https://doi.org/10.1038/s41592-024-02322-6

Visual interpretability of bioimaging deep learning models

Oded Rotem & Assaf Zaritsky

The success of deep learning in analyzing bioimages comes at the expense of biologically meaningful interpretations. We review the state of the art of explainable artificial intelligence (XAI) in bioimaging and discuss its potential in hypothesis generation and data-driven discovery.

What is interpretable machine learning?

Deep learning (DL) has become the workhorse underlying the vast majority of bioimage analysis applications. Open source code and new platforms that support usability make DL a practical tool that is accessible to the broad cell biology community. Whereas traditional machine learning depends on experts defining predetermined features to optimize a specific task, the DL data-driven approach benefits from simultaneously optimizing feature extraction and the model to solve the downstream task. Given sufficient data, DL excels at nonlinear integration of image regions, enabling it to identify complex spatial patterns in the raw bioimage data. As a result, DL models surpass traditional bioimage analysis methods and in some cases even exceed the human gold standard. However, the success of DL comes with the discomfort of relying on 'black box' models with huge parametric spaces comprising millions of parameters that introduce major challenges in human understanding of the models' inner workings and in following the models' decision process. In the broader field of machine learning, these challenges triggered the initiation of the vibrant nascent community of explainable artificial intelligence (abbreviated XAI).

The terms 'interpretability' and 'explainability' are used interchangeably in most of the XAI literature. We would like to highlight our perception of the nuances of difference between these terms. Both terms involve the active participation of a human expert in resolving the model's decision process by examining the XAI outputs. Explainability is the explicit description of the cause for a decision in a manner that a human can understand. Interpretability is turning an explanation to domain- and context-specific insight. In the context of bioimaging, this means 'translating' an image-pattern explanation to a biologically meaningful interpretation. For example, realizing that a model pays attention to a specific image region in a cell (that is, an explanation) is not necessarily sufficient to derive specific hypotheses regarding which organelles play a role in this decision (that is, an interpretation). As another example, one explanation of a model that was trained to predict melanoma metastatic efficiency from label-free cell imaging was increased light scattering within single melanoma cells[1]. Although we do not know what causes this change in light scattering, we can turn this explanation into a specific hypothesis via interpretation as a potential change in altered intracellular organelle organization, such as phase-separated droplets or lysosomes, thus setting the stage for follow-up studies to test these possibilities.

Under these definitions, interpretation is always explainable, but the converse is not necessarily true. Interpretability is especially important in sciences and medicine, where 'black box' predictions are not sufficient to decipher the fundamental 'mechanisms' underlying them. In this Comment, we discuss the motivation, current state and future potential of XAI and especially visual interpretability in the realm of bioimaging, and reflect on our experiences in this domain (Fig. 1).

Why do we need visual interpretations of deep learning models?

Not every application of DL requires visual interpretability. For example, it is perhaps acceptable to trust the 'black box' for image analysis tasks such as nucleus segmentation. In many other cases, however, there are convincing motivations for aiming to interpret the machine's predictions. The first motivation is trust. Lack of trust is one of the major obstacles to deploying (high-performing) DL models in biomedical applications. Understanding how a model reaches its decisions provides transparency that can help avoid spurious biases in the image data and in the model's decision process, and can provide clues in cases of erroneous predictions.

Example of such erroneous interpretations include classifying an image as 'wolf' rather than 'husky' due to snow in the image, biased classification of dermatology images due to background skin texture and color balance[2] and technical staining artifacts in parasite-infected red blood cells guiding the model's decision[3]. In these instances, the model is biased toward 'easy' explanations that correlated with the true label, but this 'laziness' (or overfitting), termed 'shortcut learning', causes the model to miss more complex, biologically relevant discriminate content encapsulated in the images. Such understanding can be used to improve the trustworthiness and usability of a model by understanding when models make mistakes and developing methods to mitigate these situations.

The second motivation for applying XAI in bioimaging is the possibility of using the XAI output to measure phenotypes that have been reported based on subjective observations, but that are difficult to measure without the explanation. Examples include using the XAI outcome to define a quantitative measure for a morphological component

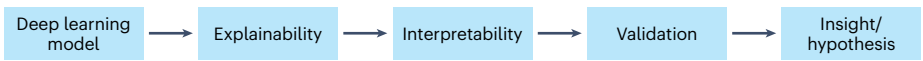

Fig. 1 | The visual interpretability process. New biological insights and/or specific hypotheses are derived from interpreting DL models explanations.

of in vitro fertilization (IVF) embryos called the blastocoel[4] and associating protein localization patterns[5] or their perturbations[6] with specific organelles. The third motivation, and in our opinion the most exciting, is the potential to extract subtle phenotypes that are invisible to the human eye and that cannot be reliably measured with standard image analysis. For example, interpreting live label-free imaged melanoma cells' enhanced protrusive activity as metastasis-driving features[1] or identifying loss of hemoglobin and crenated shapes of blood cells under parasite infection[3].

This discovery-driven interpretability is especially exciting in the biomedical domains, where a new explanation for a machine prediction can lead to a specific hypothesis that can be tested experimentally, closing the loop to establish a causal link. Is it realistic to derive new mechanistic hypotheses and draw biological conclusions from visual interpretation of DL models? These are the early days of XAI in bioimaging, and only time will tell. Developing methods for robust and user-friendly visual interpretability will enable effective exploration of potential biologically meaningful explanations and is a necessary step toward answering this fundamental question.

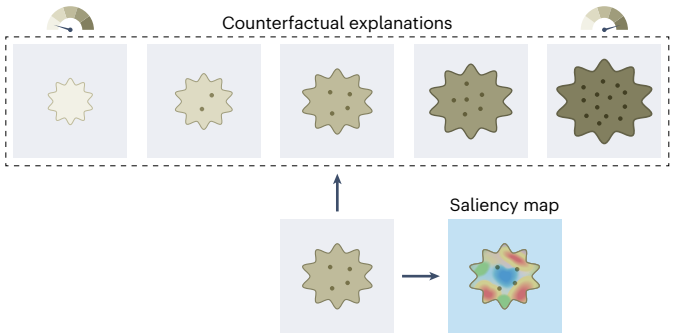

Fig. 2 | Saliency map-based (bottom right) versus counterfactual-based (top) explanations. Saliency methods highlight regions of more importance to the DL model's decision. Counterfactual methods generate artificial realistic images with exaggerated classification-driving phenotypes thus altering the model's prediction. Exaggerated phenotypes in this example are color, size and number of dark spots.

How do we visually interpret deep learning models?

Interpreting image-based DL models is much more difficult than interpreting tabular-based models because of the challenge of moving from pixels to semantic entities. In the field of computer vision, most XAI papers stop at the point of identifying trivial explanations such as open mouths or pointed ears for discriminating between dogs versus cats. In bioimaging, it is harder to interpret the semantic image properties that drive models' decisions because of difficulties in deciphering the less intuitive biological meanings. The latter requires extensive and systematic confirmation of these non-trivial interpretations. We believe that computational biologists developing XAI methods have a competitive advantage because of the inherent expectation in the field of uncovering the 'black box' and provide a plausible biological interpretation, which is a critical step toward generating new, specific hypotheses and designing experiments to test them.

Model interpretation can be in the context of a (local) instance—and/or of a (global) model. Local interpretability characterizes the key decisions for individual predictions, and global interpretability provides a holistic understanding of a model's reasoning processes. Global interpretability is useful toward understanding general rules at the cost of averaging out the full spectrum of heterogeneity of the local instance interpretations. For example, when analyzing a model for IVF embryo morphology quality prediction, global interpretation identified the embryo size, the trophectoderm (a ring of cells surrounding the embryo) and the blastocoel (a fluid-filled cavity inside the embryo) as the top classification-driving morphological features[4]. Local interpretability categorized each embryo according to the different morphological features that dominated the specific classifier's decision.

Some interpretability methods can provide both local and global interpretability. For example, the popular tabular-based SHapley Additive exPlanations (SHAP)[7] calculates the contribution of each (interpretable) feature to an individual prediction. These local interpretations can be aggregated across all individual predictions according to specific criteria (for example, a classification label) to provide a global interpretation of the model. A straightforward example of global interpretability is the exclusion ('ablation') of a feature or a group of features and the ranking of these features according to the corresponding degradation in the model's performance. Such ablation has been applied to bioimaging data to quantify the influence of the local cell density on the prediction of cell fate (apoptosis or mitosis)[8] and the influence of neighboring cell fates (delamination, division or no behavior) on the prediction of cell behavior[9]. A related approach involves systematic assessment of model performance for varying sizes of the classified image crops to probe the biologically relevant spatial scales necessary for a prediction[10].

The most common methods for visual interpretability can be broadly classified as saliency map based or counterfactual based (Fig. 2). Saliency-based (also called 'feature attribution') explanations generate 'attention maps' of the image regions that contribute the most to an individual prediction by assigning to each pixel the aggregated network's inner layers' activations or gradients. Saliency methods are suited for local (instance) interpretation when the visual explanations are localized in the image and do not require further training: for example, in correlating attention maps with their corresponding subcellular structures[11] or using attention maps derived from cell trajectory data to interpret the relative influence of neighboring cells on the motion of a given cell[12]. Counterfactual-based explanations use generative models to artificially change the image in ways that maintain a realistic image and alter the model's prediction, and they excel at understanding subtle image differences in domains that are less intuitive to humans, such as bioimaging; these will therefore will be our focus for the rest of this Comment.

Counterfactual explanations require training of generative model to optimize a lower-dimensional latent representation space that can be used to generate images of the entity of interest. These latent representations can be traversed to generate visual counterfactual explanations along specific directions in the latent space that translate to phenotypic alterations in the image space. Such latent space traversals can exaggerate classification-driving phenotypes beyond the observed data distribution while keeping the rest of the image fixed, together enabling the interpretation of subtle phenotypes. Counterfactual-based explanations are also more suited than saliency-based methods to identifying non-localized visual explanations such as shape or color. In bioimaging, this counterfactual approach has shown promise for interpreting subtle cellular phenotypes by synthetically generating image sequences that follow cell-state transitions[13], disease progression[1], cell fate decisions[8] and perturbations[3]. The approach was also able to provide distinct interpretations for different models that were

trained for the same medical task of classifying melanoma from dermoscopic images of the skin[2].

One approach to interpreting latent representations without or with minimal human intervention is correlating the latent representations with phenotypes directly measurable from the images: for example, correlating latent representations of proteins to their organelles' localization[5], correlating latent representations of cells to their local density[8] and correlating latent representations of IVF embryos to their size[4]. Of course, this interpretability approach relies on existing measurements and cannot be used to discover explanations that could not be measured in advance. We recently proposed a two-step interpretability method that used counterfactual explanations to assign semantic image properties to latent features, followed by SHAP ranking of these interpreted features to interpret individual predictions[4]. This approach combined global and local interpretability to determine the cause of classification (and misclassification) of the morphological quality of IVF embryo instances. Another recent and creative approach used counterfactual explanations to generate attention maps of the most discriminative features and measured their contribution to the prediction[14]. This was applied to reveal morphological differences between different *Drosophila melanogaster* synapses in electron microscopy images.

Outlook

The ever-increasing volume and complexity of bioimage data is making DL a crucial component of modern cell biology. We strongly believe that advancing XAI for bioimaging will propel cell biology forward by improving the trustworthiness of DL models and enabling the design of interpretability 'discovery machines' that can reveal new mechanistic understanding exceeding human intuition. We identify the following themes as having high potential to advance us in this promising direction of XAI visual-driven hypothesis generation.

Visual interpretability of image-to-image transformations. The current focus of visual interpretability is in the realm of classification models, neglecting the interpretability of DL-based image-to-image transformations that are the backbone of many bioimaging applications, ranging from image restoration to denoising, segmentation and cross-modality transformations. Current methods for the explainability of biomedical image-to-image transformations mostly rely on pixel-based uncertainty estimations and can be applied to learn what information in the input image was used for the output prediction. Interpretability of image-to-image transformations could have practical implications in identifying when a model does not perform well. For example, technical (for example, batch effects, uneven illumination) or biological (for example, rare cell states, perturbations) out-of-distribution data can lead to errors and hallucinations and should be considered during downstream analysis. We call for the development of methods for the visual interpretation of bioimage image-to-image transformations that go beyond the pixel scale to the scale of semantic objects to enable biologically meaningful interpretation.

Validating and measuring visual interpretability. Confirming visual interpretations of a non-trivial phenotype is hard because measuring the confidence in an interpretation and assessing whether one visual interpretation is better than another constitute an ill-defined problem. In practice, interpretability is usually subjectively validated in the final figure of a DL bioimaging paper, reporting the researchers' intuitions based on several 'representative' examples. More systematic validation mechanisms include assessing the robustness of the explanation by demonstrating that different models for the same task have similar interpretations and assessing sensitivity by showing that slight changes in the image space or the latent space ('adversarial attacks') that do not change the prediction also do not change the interpretation. If the interpreted phenotype can be explicitly measured, then it can be quantified by carefully (that is, avoiding confounders) correlating latent representations with phenotypes. In more complicated situations in which the interpreted phenotypes cannot be empirically measured, solutions include performing expert validation and evaluations[2,4] or altering the image (directly, not through the latent representation) according to the interpretation and then evaluating whether the model prediction was changed accordingly[2,14]. When live imaging is possible, one can examine spontaneous transitions between classification states that reduce most of the variability confounding the human ability to interpret the subtle phenotypes in snapshot images[1]. With the understanding that complex models are less interpretable, a possible approach to promote interpretability is reducing the model's complexity under the assumption that this will increase its 'interpretability potential'. For example, it is possible to constrain the training of the DL toward sparse representations and demonstrate that these representations perform well enough and are more interpretable[15]. Devising more objective and systematic metrics for visual interpretability remains an open challenge that we believe is critical for advancing the field.

More challenges and opportunities. Other promising future directions include developing better visualization approaches to ease the translation from explanation to interpretation, developing interpretability methods for more complex problems such as multi-class classification, effectively incorporating of 3D and/or temporal information in the interpretability method, and improving interpretability of high-dimensional image data (for example, single-cell spatial omics). Another opportunity lies in designing latent representations that can encourage better interpretability-driven discovery, such as disentangled representations in which each latent feature encodes a single semantic image property. Additional open questions that warrant systematic future investigations include "How much is the interpretability capacity conditioned on the accuracy of the model being interpreted?" and "Can we benefit from automated analysis of visual explanations, such as providing objective readouts indicating that a certain prediction cannot be trusted?"

Advancing in these directions will promote effective data-driven discovery by relying on DL's remarkable capacity to detect complex bioimage patterns to visually guide human interpretation and hypothesis generation.

Oded Rotem & Assaf Zaritsky
Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Be'er-Sheva, Israel.
e-mail: assafzar@gmail.com

Published online: 9 August 2024

References

1. Zaritsky, A. et al. *Cell Syst.* **12**, 733–747.e6 (2021).
2. DeGrave, A. J., Cai, Z. R., Janizek, J. D., Daneshjou, R. & Lee, S. I. *Nat. Biomed. Eng.* https://doi.org/10.1038/s41551-023-01160-9 (2023).
3. Lamiable, A. et al. *Nat. Commun.* **14**, 6386 (2023).
4. Rotem, O. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2023.11.15.566968. (2023).

5. Kobayashi, H., Cheveralls, K. C., Leonetti, M. D. & Royer, L. A. *Nat. Methods* **19**, 995–1003 (2022).
6. Razdaibiedina, A. et al. *Mol. Syst. Biol.* **20**, 521–548 (2024).
7. Lundberg, S. & Lee, S.-I. *NIPS'17: Proc. 31st International Conference on Neural Information Processing Systems* 4768–4777 (ACM, 2017).
8. Soelistyo, C. J. et al. *Nat. Mach. Intell.* **4**, 636–644 (2022).
9. Yamamoto, T., Cockburn, K., Greco, V. & Kawaguchi, K. *PLOS Comput. Biol.* **18**, e1010477 (2022).
10. Schmitt, M. S. et al. *Cell* **187**, 481–494.e24 (2024).
11. Doron, M. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2023.06.16.545359 (2023).
12. LaChance, J., Suh, K., Clausen, J. & Cohen, D. *J. PLOS Comput. Biol.* **18**, e1009293 (2022).
13. Yang, K. D. et al. *PLOS Comput. Biol.* **16**, e1007828 (2020).
14. Eckstein, N. et al. *Cell* **187**, 2574–2594 (2024).
15. Soelistyo, C. J. & Lowe, A. R. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2402.03115 (2024).

Acknowledgements
This research was supported by the Israeli Council for Higher Education (CHE) via the Data Science Research Center, Ben-Gurion University of the Negev, Israel (to AZ), and by the Rosetree trust (to AZ). We thank Nadav Rappoport, Meghan Driscoll, Orit Kliper-Gross and Kevin Dean for critically reading this Comment.

Author contributions
O.R. and A.Z. conceived and wrote this comment.

Competing interests
The authors declare no competing interests.