Virginie Uhlmann, Laurène Donati, and Daniel Sage

A Practical Guide to Supervised Deep Learning for Bioimage Analysis

Challenges and good practices

SHUTTERSTOCK.COM/KKSSR

The variety of bioimage data and their quality have dramatically increased over the last decade. In parallel, the number of proposed deep learning (DL) models for their analysis grows by the day. Yet, the adequate reuse of published tools by practitioners without DL expertise still raises many practical questions. In this article, we explore four categories of challenges faced by researchers when using supervised DL models in bioimaging applications. We provide examples in which each challenge arises and review the consequences that inadequate decisions may have. We then outline good practices that can be implemented to address the challenges of each category in a scientifically sound way. We provide pointers to the resources that are already available or in active development to help in this endeavor and advocate for the development of further community-driven standards. While primarily intended as a practical tutorial for life scientists, this article also aims at fostering discussions among method developers around the formulation of guidelines for the adequate deployment of DL, with the ultimate goal of accelerating the adoption of novel DL technologies in the biology community.

Introduction

The automation of acquisition pipelines and the development of new microscopy technologies that push the resolution limits in both time and space have dramatically increased the amount of bioimage data currently generated. Advanced analysis algorithms now allow one to produce images of ever-better quality and enable a truly quantitative assessment of complex biological structures and their interactions.

Although initially considered as a particular case of image processing and computer vision, bioimage analysis has grown over the past 10 years into an independent field of research populated by its own community of scientists. Bioimage-analysis problems offer unique challenges that call for dedicated scientific approaches and solutions, while still clearly benefiting from research developments in computer vision. Predictably, recent breakthroughs in that field and the rise of DL have inspired the development of powerful automated methods for

Digital Object Identifier 10.1109/MSP.2021.3123589
Date of current version: 24 February 2022

many of the classical problems of bioimage analysis, from classification to segmentation and image restoration.

Landmark developments in bioimaging

In 2015, the deep convolutional neural network (CNN) ResNet reached first place on the ImageNet classification leaderboard [1]. Because of its ability to accommodate hundreds to thousands of neuron layers and still train efficiently, ResNet became a reference model for classification tasks involving natural images. ResNet models are CNN encoders that have an overall size that depends on the size of each of the model layers and the number of layers considered. These models and their variants, such as DenseNet [2], were quickly adopted by the bioimage-analysis community and led to breakthroughs in the automation of classification for challenging microscopy data sets [3].

The U-Net model, presented at the MICCAI the same year as ResNet appeared, experienced an even bigger success [4]. Unlike ResNet, U-Net was tailored from the start to biomedical images and reached an astounding performance on segmentation problems. U-Net is an encoder–decoder CNN composed of a contracting (downsampling) and an expansive (upsampling) path. One of the key strategies of the U-Net model is the use of skip connections that incorporate the multiresolution aspect of visual features in images and integrate spatial information. As a result, a large number of feature maps are available in the decoder path, which allows information to be transferred efficiently. Although many other architectures were proposed in the first years of the DL era, U-Net established itself as the most efficient and versatile backbone for bioimage analysis. Its superior robustness inspired many variants and led to the first U.S. Food and Drug Administration (FDA)-approved DL algorithm for digital pathology [5]. Six years after its introduction, U-Net remains one of the most commonly used CNN architectures in bioimage analysis, with more than 5,000 citations at the start of 2021.

> **"Although many other architectures were proposed in the first years of the DL era, U-net established itself as the most efficient and versatile backbone for bioimage analysis."**

Based on this landmark supervised deep neural network model, excellent solutions were then proposed for classical bioimaging problems, such as deconvolution and denoising [6], single-molecule localization microscopy [7], [8], segmentation [9], and object detection [10], [11]. A key goal of these methods is to be generalist enough to guarantee good performance on a wide range of imaging modalities, such as fluorescence, differential interference contrast, phase-contrast, and bright-field microscopy, among others. All these methods rely on the U-Net architecture, demonstrating once more its robustness to the variety of visual appearances in the data and the diversity of bioimage-analysis problems being addressed.

In [6], for instance, the model is trained on pairs of images acquired at several light intensities and signal-to-noise ratios, and it is able to restore high-quality images from data acquired in low-light conditions. In [9], a U-Net model generates segmentation masks on a wide variety of images based on gradient predictions. This model also allows identifying individual objects in the image as star-convex polygons [10] and parametric spline curves [11]. In addition to these established tools, which work on a wide range of bioimages, numerous application-specific DL-based methods have been and are still being developed. We orient readers interested in an overview of the state of the art to [12].

Deep learning in the reuse era

While research in DL is still progressing at a quick pace, a consensus on neural network architectures has started to emerge for bioimaging applications. As such, the urgency to develop new methodologies is steadily being replaced by the need for bioimage analysts to gain proficiency in the appropriate use of existing DL models. Concretely, bioimage analysis is shifting from a setting in which biologists had to team up with computer scientists to develop new tools, to a paradigm in which life scientists are able to choose the best option for their data from a catalog of available and ready-to-use neural network models.

This situation is reminiscent of the one involving classical image processing algorithms in the early 1990s, for instance, in image segmentation. Back then, the few robust and generally applicable algorithms that stood out in the computer vision community, such as the watershed algorithm and active contours, were adopted by life scientists, adapted, and repurposed for a myriad of specific applications. As this transition now operates within the DL era, practitioners must be instructed on how to use this new technology appropriately, and developers of novel methods need to agree on a set of guidelines for their users.

Because of the obvious timeliness and importance of the topic, many review papers on the use of DL in bioimaging have been published recently [12]–[14]. These works present an overview of the state of the art in the subject, discuss major accomplishments and current limitations, review ongoing directions of research, and outline the next methodological challenges. They do not, however, focus on providing comprehensive practical advice for nonexperts in DL technology. Yet, because the bioimage-analysis community has reached a critical mass of published works that present successful supervised DL solutions, we believe that users will gain most by understanding how to reuse and adapt these existing tools in a scientifically sound manner, instead of trying to reinvent the wheel.

Terminology and key concepts

In supervised DL, the classical workflow first involves the training of a randomly initialized network with a large amount of annotated ground-truth data [Figure 1(a)]. Then, the trained model can be used in inference mode to make predictions for

new data on which it has not been trained and for which no ground-truth labels are available [Figure 1(b)]. The trained model can also be used as basis to obtain a model that is specific to a new related problem [Figure 1(c)].

In this article, we define *pretrained models* as bundles of neural network architectures, trained weights, and suitable hyperparameters. Throughout the manuscript, selecting a pretrained model therefore means selecting such a bundle for direct reuse in inference mode on one's own data. Once again, note that we will not discuss the process of training neural network architectures from scratch and tuning hyperparameters.

Neural networks can be pretrained in a range of different manners. A classical strategy consists of pretraining on large natural image databases, such as ImageNet. As already mentioned, we will, however, focus on the reuse of models that have been initially trained on a bioimage data set to solve a specific bioimage-analysis problem and on the questions this raises.

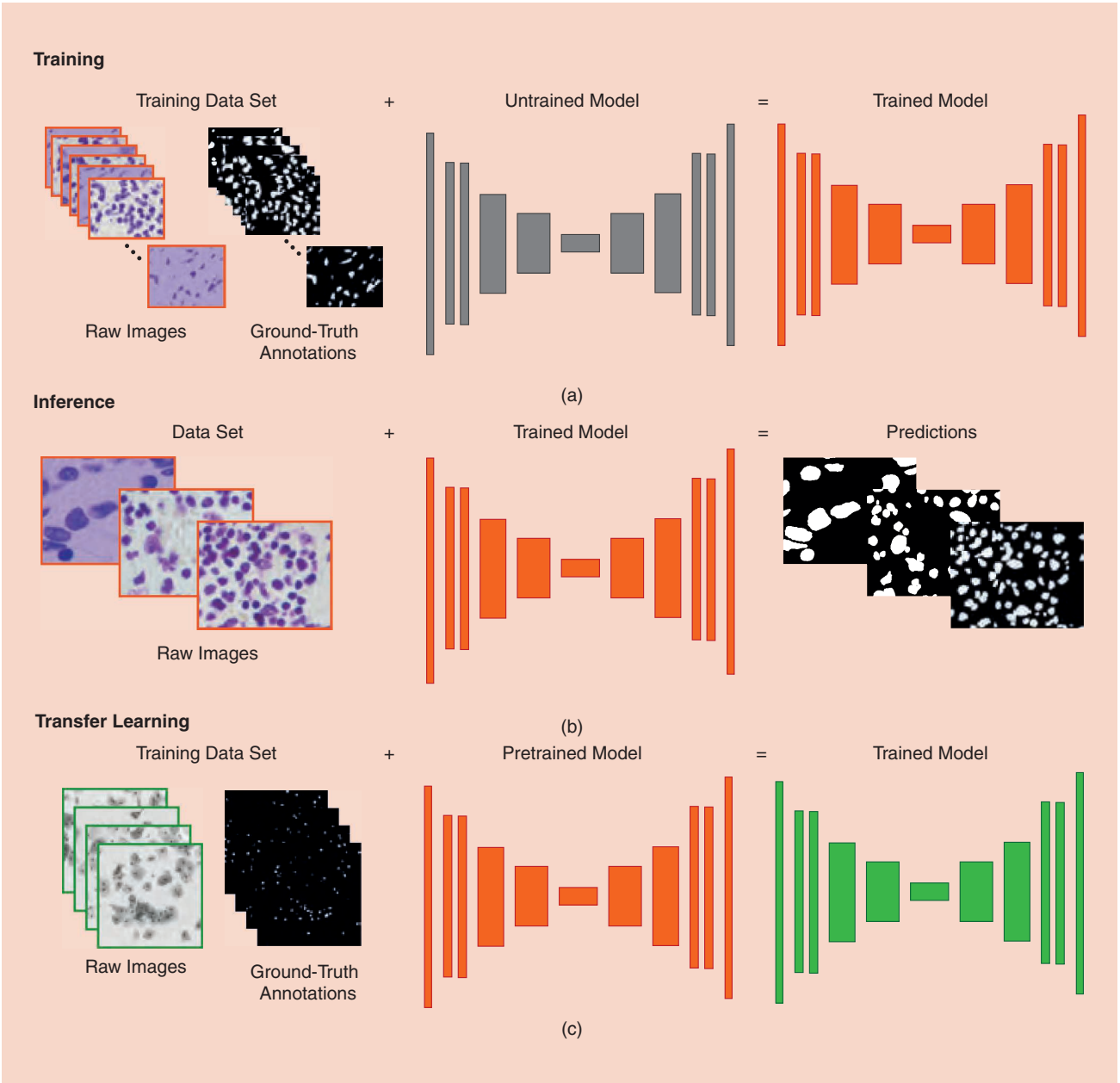


FIGURE 1. A typical supervised DL workflow for nuclei segmentation in microscopy images. (a) During training, the randomly initialized network learns to segment nuclei in a specific type of microscopy data (hematoxylin and eosin stain, outlined in orange) from a large annotated data set. (b) The trained network can then be used in inference mode to segment unseen images of the same type. (c) The weights of the same network can also be fine-tuned for the related task of segmenting nuclei in another type of image data (Hoechst fluorescent stain, depicted in green), relying on a smaller amount of representative annotated samples. (Microscopy images are samples of the BBBC038v1 data set; source: Broad Bioimage Benchmark Collection [15].)

It is important to note that bioimage-analysis tasks may be related in multiple ways. They may, for instance, exploit data acquired with the same microscope or focus on the same type of biological objects. As we shall see, the nature of the relationship between the problem on which a DL model was initially trained and the new task at hand is crucial in understanding how the model will transfer to the new problem.

Large annotated bioimaging data sets are rarely available because of the sheer amount of human resources that manual annotation requires and because the curation is particularly challenging [16]. Hence, strategies that exploit pretrained models to alleviate the need for large amounts of annotations are particularly relevant for biology. In particular, the approaches of transfer learning and fine-tuning, which we introduce later on in the manuscript, allow one to make the most of small, high-quality, curated training sets.

Strategies that exploit pretrained models to alleviate the need for large amounts of annotations are particularly relevant for biology.

Article Scope and organization

This article is targeted to life scientists who are enthusiastic about supervised DL but are not machine learning researchers or engineers by background. In contrast to the previously mentioned published reviews, our article approaches the topic of DL in bioimaging from a practitioner's perspective. We focus on the concrete questions users may ask themselves when attempting to reuse DL models to analyze their data and propose an overview of good practices they can follow. More precisely, our focus is on the supervised learning setting, which consists of situations in which the network is trained from a labeled set of ground-truth examples. There is ongoing research around other learning paradigms, such as semisupervised, unsupervised, and reinforcement learning, but we choose not to cover them here because of their current lower adoption by end users in the life sciences. We thus concentrate on questions that arise from the reuse of available pretrained supervised models for direct inference rather than on the process of model building and training or the design of neural network architectures.

Adopting a tutorial tone, we aim to address both practitioners without DL expertise (to give them concrete tips and pointers toward helpful resources) and method developers (to help them identify how to make their tools more accessible and reusable to life scientists). Eventually, our hope is that this article will help foster further discussions and exchanges between the engineering and biology communities.

Table 1. The key challenges and risks for the everyday user of DL in bioimage analysis.

Challenge	Cause	Example	Risks
Choice of pretrained model	Many different models to choose from for the same task	Which model is best to segment individual cells in bright-field microscopy images?	Inconsistent results
Data set shift	Difference in distributions and/or domains between training and inference	Can I use this model on my data although they are slightly different from the ones used for training?	Performance degradation and hallucination of results
Trust in results	Black-box nature of DL	How much can I trust an accuracy of 98%?	Overconfidence or unwarranted skepticism
Overuse of DL	DL preferred over equally performing classical alternatives	Is DL the best tool for my particular problem?	Unnecessary complication of the analysis pipeline

Challenges and risks for the practitioner

Life scientists who want to reuse pretrained DL models to analyze their own bioimage data sets will stumble upon a number of practical questions: which pretrained model to choose, how to appropriately reuse it (what is the range of applicability of the chosen pretrained model), and, ultimately, what is the validity of the results (can the model's prediction be trusted). By understanding the complications that these different considerations bring, users may also want to reconsider their premise and question whether DL is even really needed at first. To navigate these various aspects, we have divided the practical challenges that practitioners may face on an everyday basis into four main categories, as shown in Table 1: 1) the choice of pretrained models, 2) data set shift, 3) trust in results, and 4) overuse of DL models. Each category is further discussed in its individual section.

Choice of pretrained model

Numerous custom DL approaches has been proposed over recent years to solve similar analysis tasks in bioimaging. Hence, the first challenge a practitioner is likely to face when wanting to reuse an existing pretrained model for his/her own data is to identify the most appropriate one. To start, end users need to identify a series of DL models that are appropriate for their specific applications. After thorough consideration of the available literature, code bases, and expert recommendations, the follow-up is then to choose the one model they should use among the many that could be identified.

This task is not straightforward in practice. In contrast to computer vision and medical imaging (see grand-challenge.org), few problems in bioimaging have well-established benchmark data sets on which performance can be objectively ranked and comparatively assessed. Reconstruction of single-molecule localization superresolution microscopy images (srm.epfl.ch), nuclei segmentation (bbbc.broadinstitute.org/BBBC038), and cell tracking (celltrackingchallenge.net) are notable examples for which benchmarks exist. The bioimaging community is poised to further expand with community-compiled resources. When available, such benchmarks provide valuable indications on the relative performance of the different approaches addressing a similar task as well as insights on their strengths and weaknesses.

Following good scientific practice, most methods are openly available in nonproprietary format on version-control platforms, such as GitHub (github.com) or Zenodo (zenodo.org). Unfortunately, the richness of the documentation and the training information deposited with these models varies dramatically. At one end of the spectrum, users can access exemplar repositories that permit them to fully reproduce published results by providing information that is not necessarily embedded in the code, such as execution environment or external dependencies, in the form of self-contained Docker images (www.docker.com). At the opposite end, users may access minimalist platforms hosting uncommented software with little to no information about its execution, practice data, or user guidance.

Pretrained models that do not provide access to their original training set should be considered incomplete and hence avoided.

The repositories that host well-documented code and provide example pipelines, for instance, in the form of Jupyter notebooks (jupyter.org), enable users to forge their own opinion through quick testing. However, seemingly promising methods that either 1) address problems for which there exist no community-accepted benchmark data sets or 2) lack appropriate documentation to enable code reuse or practice data to reproduce published results pose a particularly difficult challenge to the end user. For example, a lack of details about the initial training strategy and the nature of the data used for training makes it especially difficult to evaluate a possible data set shift (see the section "Data Set Shift"). Another difficulty is that popular models often keep on being developed after publication. Practitioners must then be careful to select the appropriate version of the model, keeping in mind that it may have an impact on the model performance.

The adoption of a well-thought-out approach to navigate the choice of pretrained models is crucial. Indeed, the use of a pretrained model that is not adequately suited to the task at hand may have dire consequences on end results, ranging from poor performance to blatantly erroneous predictions, as we further discuss and exemplify in the section "Trust in Results." While less problematic from a methodological perspective but no less vexing for the end user, different suitable models may generate different, possibly conflicting, predictions. Such inconsistencies can be particularly daunting to untangle when produced by a collection of apparently reliable methods. This again raises questions about trust and interpretability, also discussed in the section "Trust in Results." To complicate things further, the adoption of a seemingly appropriate model—for instance, one that is designed for the same problem and image data—that was initially trained on an insufficient amount of data may also result in a subpar performance.

Data set shift

Once one relevant pretrained model (or a set of pretrained models) has been identified, the next question revolves around the understanding of whether a direct reuse in inference mode is appropriate, which is closely tied to the generalization capability of the network. A key complication in bioimaging is that experiments are rarely standardized. They instead exhibit an extreme variability at all steps, from sample preparation to imaging modality and conditions, scales, and biological phenotypes of interest. Hence, the challenge for the practitioner lies in understanding the extent to which a model designed for a specific task (e.g., the segmentation of cells stained with a membrane marker) can be reused in the context of a seemingly identical but slightly different task (e.g., the segmentation of membrane-stained cells with a different marker). Unfortunately, even when the new data set on which inference is performed is very close to the one used for training, the performance of a DL model may degrade unpredictably, from slightly worse-than-expected results to completely irrelevant ones.

The degradation of the model performance due to discrepancies in the data set used for training and the one used for inference is commonly referred to as *data set shift* [18]. Data set shift is studied extensively in probabilistic modeling and can be characterized in various ways; we refer the interested reader to [19] for a comprehensive overview. In Figure 2, we provide an illustration of the three main categories of data set shifts, namely:

- the covariance shift, which reflects a discrepancy in the distributions of the input features seen at training and inference stages (e.g., if the resolution, sample preparation, or acquisition parameters change between training and inference)
- the prior probability shift, which reflects a discrepancy in the distributions of the classes present at training and inference stages (e.g., if the class balance in the training data does not match that in the inference data)
- the concept shift, which reflects a discrepancy in the relationship between the input features and the network prediction (e.g., if the network trained on images from a specific modality is used to predict the same objects imaged with a different modality).

Some types of data set shift can be more easily identified than others. For instance, a risk of concept shift can be directly identified whenever a pretrained model is used to handle image data that have been acquired in a different way than the ones on which the model was trained. On the other hand, covariate and prior probability shifts will often be more subtle and require a thorough investigation of the properties of the original training set (e.g., in terms of object appearances or class balance). For this reason, methods developers committed to open science should always strive to make the original training set available. On the user side, pretrained models that do not provide access to their original training set should be considered incomplete and hence avoided.

While a good understanding of the types and consequences of data set shifts is undeniably useful, transfer learning and fine-tuning offer concrete mitigation strategies that practitioners can adopt. Subsequent questions revolve around the

amount of data needed for transfer learning or fine-tuning and the required degree of similarity with the data used for training. We introduce these various concepts in more detail in the section "Transfer Learning and Fine-Tuning."

Trust in results

Once a supervised model has been chosen and possible data set shifts accounted for, the next challenge is to assess the reliability of the resulting predictions. The very nature of DL models makes it difficult to determine the role each feature plays in the inference process, a phenomenon colloquially referred to as the "black-box effect." DL methods typically offer fewer theoretical guarantees than classical algorithms, for instance, with regard to convergence to their solution. In some cases, inappropriately used supervised DL models can even hallucinate artificial results, as exemplified in Figure 3: an image restoration model trained on images of microtubules will be strongly biased to produce tube-like structures, even when the input image contains blob-like nuclear pore complexes. Concrete examples of questions that practitioners may ask themselves at this stage thus include to what extent they can trust the predictions of a model obtaining a high accuracy after fine-tuning; whether all automatically inferred results must be scrutinized and assessed manually; or what to do when the output of the model looks qualitatively good, but the mechanisms leading to it are not immediately interpretable.

All things being considered, adequately chosen supervised models that have been fine-tuned appropriately can efficiently automate many bioimage-analysis tasks that are otherwise performed manually. They can also help reveal patterns in data that are too complex for the human eye to perceive. The biggest risks around trust in DL models therefore include both blind trust in results, leading to flawed conclusions, and overskepticism, leading to unnecessarily scrutinized results and possibly disregarded or downplayed findings. Hence, one's ability to identify the factors that may impact the performance of a pretrained model and the understanding of how they can be mitigated are crucial for a right balance between trust and questioning.

Beyond data set shift, performance may also be affected by confounders that are of no direct biological interest. For instance, a classifier may be able to successfully distinguish between two different biological conditions using background image content only, if the same background noise statistics happen to be shared among samples of the same type [21]. Although practitioners may rightfully find it difficult to identify sources of bias in DL predictions without having been trained as a machine learning expert, general principles and accessible resources can be exploited to start on the right track, as discussed in the section "Validation."

In the life sciences community, a model zoo specifically dedicated to pretrained models for bioimage analysis is currently being developed: the Bioimage Model Zoo.

Overuse of deep learning

Considering the popularity and all-purpose nature of DL solutions, a last-but-not-least important aspect to consider is the overuse of DL models in situations where conceptually simpler and computationally lighter traditional image processing or machine learning solutions would perform equally well. Although often arising only after the complications inherent to the use of DL (discussed in the sections "Choice of Pretrained Model," "Data Set Shift," and "Trust in Results") have become apparent, the question of whether DL is

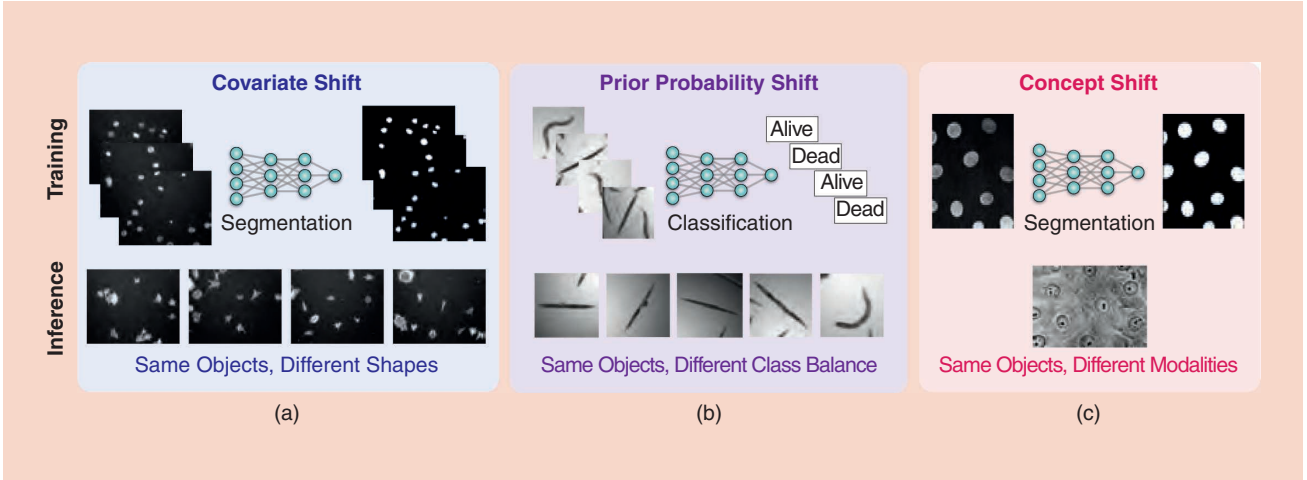
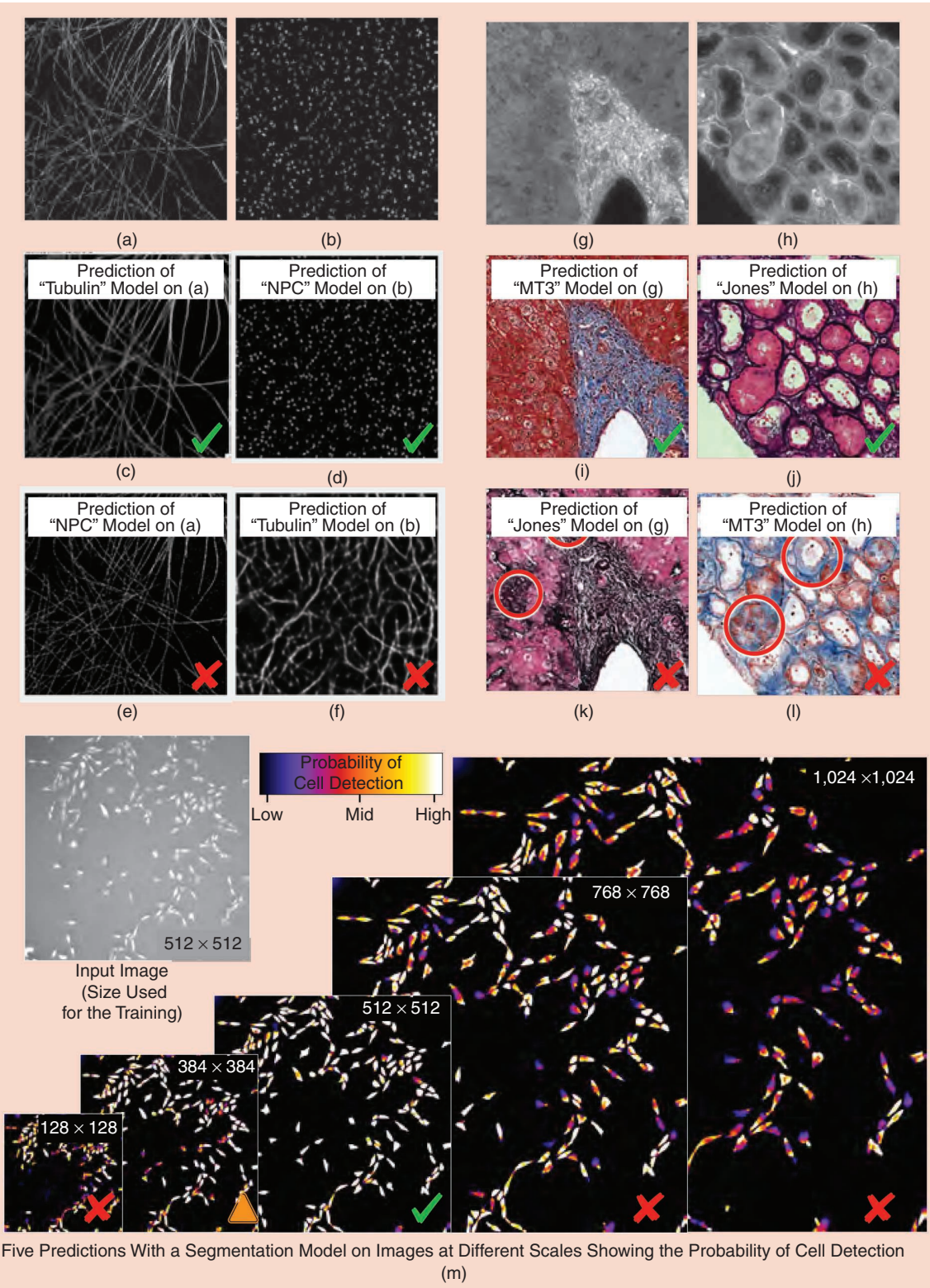


FIGURE 2. Causes of data set shift with illustrative examples. (a) A covariate shift happens when the distribution of features in the data used for training and inference differ. In our example, a network trained on a batch of images featuring exclusively round cells is used to segment another batch where the cells exhibit much more complicated shapes. (b) A prior probability shift is caused by a change in the class distributions between training and inference. In our example, a classifier trained on a balanced two-class problem is used to infer on an imbalanced data set. (c) A concept shift occurs when the relationship between the input features and network output(s) changes between training and inference. In our example, a network trained on fluorescence light microscopy images is used to segment phase-contrast microscopy images of the same sample. (Image sources: [17], BBBC010, bbbc.broadinstitute.org/BBBC010 and Cell Image Library:11831, www.cellimagelibrary.org.)



Five Predictions With a Segmentation Model on Images at Different Scales Showing the Probability of Cell Detection

FIGURE 3. Inference of pretrained models on appropriate and inappropriate input images. (a)–(f) Predictions of two models provided in ANNA-PALM (annapalm.pasteur.fr). (a) The "Tubulin" model was trained on images of a microtubule structure, and (b) the "NPC" model was trained on images of a nuclear pore complex. The application of these models to structures on which they were trained [(c) and (d)] and to different ones [(e) and (f)] reveals the strong bias of the model toward the specific type of structure encountered during training. (g)–(l) Predictions of two trained models, "Masson Trichrome MT3" and "Jones," for virtual staining [20]. The models were used to virtually stain objects seen during training [(i) and (j)] and (k) different objects [(k) and (l)]. Incorrect outputs are highlighted with red circles. (m) Illustration of the dependency on object size in a U-Net-based segmentation model. The input image is a phase-contrast microscopy image of stem pancreatic cells. The output is a probability map that indicates how well the cells are detected at a given position. The quality of the results is strongly influenced by the image size and, more specifically, by how similar this size is to the size of the training data.

actually needed should ideally always come first. For instance, a practitioner desiring to segment nuclei on DAPI-stained confocal fluorescence microscopy images may start annotating their data set to generate ground truth for fine-tuning, without first checking whether simple automated thresholding, watershed, or a random-forest pixel classifier would suffice.

The use of DL when classical image processing solutions (or non-DL machine learning tools) would provide equally good results at little extra cost or effort has direct consequences. It may unnecessarily complicate the analysis pipeline and incur many of the previously mentioned issues that could have been avoided otherwise.

Community efforts to produce large-scale publicly available annotated datasets facilitate the practitioners' choice among the wealth of available solutions.

DL does, however, provide efficient solutions to challenging bioimage-analysis tasks, in part due to its impressive adaptability to the data. When relying on appropriate pretrained models, it also offers massive simplification over many handcrafted pipelines that depend on fine, manual parameter tuning. The identification of whether a considered problem requires DL entirely depends on its complexity, which is difficult to assess for researchers without advanced expertise in image processing or computer vision. In spite of this, assessing whether DL is the most appropriate solution to the problem at hand should be the de facto first step of any bioimage-analysis pipeline.

We summarize in Figure 4 the various challenges we discussed throughout this section. We also indicate at which step they should be asked on the user's path from the initial bioimaging data set to the desired end results.

A guide of good practices

It would be illusive to pretend one can provide a one-size-fits-all solution to any of the challenges flagged previously. However, a wealth of tools and strategies is available to help practitioners navigate the practical questions outlined in the preceding section, "Challenges and Risks for the Practitioner." Here, we enunciate a set of good practices, summarized in a few essential points at the end of each section, that help one address the different categories of challenges. We list existing resources and identify possible future directions of efforts from the community to support the development of usage standards.

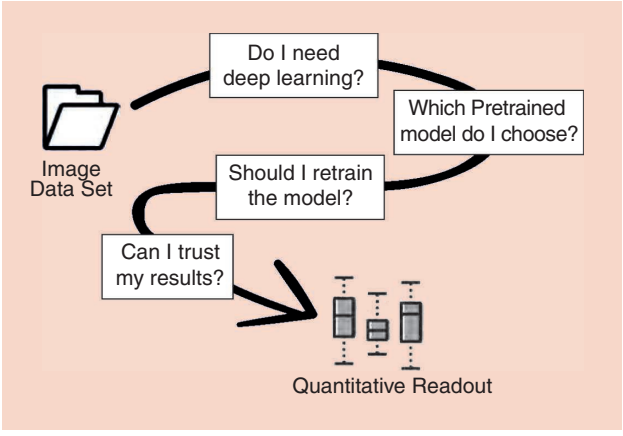


FIGURE 4. A recap of the questions that arise, in their ideal chronological order, when considering the use of supervised DL to solve a bioimage-analysis problem.

Resources to choose appropriate pretrained models

The computer-vision community addressed the challenge of centralizing the distribution of pretrained DL models through the development of model zoos (modelzoo.co). Model zoos are websites that host a collection of curated code and pretrained models for a wide range of platforms and uses. By offering a clearly identifiable entry point to the practitioner, model zoos ensure that the basic required amount of information and documentation will be available to guarantee reproducibility. Typically, model zoos allow users to backtrack the specific nature of the data set and the training strategy.

In the life sciences community, a model zoo specifically dedicated to pretrained models for bioimage analysis is currently being developed: the Bioimage Model Zoo (bioimage.io). Led by a consortium of method developers, this community-driven initiative aims at providing a central repository for published DL models for a large panel of bioimaging applications. The Bioimage Model Zoo is primarily targeted to end users and focuses on model interoperability and easy testing. It hosts fully documented pretrained models that include trained weights, a description of the model architecture, example inputs and outputs, and a configuration specification file to allow at least one of the Bioimage Model Zoo's consumer software types to load and run the model. At the time of writing, the consumer software includes ilastik (www.ilastik.org), ZeroCostDL4Mic (github.com/HenriquesLab/ZeroCostDL4Mic), ImJoy (imjoy.io), ImageJ/Fiji (fiji.sc), and DeepImageJ (deepimagej.github.io/deepimagej). While still in its early days, the Bioimage Model Zoo is evolving quickly and is poised to become a reference resource for the search of pretrained models dedicated to bioimage analysis. It therefore is a good entry point for practitioners in their quest for models that can be tested right away in their favorite software.

Once a set of pretrained models has been identified, end users without programming expertise have several options to compare the performance of the models on their own data. The DeepImageJ plug-in offers a unifying interface to easily exploit pretrained models that solve various types of bioimage-analysis problems in inference mode (available in the Bioimage Model Zoo for instance) through the ImageJ platform. In the same vein, the Fiji plug-in CSBDeep (csbdeep.bioimagecomputing.com) and the ZeroCostDL4Mic toolbox propose to simplify the training and transfer learning steps of most popular models for image segmentation, restoration, and object detection. ZeroCostDL4Mic provides user-friendly Python notebooks that package DL models for bioimage analysis into an entirely human-readable pipeline. While not being a model zoo in itself, ZeroCostDL4Mic offers an enticing platform in

which new models are likely to be quickly incorporated upon their release. Finally, web apps, such as ML-SIM (ml-sim.com) and CDeep3M-Preview (cdeep3m.crbs.ucsd.edu/cdeep3m) also provide an easy way to test a variety of DL models for specific bioimage-analysis tasks, namely reconstruction of structured illumination microscopy images and large-scale segmentation of large electron and light microscopy data sets. For readers specifically interested in bioimage segmentation, the recent review [22] provides an exhaustive list of open source DL software dedicated to that problem.

All of the aforementioned tools are useful to quickly test and explore the suitability of different DL models. To quantitatively assess and compare the performance of candidate models on their data, users can 1) carefully annotate a small set of images that are representative of the rest of the data set (using the tools described in the section "Transfer Learning and Fine-Tuning") and 2) monitor the level of agreement between predictions from each DL model and their high-quality, expert-generated ground-truth annotations. It is worth keeping in mind that small differences in performance may not be statistically meaningful and that choosing the best-performing model may therefore not always be sound. Ideally, the choice of the model should be driven by statistics, relying on multiple comparison tests [23]. The full process of choosing a pretrained model, along with the many steps involved, is illustrated in Figure 5.

"Through the exploration of basic solutions available at a very low entry cost, users can develop a sense of the difficulty of their problem."

In a longer-term perspective, the existence of a model zoo dedicated to bioimage analysis has many positive outlooks for life scientists beyond a simplification of the selection process. Such a platform could facilitate the establishment of a community-driven rating of reliable models, ensure the appropriate versioning of the deposited models, and set guidelines to enforce the proper documentation of the resources made available by the zoo. In parallel, community efforts to produce large-scale publicly available annotated data sets for a wide variety of bioimage-analysis problem, as pioneered by the Broad Bioimage Benchmark Collection (bbbc.broadinstitute.org/image_sets), and benchmarking platforms, such as BIAflows (biaflows.neubias.org), can provide essential information that further facilitates the practitioners' choice among the wealth of available solutions.

The recommended good practice is as follows:

1) Search adequate pretrained models for your task in model zoos or in the recent literature.
2) Pay close attention to the quality of the associated documentation and training information in view of a possible fine-tuning process (see the section "Transfer Learning and Fine-Tuning").
3) Relying on a small, high-quality set of ground-truth annotations, carry out a scout quantitative comparison of the performance of the candidate models on your own data.

Transfer learning and fine-tuning

To assess the risk of data set shift in practice, the starting point is to evaluate the differences between the data set used for training the DL model and the new one on which it is to be deployed (see also the section "Data Set Shift"). If discrepancies between the two problems are identified, the pretrained model should not be directly used for inference but, instead, should be adapted to the new problem at hand through transfer learning and fine-tuning. These two related strategies, sometimes used interchangeably, aim to exploit what a DL model previously learned from a first problem to facilitate learning in a new, similar problem. In both approaches, the use of a small amount of training examples is sufficient as the model is not retrained from scratch, hence bypassing the discouragingly high data needs of DL. In the overwhelming majority of cases, good practice dictates that pretrained models should be adapted before reuse; as a first resource, we orient the reader interested in concrete examples to [24].

Classically, transfer learning in a pretrained DL model is carried out by "freezing" the weights of some of the layers (e.g., the encoder path of a U-Net) and retraining the remaining layers (e.g., the decoder path of a U-Net) with a few ground-truth annotations from the new data to be analyzed. Fine-tuning, in contrast, generally consists in retraining the entire model using the newly generated training set. The recent review [25] discusses these concepts in more technical terms, touching upon the related question of domain adaptation, and is another excellent entry point for readers interested in the topic.

Both transfer learning and fine-tuning require the availability of a small, high-quality training set that reflects the nature of the new problem at hand. To prepare such a data set, a first good rule of thumb is to identify any confounding factor or possible source of batch effect that may affect the statistical analysis of the biological data. Batch effects are changes in data that are due to factors with no link to the biological problem of interest. They can have many sources, from discrepancies in the acquisition protocols across labs (e.g., different microscopy setups) to subtle changes in the acquisition conditions during a single experiment (e.g., changes in the brightness of the microscope lamp). Once such effects have been accounted for, the curated data set can be used to adapt the pretrained model to the new problem by transfer learning and/or fine-tuning, thereby mitigating the risk of data set shift. Note that transfer learning strategies have also been proposed for situations where no ground truth can be produced for the new data set. While this "unsupervised transfer learning" paradigm, technically referred to as *transductive learning*, is beyond the scope of this article, we encourage interested readers to explore the in-depth review [26].

Various open source tools, such as YaPiC (yapic.github.io/yapic), ImJoy, and DeepCell (www.deepcell.org) enable the fine-tuning of a pretrained model given a set of problem-specific annotations. At the time of writing, many other general-purpose

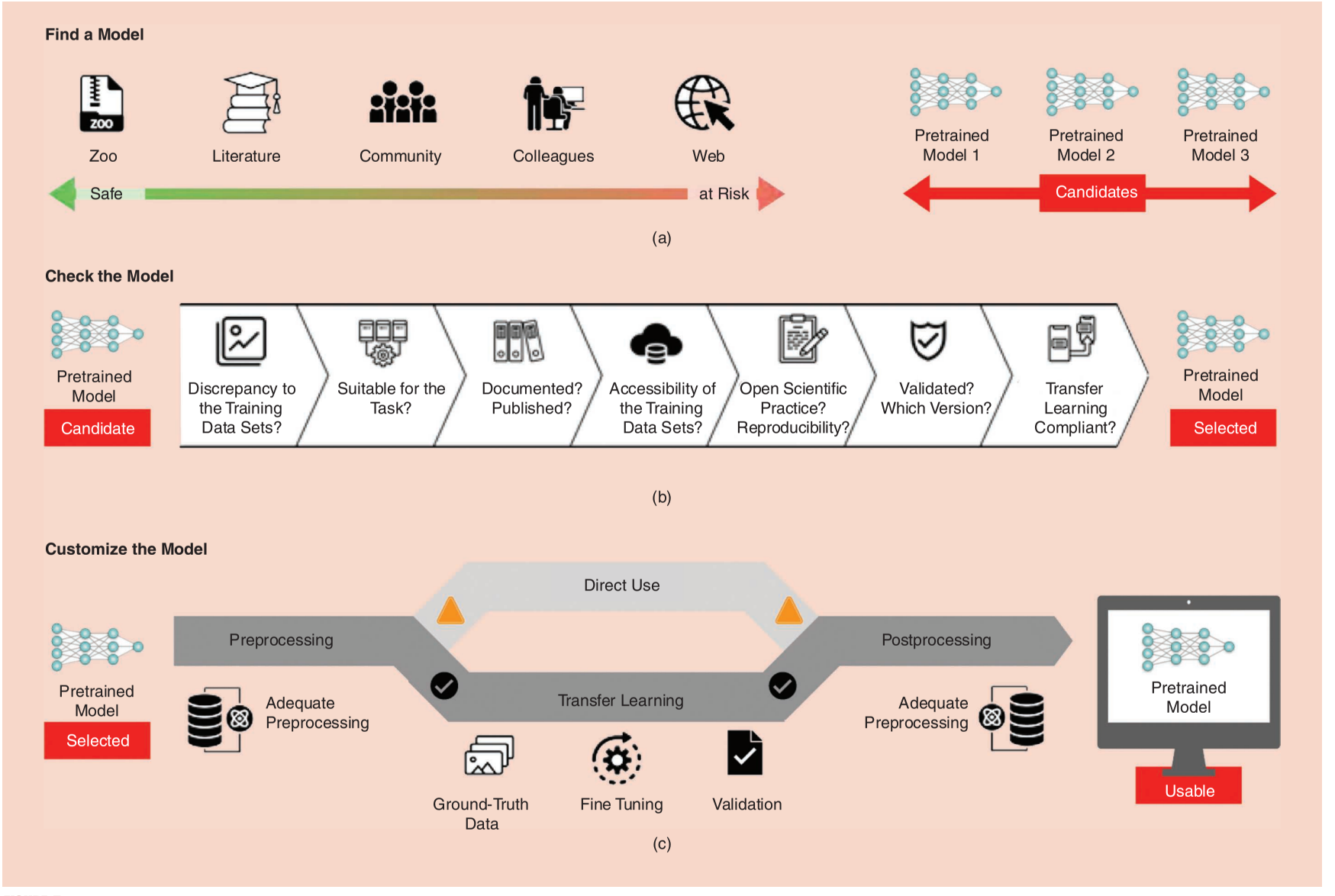


FIGURE 5. Choosing a pretrained model. (a) First, candidate models for the considered task are identified from different sources, the reliability of which may need to be assessed. (b) Each candidate model is then checked for a variety of criteria, facilitating the selection of a subset of pretrained models. (c) Finally, the model is adapted to the problem at hand, ideally through fine-tuning, as described in the section "Transfer Learning and Fine-Tuning." Models going through this entire process can then be safely used to infer on the user's specific data.

image-analysis softwares have announced that this functionality would soon be supported. Alternatively, a handful of all-purpose models trained on a large variety of images, such as CellPose (www.cellpose.org) and NucleAIzer (www.nucleaizer.org), can be reasonably used without fine-tuning. The reason behind this exception lies in the fact that these two models are regularly retrained by their developers on user-submitted data. However, care should be taken to restrict their use to the problem and type of data (in terms of, for instance, imaging modality and objects of interests) these models have encountered during their extensive training.

Resources that facilitate the annotation process are also essential to enable the wider use of transfer learning and fine-tuning. Solutions currently range from basic annotation tools, as found in ImageJ/Fiji and in the napari platform (github.com/napari/napari), to classical machine learning, as provided through the ilastik software. The advent of powerful annotation tools and all-purpose and user-friendly interfaces that make the transfer learning process accessible to practitioners will be key to the democratization of the practice. The combination of these two resources will also bring new insights on the impact that the quality and the amount of problem-specific annotations have on the transfer learning process and, ultimately, on the performance of the DL model.

The recommended good practice is as follows:

1) Identify possible confounding factors in your data set.
2) Design a small, high-quality data set that reflects the nature of your problem. If possible, exploit classical image processing and classical/shallow machine learning tools to facilitate the annotation process.
3) Fine-tune your pretrained model with this curated data set.

Validation

The question of trust in the results produced by DL models is in general multifaceted and goes far beyond the scope of this article. In particular, questions related to the robustness and stability of supervised deep neural networks are being actively investigated in artificial intelligence research. For the end users, reliable ways to investigate and assess the validity of their supervised models include an adequate cross-validation strategy, the monitoring of loss curves, and the reporting of standard metrics. Recent tutorials, such as [27], offer a good introduction to the key technical concepts required for the design of a sound validation procedure. In addition, end users without DL or image-analysis expertise can reach out to experts and collect advice in the online forum forum.image.sc or in user/developer workshops, such as those provided in the I2K conference series (imagej.net/Conference). One avenue to strengthen these practices includes the continuous expansion of the interdisciplinary platform image.sc in its mission to bring together the broader community of computer scientists, software developers, and end users. In addition, many institutions regularly organize internal events, such as hackathons or "consulting hours," to guide users through the application of DL in bioimaging. Dedicated recurring international meetings could also help further consolidate and facilitate interdisciplinary exchanges.

Some members adopted deep learning immediately, while others swiftly rejected its use on the basis of its "black-box" nature and limited explainability.

Several ongoing efforts aim at addressing the problem of explainability of DL models [28]. Explainable DL models, such as xDNN [29], and visual analytic frameworks, such as explAIner [30], aspire to facilitate the in-depth monitoring and investigation of what models are learning to increase their transparency and allow the identification of confounders and possible batch effects. Developed primarily for computer vision, these methods have yet to be adapted to bioimaging problems.

The recommended good practice is as follows:

1) Establish a cross-validation procedure, monitor loss curves, and report standard metrics.
2) Get the predictions of your model reviewed by experts in the field.
3) Ask for feedback on your validation strategy on interdisciplinary online platforms (e.g., image.sc), in user/developer workshops, or in dedicated initiatives within your institution.
4) Publish your code and your results in a fully open manner to support reuse.
5) Rate and give feedback on the pretrained model you used in the Bioimage Model Zoo.

Start simple

While fully acknowledging the many successes of DL in bioimaging, it remains sound practice to only progressively increase complexity when designing a bioimage-analysis pipeline. Concretely, this implies starting from simple, classical algorithms, and not necessarily with the latest DL model. Many GUI-based platforms allow for a quick test of classical image-analysis methods and non-DL machine learning strategies, including ImageJ/Fiji (fiji.sc), Icy (icy.bioimageanalysis.org), ilastik (www.ilastik.org), Weka (www.cs.waikato.ac.nz/ml/weka), CellProfiler (cellprofiler.org), and QuPath (qupath.github.io). These resources either maintain a collection of classical algorithms over the course of their different releases or rely on external contributions in the form of plug-ins. All of them are open source and targeted at users with little to no expertise in image analysis, making them a particularly well-suited entry point for researchers wanting to analyze their own data.

Through the exploration of basic solutions made available at a very low entry cost (e.g., automated thresholding for segmentation), users can develop a sense of the difficulty of their problem and explore the many available solutions in a progressive manner. Such a trial-and-error approach is no waste of time but instead provides crucial hints at the particularities of one's problem at hand. Furthermore, assessing and understanding the quality of the results obtained with classical methods provides clues on the complexity of the considered image-analysis problem, which can help users be on the watch when moving on to supervised DL models for which the precise formulation of the desired outcome is crucial to generate ground-truth data.

The benefit of this "starting simple" approach is thus to opt for a DL-based approach only when the right conditions are met: 1) the challenges of the considered problem have been well understood, 2) the shortcomings of classical approaches have been clearly identified, 3) an appropriate pretrained model has been isolated, and 4) a robust validation strategy has been devised. As such, it helps one to evaluate where it is most appropriate to invest efforts for solving the problem at hand, either in the tuning of classical algorithms or in the generation of high-quality annotated data sets.

The community of bioimage analysts is providing invaluable training, documentation, and assistance with classical image-analysis tools through initiatives such as the Network of European Bioimage Analysts (eubias.org/NEUBIAS) and the Center for Open Bioimage Analysis (openbioimageanalysis.org/training). The well-established trend toward the development of user-friendly solutions for bioimage analysis further consolidates this positive perspective, giving life scientists precious tools to develop interdisciplinary technical expertise. We visually recapitulate the main resources introduced in this section in Figure 6. Additionally, we provide the full list of every tool and method mentioned in the article in our supplementary downloadable material (available at https://doi.org/10.1109/MSP.2021.3123589) along with their companion references and websites when applicable.

The recommended good practice is as follows:

1) Start by exploring classical image analysis and non-DL machine learning solutions that have been extensively validated.
2) When classical approaches do not yield satisfying results, identify possible causes (e.g., complexity of the data, sensitivity to the algorithm's input parameters) and keep them in mind when moving on to more sophisticated DL methods.

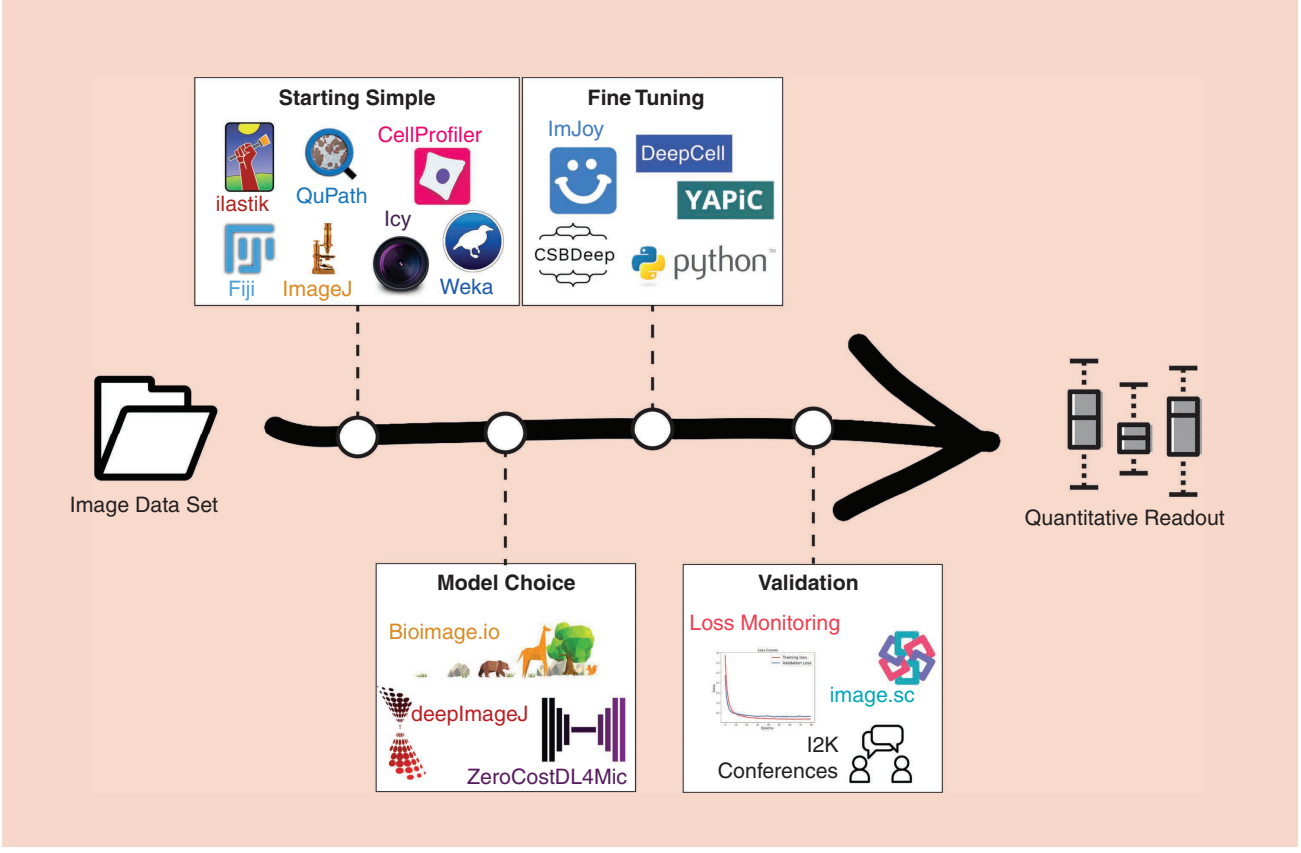


FIGURE 6. A recap of the main resources for each step of a bioimage-analysis path with pretrained supervised DL models, grouped according to the best practice they help support. References and links to the tools listed on the figure are provided in the text. The selection appearing on the figure is not meant to be representative of all existing resources but, instead, lists the most general and widely used ones. (Source of the Python logo: Python Foundation.)

Open questions and outlook

Many exciting resources are currently being developed by the bioimaging community to facilitate the adoption and reuse of supervised DL models in a scientifically rigorous manner. However, there remain several facets of the DL machinery that can strongly impact the everyday use of pretrained models and yet cannot be easily explored by end users at the time of writing this article. For instance, the development of a user-friendly validation and explainability platform would without a doubt significantly contribute to strengthen the trust in DL-based approaches by allowing end users to safely exploit them at their full potential.

A question that is being actively investigated in DL research is that of the robustness of the models, particularly in the context of adversarial attacks. Adversarial attacks are visually imperceptible but precisely structured perturbations in the input image that radically throw off the performance of a model. They ensue from the very high dimensionality of the spaces on which DL models are operating, of which only a very small fraction is "seen" during training. In such high-dimensional spaces, even a minor perturbation can suffice to let the network reach a part of the space it has never seen before and produce unexpected outputs. In classification tasks, for instance, changes as small as the removal of a small patch from the input image have been shown to lead to entirely wrong predictions that nevertheless yield extremely high confidence scores.

The characterization of the extent to which a network can be subject to adversarial attacks and the type of input perturbations that produce such effects is crucial for bioimaging tasks. Even though a DL model designed to analyze microscopy data is admittedly less at risk of being voluntarily attacked by human-designed sets of perturbed inputs, it is nevertheless very likely to encounter unexpected real-world perturbations in the data it will process. To defend DL models against such "natural" adversarial attacks, one strategy consists in the inclusion of all of the relevant perturbations in the training set. Unfortunately, the identification of all possible sources of variability is extremely challenging, if not impossible. Hence, the development of user-friendly frameworks that allow for the evaluation of the robustness of models and facilitate the design of defense strategies is an urgent need for the bioimaging community.

On another front, an exciting perspective of modern biology is the possibility to combine the information from visual (image-based) and omics data. The incorporation of these many quantitative readouts through multimodal DL models is attracting a lot of attention. As research progresses in this direction, new reference architectures and training strategies for the processing of multimodal data sets are likely to emerge. Yet, the use of these advanced DL models will come with challenges similar to those discussed in the section "Challenges and Risks for the Practitioner," possibly aggravated by the deeper complexity of the problem. The establishment of good practices for the safe reuse of pretrained models may thus have an even-stronger importance in the coming multimodal era.

Time and again, the bioimage-analysis community has demonstrated its extraordinary ability to rapidly develop tools that are tailored to the needs and specificity of its research field. We believe that the trend is bound to continue and that reliable open source resources helping users to navigate the challenges they encounter will emerge in the coming years. Equipped with these user-friendly platforms and a clear set of community-driven best practices, we are confident that practitioners will be well armed to exploit the full potential of DL methods in a safe way.

Concluding remarks

The past decade has seen the development of numerous supervised DL models for tasks in bioimage analysis. Following open-science principles, many of these pretrained models are freely available for reuse. To fully exploit their potential in a scientifically sound manner, efforts are required at the practitioner level (to use pretrained models in an informed way, understand their limitations, and interpret results appropriately) and at the community level (to share experiences, create a culture of best practices, and increase confidence in DL-based predictions). Throughout this manuscript, we have outlined several questions and challenges that practitioners are likely to encounter in their use of DL models. We have reviewed concrete examples of strategies and provided selected pointers to open source resources that are generic enough to be relevant in many different types of bioimage-analysis problems, from image restoration to classification and segmentation.

In its early days, DL has been polarizing within the bioimaging community. Some members adopted it immediately, while others swiftly rejected its use on the basis of its "black-box" nature and limited explainability. By proposing a concrete set of good practices that generally apply to the use of DL in a range of bioimage-analysis problems, our hope is twofold: to reassure skeptics and provide them with a strategy that minimizes the risks when experimenting with DL, and to equip long-time DL enthusiasts with additional safeguards on their exploratory journey.

Acknowledgments

We thank Michael Unser and Johannes Hugger for their helpful comments and discussions on the manuscript. Virginie Uhlmann is supported by European Bioinformatics Institute core fundings. This article has supplementary downloadable material available at https://doi.org/10.1109/MSP.2021.3123589, provided by the authors.

Authors

Virginie Uhlmann (uhlmann@ebi.ac.uk) received her Ph.D. diploma in electrical engineering, which was awarded the 2020 Asea Brown Boveri award, from the École polytechnique fédérale de Lausanne, Switzerland. She is leading a research group at the European Bioinformatics Institute,

Cambridge, CB10 1SD, U.K. Her main research interest is quantitative bioimage analysis with a blend of mathematical models and machine learning approaches, focusing on computational geometry and statistical shape analysis. She is an associate editor of the Public Library of Science's PLOS Computational Biology and Cambridge University Press's Biological Imaging, a member of the IEEE Signal Processing Society's Bio Imaging and Signal Processing Technical Committee, and a Member of IEEE.

Laurène Donati (laurene.donati@epfl.ch) received her Ph.D. degree in electrical engineering from the École polytechnique fédérale de Lausanne (EPFL). She is the executive director of the EPFL Center for Imaging (imaging.epfl.ch), Lausanne, 1015, Switzerland, which aims to promote interdisciplinary expertise and cutting-edge research in imaging at EPFL. Her thesis focused on the development of new reconstruction methods for cryoelectron microscopy, including novel unsupervised learning-based techniques for single-particle analysis.

Daniel Sage (daniel.sage@epfl.ch) received his Ph.D. degree in signal and image processing from the Institut National Polytechnique de Grenoble, France. He was a consulting engineer at a private company developing industrial vision systems oriented to quality control before joining the Biomedical Imaging Group at École polytechnique fédérale de Lausanne, Lausanne, 1015, Switzerland, in 1998 as the head of software development. He is involved in numerous research projects in computational bioimaging, including super-resolution microscopy, tracking, deconvolution, and image quantification. He is also involved in open source software development for the life sciences community, using both engineering and machine learning methods.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[2] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 4700–4708.

[3] W. Ouyang *et al*., "Analysis of the human protein atlas image classification competition," *Nature Methods*, vol. 16, no. 12, pp. 1254–1261, 2019, doi: 10.1038/s41592-019-0658-6.

[4] T. Falk *et al*., "U-Net: Deep learning for cell counting, detection, and morphometry," *Nature Methods*, vol. 16, no. 1, pp. 67–70, 2019, doi: 10.1038/s41592-018-0261-2.

[5] A. Echle, N. T. Rindtorff, T. J. Brinker, T. Luedde, A. T. Pearson, and J. N. Kather, "Deep learning in cancer pathology: A new generation of clinical biomarkers," *Brit. J. Cancer*, vol. 124, no. 4, pp. 686–696, 2020, doi: 10.1038/s41416-020-01122-x.

[6] M. Weigert *et al*., "Content-aware image restoration: Pushing the limits of fluorescence microscopy," *Nature Methods*, vol. 15, no. 12, pp. 1090–1097, 2018, doi: 10.1038/s41592-018-0216-7.

[7] A. Speiser *et al*., "Deep learning enables fast and dense single-molecule localization with high accuracy," *Nature Methods*, vol. 18, pp. 1082–1090, Sep. 2021, doi: 10.1038/s41592-021-01236-x.

[8] E. Nehme, L. E. Weiss, T. Michaeli, and Y. Shechtman, "Deep-STORM: Super-resolution single-molecule microscopy by deep learning," *Optica*, vol. 5, no. 4, pp. 458–464, 2018, doi: 10.1364/OPTICA.5.000458.

[9] C. Stringer, T. Wang, M. Michaelos, and M. Pachitariu, "Cellpose: A generalist algorithm for cellular segmentation," *Nature Methods*, vol. 18, no. 1, pp. 100–106, 2021, doi: 10.1038/s41592-020-01018-x.

[10] U. Schmidt, M. Weigert, C. Broaddus, and G. Myers, "Cell detection with star-convex polygons," in *Proc. Med. Image Comput. Comput. Assisted Intervention Conf. (MICCAI'18)*, Granada, Spain, Sep. 16–20, 2018, pp. 265–273, doi: 10.1007/978-3-030-00934-2_30.

[11] S. Mandal and V. Uhlmann, "SplineDist: Automated cell segmentation with spline curves," in *Proc. IEEE Int. Symp. Biomed. Imag. (ISBI'21)*, Nice, France, Apr. 13–16, 2021, pp. 1082–1086, doi: 10.1109/ISBI48211.2021.9433928.

[12] E. Meijering, "A bird's-eye view of deep learning in bioimage analysis," *Comput. Structural Biotechnol. J.*, vol. 18, pp. 2312–2325, Aug. 2020, doi: 10.1016/j.csbj.2020.08.003.

[13] D. P. Hoffman, "The promise and peril of deep learning in microscopy," *Nature Methods*, vol. 18, no. 2, pp. 131–132, 2021, doi: 10.1038/s41592-020-01035-w.

[14] E. Moen, D. Bannon, T. Kudo, W. Graf, M. Covert, and D. Van Valen, "Deep learning for cellular image analysis," *Nature Methods*, vol. 16, no. 12, pp. 1233–1246, 2019, doi: 10.1038/s41592-019-0403-1.

[15] J. C. Caicedo *et al*., "Nucleus segmentation across imaging experiments: The 2018 data science bowl," *Nature Methods*, vol. 16, no. 12, pp. 1247–1253, 2019, doi: 10.1038/s41592-019-0612-7.

[16] J. Ellenberg, J. R. Swedlow, M. Barlow, C. E. Cook, U. Sarkans, A. Patwardhan, A. Brazma, and E. Birney, "A call for public archives for biological image data," *Nature Methods*, vol. 15, no. 11, pp. 849–854, 2018, doi: 10.1038/s41592-018-0195-8.

[17] J. M. Phillip, K.-S. Han, W.-C. Chen, D. Wirtz, and P.-H. Wu, "A robust unsupervised machine-learning method to quantify the morphological heterogeneity of cells and nuclei," *Nature Protocols*, vol. 16, no. 2, pp. 754–774, 2021, doi: 10.1038/s41596-020-00432-x.

[18] J. Quiñonero-Candela, M. Sugiyama, N. D. Lawrence, and A. Schwaighofer, *Dataset Shift in Machine Learning*. Cambridge, MA, USA: MIT Press, 2009.

[19] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognit.*, vol. 45, no. 1, pp. 521–530, 2012, doi: 10.1016/j.patcog.2011.06.019.

[20] Y. Rivenson *et al*., "Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning," *Nature Biomed. Eng.*, vol. 3, no. 6, pp. 466–477, 2019, doi: 10.1038/s41551-019-0362-y.

[21] L. Shamir, "Assessing the efficacy of low-level image content descriptors for computer-based fluorescence microscopy image analysis," *J. Microsc.*, vol. 243, no. 3, pp. 284–292, 2011, doi: 10.1111/j.1365-2818.2011.03502.x.

[22] A. M. Lucas, P. V. Ryder, B. Li, B. A. Cimini, K. W. Eliceiri, and A. E. Carpenter, "Open-source deep-learning software for bioimage segmentation," *Mol. Biol. Cell*, vol. 32, no. 9, pp. 823–829, 2021, doi: 10.1091/mbc.E20-10-0660.

[23] S. Midway, M. Robertson, S. Flinn, and M. Kaller, "Comparing multiple comparisons: Practical guidance for choosing the best multiple comparisons test," *PeerJ*, vol. 8, p. e10387, Dec. 2020, doi: 10.7717/peerj.10387.

[24] A. Kensert, P. J. Harrison, and O. Spjuth, "Transfer learning with deep convolutional neural networks for classifying cellular morphological changes," *SLAS Discovery*, vol. 24, no. 4, pp. 466–475, 2019, doi: 10.1177/2472555218818756.

[25] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, 2020, doi: 10.1109/JPROC.2020.3004555.

[26] W. M. Kouw and M. Loog, "A review of domain adaptation without target labels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 766–785, 2019, doi: 10.1109/TPAMI.2019.2945942.

[27] L. L. Vercio *et al*., "Supervised machine learning tools: A tutorial for clinicians," *J. Neural Eng.*, vol. 17, no. 6, p. 062001, 2020, doi: 10.1088/1741-2552/abbff2.

[28] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau, "Visual analytics in deep learning: An interrogative survey for the next frontiers," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 8, pp. 2674–2693, 2018, doi: 10.1109/TVCG.2018.2843369.

[29] P. Angelov and E. Soares, "Towards explainable deep neural networks (xDNN)," *Neural Netw.*, vol. 130, pp. 185–194, Oct. 2020, doi: 10.1016/j.neunet.2020.07.010.

[30] T. Spinner, U. Schlegel, H. Schäfer, and M. El-Assady, "explAIner: A visual analytics framework for interactive and explainable machine learning," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 1, pp. 1064–1074, 2019, doi: 10.1109/TVCG.2019.2934629.

SP