce, J., Schuck, T., Grossman, M., Clark, C.M., et al. (2006). Ubiquitinated TDP-43 in Frontotemporal Lobar Degeneration and Amyotrophic Lateral Sclerosis. Science *314*, 130–133.

Pani, G., De Vos, W.H., Samari, N., de Saint-Georges, L., Baatout, S., Van Oostveldt, P., and Benotmane, M.A. (2014). MorphoNeuroNet: An automated method for dense neurite network analysis. Cytometry A *85*, 188–199.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.

Roselli, F., and Caroni, P. (2015). From Intrinsic Firing Properties to Selective Neuronal Vulnerability in Neurodegenerative Diseases. Neuron *85*, 901–910.

Ross, C.A. (1997). Intranuclear neuronal inclusions: a common pathogenic mechanism for glutamine-repeat neurodegenerative diseases? Neuron *19*, 1147–1150.

Rutherford, N.J., Zhang, Y.-J., Baker, M., Gass, J.M., Finch, N.A., Xu, Y.-F., Stewart, H., Kelley, B.J., Kuntz, K., Crook, R.J.P., et al. (2008). Novel Mutations in TARDBP (TDP-43) in Patients with Familial Amyotrophic Lateral Sclerosis. PLoS Genet *4*, e1000193.

de Santos-Sierra, D., Sendiña-Nadal, I., Leyva, I., Almendral, J.A., Ayali, A., Anava, S., Sánchez-Ávila, C., and Boccaletti, S. (2015). Graph-based unsupervised segmentation algorithm for cultured neuronal networks' structure characterization and modeling. Cytometry A *87*, 513–523.

Saudou, F., Finkbeiner, S., Devys, D., and Greenberg, M.E. (1998). Huntingtin acts in the nucleus to induce apoptosis but death does not correlate with the formation of intranuclear inclusions. Cell *95*, 55–66.

Saxena, S., and Caroni, P. (2011). Selective Neuronal Vulnerability in Neurodegenerative Diseases: from Stressor Thresholds to Degeneration. Neuron *71*, 35–48.

Scott, S.L. (2014). BoomSpikeSlab: MCMC for Spike and Slab Regression.

Sharma, P., Ando, D.M., Daub, A., Kaye, J.A., and Finkbeiner, S. (2012). Chapter seventeen - High-Throughput Screening in Primary Neurons. In Methods in Enzymology, P.M. Conn, ed. (Academic Press), pp. 331–360.

Sisodia, S.S. (1998). Nuclear inclusions in glutamine repeat disorders: are they pernicious, coincidental, or beneficial? Cell *95*, 1–4.

Skibinski, G., and Finkbeiner, S. (2013). Longitudinal measures of proteostasis in live neurons: Features that determine fate in models of neurodegenerative disease. FEBS Lett. *587*, 1139–1146.

Skibinski, G., Nakamura, K., Cookson, M.R., and Finkbeiner, S. (2014). Mutant LRRK2 Toxicity in Neurons Depends on LRRK2 Levels and Synuclein But Not Kinase Activity or Inclusion Bodies. J. Neurosci. *34*, 418–433.

Statistics, L.B., and Breiman, L. (2001). Random Forests. In Machine Learning, pp. 5–32.

Suzuki, M., McHugh, J., Tork, C., Shelley, B., Klein, S.M., Aebischer, P., and Svendsen, C.N. (2007). GDNF Secreting Human Neural Progenitor Cells Protect Dying Motor Neurons, but Not Their Projection to Muscle, in a Rat Model of Familial ALS. PLoS ONE *2*, e689.

Suzuki, M., McHugh, J., Tork, C., Shelley, B., Hayes, A., Bellantuono, I., Aebischer, P., and Svendsen, C.N. (2008). Direct Muscle Delivery of GDNF With Human Mesenchymal Stem Cells Improves Motor Neuron Survival and Function in a Rat Model of Familial ALS. Mol. Ther. *16*, 2002–2010.

Thevenaz, P., Ruttimann, U.E., and Unser, M. (1998). A pyramid approach to subpixel registration based on intensity. IEEE Trans. Image Process. *7*, 27–41.

Tripathi, V.B., Baskaran, P., Shaw, C.E., and Guthrie, S. (2014). Tar DNA-binding protein-43 (TDP-43) regulates axon growth in vitro and in vivo. Neurobiol. Dis. *65*, 25–34.

Tüchler, R. (2008). Bayesian Variable Selection for Logistic Models Using Auxiliary Mixture Sampling. J. Comput. Graph. Stat. - J COMPUT GRAPH STAT *17*, 76–94.

Varma, H., Lo, D.C., and Stockwell, B.R. (2008). High Throughput Screening for Neurodegeneration and Complex Disease Phenotypes. Comb. Chem. High Throughput Screen. *11*, 238–248.

Wang, W., Li, L., Lin, W.-L., Dickson, D.W., Petrucelli, L., Zhang, T., and Wang, X. (2013). The ALS disease-associated mutant TDP-43 impairs mitochondrial dynamics and function in motor neurons. Hum. Mol. Genet. *22*, 4706–4719.
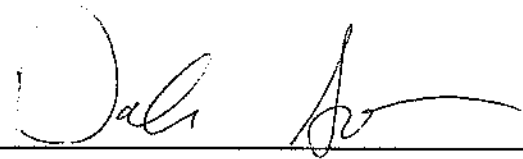
Wu, C., Schulte, J., Sepp, K.J., Littleton, J.T., and Hong, P. (2010). Automatic Robust Neurite Detection and Morphological Analysis of Neuronal Cell Cultures in High-content Screening. Neuroinformatics *8*, 83–100.

Zhang, M., Luo, G., Zhou, Y., Wang, S., and Zhong, Z. (2014). Phenotypic Screens Targeting Neurodegenerative Diseases. J. Biomol. Screen. *19*, 1–16.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature _____ Date 9-3-2015