

The Emerging Issues in Bioimaging AI Publications and Research

Jianxu Chen^{*1}, Florian Jug^{*2}, Susanne Rafelski^{*3}, and Shanghang Zhang^{*4}

1 ISAS – Dortmund, DE. jianxu.chen@isas.de

2 Human Technopole – Milano, IT. florian.jug@fht.org

3 Allen Institute for Cell Science – Seattle, US. susanner@alleninstitute.org

4 Peking University, CN. shanghang@pku.edu.cn

Abstract

This report documents the program and outcomes of Dagstuhl Seminar “The Emerging Issues in Bioimaging AI Publications and Research” (24042) held on January 21–24, 2024. The fast advancement of computational techniques, particularly those based on artificial intelligence (AI), has significantly propelled the field of computational biology. With the rapid development, new issues are emerging in bioimaging AI publications and research. For example, how can we properly validate the AI methods used in quantitative biological analysis? Also, the ethical aspects of these developments remain underexplored, lacking clear definitions and recognition within the community. The goal of this interdisciplinary seminar was to bring together experts from various fields, including experimental biology, computational biology, bioimage analysis, computer vision, and AI research, to identify, discuss and address the emerging issues in current bioimaging AI research and publications.

Seminar January 21–24, 2024 – <https://www.dagstuhl.de/24042>

2012 ACM Subject Classification Applied computing → Imaging; Computing methodologies → Artificial intelligence

Keywords and phrases artificial intelligence, bioimaging, open source, publication ethics, trustworthy ai

Digital Object Identifier 10.4230/DagRep.14.1.90

1 Executive Summary

Jianxu Chen (ISAS – Dortmund, DE)

Florian Jug (Human Technopole – Milano, IT)

Susanne Rafelski (Allen Institute for Cell Science – Seattle, US)

Shanghang Zhang (Peking University, CN)

License  Creative Commons BY 4.0 International license

© Jianxu Chen, Florian Jug, Susanne Rafelski, and Shanghang Zhang

Seminar Structure and Organization

The seminar was divided into three specific directions: ethical considerations in bioimaging AI research and publications, performance reporting on bioimaging AI methods in publications and research, and future research directions of bioimaging AI focusing on validation and robustness. The seminar was structured into two parts: the first half focused on presentations and information sharing related to these three major directions to align experts from different fields, and the second half concentrated on in-depth discussions of these topics.

* Editor / Organizer



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

The Emerging Issues in Bioimaging AI Publications and Research, *Dagstuhl Reports*, Vol. 14, Issue 1, pp. 90–107
Editors: Jianxu Chen, Florian Jug, Susanne Rafelski, and Shanghang Zhang



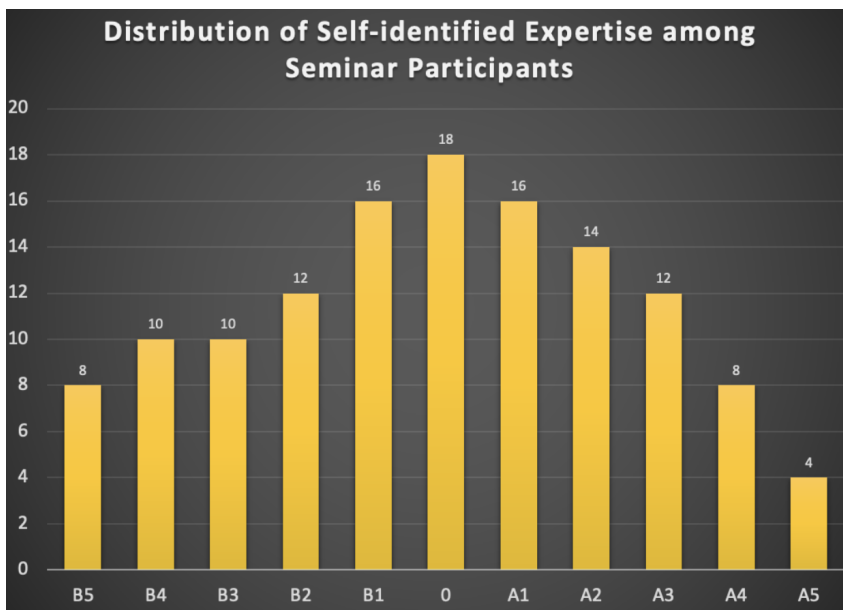
DAGSTUHL
REPORTS

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Given the highly interdisciplinary nature of the seminar, we took two specific steps to facilitate smooth communication and discussion among researchers with diverse backgrounds.

First, about six to eight weeks before the seminar, we sent out a survey to gather potential topics each participant could present within the seminar’s overarching theme. We collaborated with several participants to choose or adjust their presentation topics to ensure the effectiveness in this interdisciplinary setting. Based on the survey responses, the presentation and information-sharing portion (the first half) of our seminar began with two keynotes from editors who handle bioimaging AI papers, sharing their insights and the existing efforts by publishers. We then organized all presentations to progress from a focus on biology to bioimaging AI, and finally to AI, ensuring coverage of the full spectrum of necessary knowledge for our in-depth discussions in the second half.

Second, at the beginning of the seminar, we allocated two minutes for each participant for a quick introduction and to briefly rate their experience and expertise on a scale in the range of [B5, B4, B3, B2, B1, 0, A1, A2, A3, A4, A5], with B5 representing pure biology and A5 representing pure AI. Participants could select a single value, multiple values, or a range of values. This was not intended to stereotype participants but to facilitate easier communication. For example, if a participant with experience in the range of B5 to B3 spoke with two others during a coffee break, one with experience from B3 to A1 and the other from A3 to A5, different communication strategies would be necessary for effective discussions. The distribution of self-identified experience is summarized in the histogram below (see Fig. 1).



■ **Figure 1** Histogram of the distribution of expertised self-identified by seminar participants.

Presentations, discussions and outcomes

Overview of the scientific talks

The seminar began with presentations by editors from Nature Methods and Cell Press, who shared their insights on existing and emerging issues in bioimaging AI publications. Following this, general bioimage analysis validation issues were discussed from both a biological application perspective and an algorithmic metric perspective. These presentations were succeeded by specific application talks demonstrating how AI-based bioimage analysis is utilized and validated in high-throughput biological applications [1]. The remainder of day one focused on bioimaging AI validation through explainable AI [2], [3], [4] and existing tools [5], as well as community efforts in deploying FAIR (Findable, Accessible, Interoperable, Reusable) AI tools for bioimage analysis [6].

The second day commenced with several theoretical AI talks introducing key concepts related to model robustness, fairness, and trustworthiness [7]. These were followed by two presentations showcasing state-of-the-art AI algorithms applied in bioimaging [8], [9], and an overview of the application of foundation models in bioimaging [10]. The scientific presentation portion of the seminar concluded with a talk about the pilot work initiated by the EMBO (European Molecular Biology Organization) Press on research integrity and AI integration in publishing and trust. This talk also served as a transition into the in-depth discussions that comprised the second part of the seminar.

Summary of discussions and key outcomes

After the scientific presentation part of the seminar, the participants naturally reach the agreement on doing the discussion in a four-quadrant manner, as illustrated below in Fig. 2.

	“Users” of bioimaging AI	“Makers” of bioimaging AI
In-domain technical considerations	1	2
out-of-domain implications	3	4

■ **Figure 2** The four-quadrant for organizing the in-depth discussion.

Here are some examples of what emerges from discussions in each quadrant.

I. What are some technical considerations that users of AI should pay attention to?

When using a specific bioimage analysis model, it is crucial for users to have clear biological questions that align with the technical limitations of the bioimaging AI models. This is known as application-appropriate validation [11]. For example, the trustworthiness or validity of an AI-based microscopy image denoising model may differ significantly between a study that requires merely counting the number of nuclei in an image and one that aims to quantify the morphological properties of the nuclei.

II. What are some technical considerations that makers of AI should pay attention to?

When developing a bioimaging AI model, comprehensive evaluations and ablation studies are essential to explicitly demonstrate the model’s limitations or potential failures. For instance, evaluating a cell segmentation model under different conditions, such as various magnifications, signal-to-noise ratios, cell densities, and possibly different microscope modalities, is highly

beneficial. Providing a clear and detailed definition of the conditions under which the model has been evaluated helps users determine whether the model can be directly applied to their images or if it needs retraining or fine-tuning.

III. What are some important things the users of AI should make sure the makers of AI are aware of or should make clear to the makers of AI?

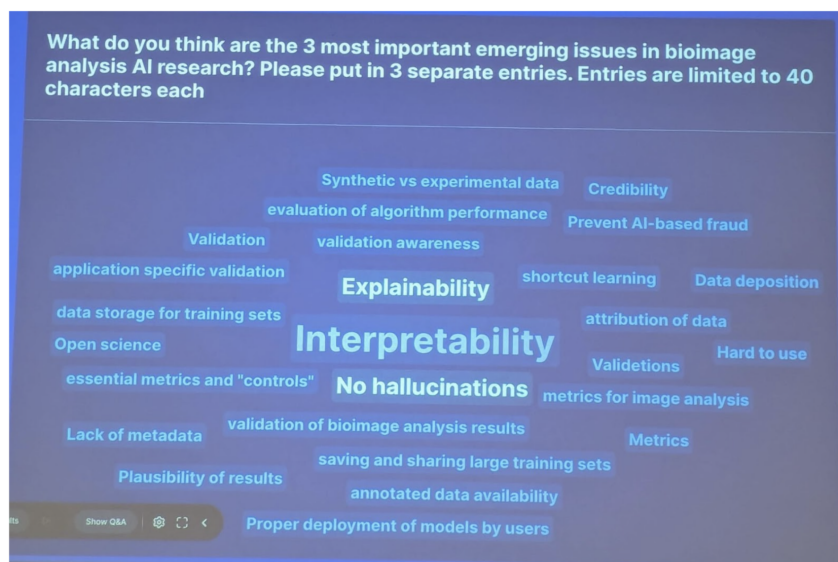
One example is the inherent presence of batch effects in biology, such as variations in fluorescence microscopy image quality due to different batches of dyes or slight morphological differences in cells from different colony positions. For effective interdisciplinary collaboration, it would be very helpful if biologists can clearly describe data acquisition processes and potential batch effects. This enables AI developers to consider these factors in their training sets, validation strategies, and model designs.

IV. What do the makers of AI need to make sure the users of AI know?

There is a lot of information that AI method developers need to help biologists think together. For example, in some collaborative projects, AI researchers need to guide their biologist collaborators how to best provide their data. For instance, the data to be analyzed to answer biological questions can be different from special data acquired merely for training the AI models, which could be referred to as a “training assay” [12], i.e., special experimental assays only made for effective model training.

The discussions highlighted in the four quadrants are only examples from the seminar. A follow-up “white paper”-like manuscript based on the full discussions is being planned as a resource for the bioimaging AI community.

A specific topic that emerged from the seminar was the interpretability and explainability of bioimaging AI models. This is evident from the word cloud generated during the discussions, as shown in Fig. 3. A follow-up seminar specifically focusing on this topic is being planned for the bioimaging AI community in the coming years.



■ **Figure 3** Word cloud generated during the discussion about the emerging issues in bioimaging AI publications and research.

Besides the science program, the seminar provided valuable opportunities for social connections and networking. Due to the pandemic, many researchers had previously only met virtually, making the in-person interactions feel like a reunion. The diversity of research fields

among participants, including wet-lab biologists and machine learning theorists with little biology experience, created unique networking opportunities. They would otherwise have rare opportunities to meet in traditional conferences. Biologists expressed that they gained new insights into the theories behind machine learning methods they had used, motivating them to rethink their future research designs. Conversely, machine learning researchers showed strong interest in collaborating with the bioimaging community to address fundamental challenges such as robustness and explainability.

Conclusions

This Dagstuhl Seminar on “The Emerging Issues in Bioimaging AI Publications and Research” successfully united a diverse group of experts from experimental biology, computational biology, bioimage analysis, computer vision, and AI research. The seminar facilitated in-depth discussions on ethical considerations, performance reporting, and future research directions in bioimaging AI, highlighting the crucial need for interdisciplinary collaboration and communication.

Through structured presentations and interactive discussions, participants underscored the importance of clear communication between AI developers and users, comprehensive model validation, and awareness of biological batch effects. The seminar emphasized the necessity for application-appropriate validation and detailed reporting of AI model conditions to enhance the trustworthiness and applicability of bioimaging AI methods. Furthermore, the seminar provided a valuable platform for social interactions and networking, bridging gaps between researchers from different fields and fostering new collaborations.

In conclusion, the seminar not only advanced discussions on critical issues in bioimaging AI publications but also laid the foundation for ongoing collaboration and innovation in the field. Planned follow-up activities will further contribute to the development and ethical application of AI in bioimaging research. The success of this seminar underscores the importance of continuous communication and cooperation in addressing the emerging challenges in bioimaging AI publications and research.

Acknowledgement

We are grateful to all seminar participants for their insightful contributions and the engaging discussions they fostered, especially in the interdisciplinary setting with a wide spectrum of expertise. We also sincerely thank the Dagstuhl Scientific Directorate for the opportunity to organize this event. Finally, our deepest appreciation goes to the exceptional Dagstuhl staff whose support was instrumental in making the seminar a success.

References

- 1 Z. Cibir et al., “ComplexEye: a multi-lens array microscope for high-throughput embedded immune cell migration analysis,” *Nat. Commun.*, vol. 14, no. 1, p. 8103, Dec. 2023, doi: 10.1038/s41467-023-43765-3.
- 2 Christopher J. Soelistyo and Alan R. Lowe, “Discovering Interpretable Models of Scientific Image Data with Deep Learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2024*, pp. 6884–6893. [Online].
- 3 Christopher J. Soelistyo, Guillaume Charras, and Alan R. Lowe, “Virtual Perturbations to Assess Explainability of Deep-Learning Based Cell Fate Predictors,” in *Proceedings of the*

- IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2023, pp. 3971–3980. [Online].
- 4 D. Schuhmacher et al., “A framework for falsifiable explanations of machine learning models with an application in computational pathology,” *Med. Image Anal.*, vol. 82, p. 102594, Nov. 2022, doi: 10.1016/j.media.2022.102594.
 - 5 L. M. Moser et al., “Piximi – An Images to Discovery web tool for bioimages and beyond.” Jun. 04, 2024. doi: 10.1101/2024.06.03.597232.
 - 6 W. Ouyang et al., “BioImage Model Zoo: A Community-Driven Resource for Accessible Deep Learning in BioImage Analysis,” *Bioinformatics*, preprint, Jun. 2022. doi: 10.1101/2022.06.07.495102.
 - 7 D. Guo, C. Wang, B. Wang, and H. Zha, “Learning Fair Representations via Distance Correlation Minimization,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 2139–2152, Feb. 2024, doi: 10.1109/TNNLS.2022.3187165.
 - 8 Saumya Gupta, Yikai Zhang, Xiaoling Hu, and Prateek Prasanna, “Topology-aware uncertainty for image segmentation,” presented at the Advances in Neural Information Processing Systems, 2024.
 - 9 G. Dai et al., “Implicit Neural Image Field for Biological Microscopy Image Compression.” arXiv, 2024. doi: 10.48550/ARXIV.2405.19012.
 - 10 A. Archit et al., “Segment Anything for Microscopy,” *Bioinformatics*, preprint, Aug. 2023. doi: 10.1101/2023.08.21.554208.
 - 11 J. Chen, M. P. Viana, and S. M. Rafelski, “When seeing is not believing: application-appropriate validation matters for quantitative bioimage analysis,” *Nat. Methods*, vol. 20, no. 7, pp. 968–970, Jul. 2023, doi: 10.1038/s41592-023-01881-4.
 - 12 J. Chen et al., “The Allen Cell and Structure Segmenter: a new open source toolkit for segmenting 3D intracellular structures in fluorescence microscopy images,” *Cell Biology*, preprint, Dec. 2018. doi: 10.1101/491035.

2 Table of Contents

Executive Summary

Jianxu Chen, Florian Jug, Susanne Rafelski, and Shanghang Zhang 90

Overview of Talks

Topological Uncertainty and Representation in Biomedical Image Analysis
Chao Chen 97

Application-appropriate validation matters for quantitative bioimage analysis
Jianxu Chen, Matheus Palhares Viana, and Susanne Rafelski 98

Metrics reloaded: Recommendations for image analysis validation
Evangelia Christodoulou 98

Working towards pick 5: strategies for scaling and distributing user-friendly containers
Beth Cimini 99

Implicit Neural Representation (INR) for Biological Image Compression and Neural Plasticity Learning
Gaole Dai 100

An overview of Cell Press policies on image presentation, data and code sharing, and AI use
Andrew Hufton 100

Discovering interpretable models of scientific image data with deep learning
Alan Lowe 101

Improving trustworthiness of ML in bioimaging through experimentally testable explanations
Axel Mosig 101

Segment Anything for Microscopy
Constantin Pape 103

High-volume, label-free imaging for quantifying single-cell dynamics in iPSC colonies
Anne Plant 103

Publishing microscopy and AI in Nature Methods
Rita Strack 104

Frequency shortcuts learning and generalization in computer vision
Nicola Strisciuglio 104

Visual interpretability of deep learning models in cell imaging
Assaf Zaritsky 105

Foundation models for biomedical image analysis
Shanghang Zhang 105

Algorithmic Fairness, Robust Generalization and Trustworthy Machine Learning
Han Zhao 106

Participants 107

3 Overview of Talks

3.1 Topological Uncertainty and Representation in Biomedical Image Analysis

Chao Chen (Stony Brook University, US)

- License** © Creative Commons BY 4.0 International license
© Chao Chen
- Main reference** Xiaoling Hu, Fuxin Li, Dimitris Samaras, Chao Chen: “Topology-Preserving Deep Image Segmentation”, in Proc. of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pp. 5658–5669, 2019.
URL <https://proceedings.neurips.cc/paper/2019/hash/2d95666e2649fcfc6e3af75e09f5adb9-Abstract.html>
- Main reference** Saumya Gupta, Yikai Zhang, Xiaoling Hu, Prateek Prasanna, Chao Chen: “Topology-Aware Uncertainty for Image Segmentation”, in Proc. of the Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 – 16, 2023, 2023.
URL http://papers.nips.cc/paper_files/paper/2023/hash/19ded4cfc36a7feb7fce975393d378fd-Abstract-Conference.html
- Main reference** Shahira Abousamra, Rajarsi Gupta, Tahsin M. Kurç, Dimitris Samaras, Joel H. Saltz, Chao Chen: “Topology-Guided Multi-Class Cell Context Generation for Digital Pathology”, in Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pp. 3323–3333, IEEE, 2023.
URL <https://doi.org/10.1109/CVPR52729.2023.00324>

Modern analytics is facing highly complex and heterogeneous data. While deep learning models have pushed our prediction power to a new level, they are not satisfactory in some crucial merits such as transparency, robustness, data-efficiency, etc. To address these challenges, I am generally interested in incorporating mathematical modeling of topology, geometry and dynamics seamlessly into the learning pipeline. Such model-informed learning approach will be more transparent, steerable and less annotation-hungry.

In this talk, I will focus on our recent work on combining topological reasoning with learning to solve problems in biomedical image analysis. With advanced imaging techniques, we are collecting images of various complex structures such as neurons, vessels, tissues and cells. These structures encode important information about underlying biological mechanisms. To fully exploit these structures, we propose to enhance learning pipelines with topology, the branch of abstract mathematics that deals with structures such as connections, loops and branches. Under the hood is a formulation of the topological computation as a robust and differentiable operator, based on the theory of persistent homology. This inspires a series of novel methods for segmentation, uncertainty estimation, generation, and analysis of these topology-rich biomedical structures.

3.2 Application-appropriate validation matters for quantitative bioimage analysis

Jianxu Chen (ISAS – Dortmund, DE), Matheus Palhares Viana (Allen Institute for Cell Science – Seattle, US), and Susanne Rafelski (Allen Institute for Cell Science – Seattle, US)

License  Creative Commons BY 4.0 International license

© Jianxu Chen, Matheus Palhares Viana, and Susanne Rafelski

Main reference Jianxu Chen, Matheus P. Viana, and Susanne Rafelski: “When seeing is not believing: application-appropriate validation matters for quantitative bioimage analysis”. *Nat Methods* 20, 968–970 (2023).

URL <https://doi.org/10.1038/s41592-023-01881-4>

A critical step towards biologically reliable analysis of microscopy image-based assays is rigorous quantitative validation with metrics and measurements appropriate for the particular biological application. Currently, however, no community standards or publication guidelines exist on how to conduct the appropriate validation of work involving quantitative analysis of microscopy images, including deep-learning based approaches. In this presentation, we discussed this challenge for both classical and modern deep-learning based image analysis approaches as well as possible solutions for automating and streamlining the validation process. First, to introduce the concept of “application-appropriate validation”, we showed a true story of how inappropriate validation of segmentations in a quantitative analysis of mitochondrial network morphology led to wrong biological conclusions. Second, besides segmentation, we showed another example of application-appropriate validation for label-free predictions. The commonly used metrics, e.g., Pearson correlation or structure similarity, could be misleading when not taking the downstream biological application into account. Finally, we discussed a list of key considerations for interpretable quantification of microscopy image-based assays, from understanding the underlying biological questions, understanding the limits of assays, understanding the validation requirement for interpretation, to the estimation of time and effort one could afford, with an emphasis on future community efforts in standardization, dissemination and interdisciplinary connections.

3.3 Metrics reloaded: Recommendations for image analysis validation

Evangelia Christodoulou (DKFZ – Heidelberg, DE)

License  Creative Commons BY 4.0 International license

© Evangelia Christodoulou

Main reference Lena Maier-Hein, Annika Reinke, Evangelia Christodoulou, Ben Glocker, Patrick Godau, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael A. Riegler, Manuel Wiesenfarth, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzl, A. Emre Kavur, Tim Rädtsch, Minu Dietlinde Tizabi, Laura Ación, Michela Antonelli, Tal Arbel, Spyridon Bakas, Peter Bankhead, Arriel Benis, M. Jorge Cardoso, Veronika Cheplygina, Beth A. Cimini, Gary S. Collins, Keyvan Farahani, Bram van Ginneken, Daniel A. Hashimoto, Michael M. Hoffman, Merel Huisman, Pierre Jannin, Charles E. Kahn, Alexandros Karargyris, Alan Karthikesalingam, Hannes Kenngott, Annette Kopp-Schneider, Anna Kreshuk, Tahsin M. Kurç, Bennett A. Landman, Geert Litjens, Amin Madani, Klaus H. Maier-Hein, Anne L. Martel, Peter Mattson, Erik Meijering, Bjoern H. Menze, David Moher, Karel G. M. Moons, Henning Müller, Felix Nickel, Brennan Nichyporuk, Jens Petersen, Nasir M. Rajpoot, Nicola Rieke, Julio Saez-Rodriguez, Clarisa Sánchez Gutiérrez, Shravya Shetty, Maarten van Smeden, Carole H. Sudre, Ronald M. Summers, Abdel A. Taha, Sotirios A. Tsaftaris, Ben Van Calster, Gaël Varoquaux, Paul F. Jäger: “Metrics reloaded: Pitfalls and recommendations for image analysis validation”, *CoRR*, Vol. abs/2206.01653, 2022.

URL <https://doi.org/10.48550/ARXIV.2206.01653>

Increasing evidence shows that flaws in machine learning (ML) algorithm validation are an underestimated global problem. Particularly in automatic biomedical image analysis, chosen performance metrics often do not reflect the domain interest, thus failing to adequately

measure scientific progress and hindering translation of ML techniques into practice. To overcome this, our large international expert consortium created Metrics Reloaded, a comprehensive framework guiding researchers in the problem-aware selection of metrics. Following the convergence of ML methodology across application domains, Metrics Reloaded fosters the convergence of validation methodology. The framework was developed in a multi-stage Delphi process and is based on the novel concept of a problem fingerprint – a structured representation of the given problem that captures all aspects that are relevant for metric selection, from the domain interest to the properties of the target structure(s), data set and algorithm output. Based on the problem fingerprint, users are guided through the process of choosing and applying appropriate validation metrics while being made aware of potential pitfalls. Metrics Reloaded targets image analysis problems that can be interpreted as a classification task at image, object or pixel level, namely image-level classification, object detection, semantic segmentation, and instance segmentation tasks. To improve the user experience, we implemented the framework in the Metrics Reloaded online tool, which also provides a point of access to explore weaknesses, strengths and specific recommendations for the most common validation metrics. The broad applicability of our framework across domains is demonstrated by an instantiation for various biological and medical image analysis use cases.

3.4 Working towards pick 5: strategies for scaling and distributing user-friendly containers

Beth Cimini (Broad Institute of MIT & Harvard – Cambridge, US)

License  Creative Commons BY 4.0 International license
© Beth Cimini

In the current scientific software environment, we have identified 5 axes that should be measured for any software or code distribution system with which one plans to share code.

Reproducible – can you tell what went in there and why and how?

Easy to create – how much extra work/knowledge is needed on the developer side to package?

Easy to run – how much extra work/knowledge is needed on the user side to run?


Long lasting – will the thing I made today work tomorrow?

Scalable – if my experiments get bigger (in terms of individual image size and/or parallelization of many images, can I still use my solution?)

In this talk, we discuss the merits of packaged applications, virtual environment spec files, online workflow tools like Galaxy, as well as software containers. We discuss strategies that can be used alongside software containers to make them maximally user friendly, and discuss possible strategies to make containers score high on all 5 axes.

3.5 Implicit Neural Representation (INR) for Biological Image Compression and Neural Plasticity Learning

Gaole Dai (*Peking University, CN*)

License  Creative Commons BY 4.0 International license
© Gaole Dai

The presentation introduces Implicit Neural Representation (INR), which is employed for various tasks such as image compression and 3D reconstruction. Specifically, we explore the distinctive attributes of INR in bioimage data compression. Additionally, we investigate the integration of the coordinate-to-value learning approach of INR into conventional Artificial Neural Networks (ANNs). By assigning specific coordinates to each cell/synapse in the ANN and integrating them with the INR network, we obtain tailored adjustment values for each location. We find that this type of adjustment exhibits neural plasticity, a characteristic unique to biological networks, making it highly valuable in Parameter Efficient Fine-Tuning (PEFT) tasks.

3.6 An overview of Cell Press policies on image presentation, data and code sharing, and AI use

Andrew Hufton (*Patterns, Cell Press – Würzburg, DE*)

License  Creative Commons BY 4.0 International license
© Andrew Hufton

The Cell Press journals, including *Patterns* (<https://www.cell.com/patterns/>), have high standards for the transparency and reproducibility of research presented at our journals. In my talk, I presented a brief overview of our policies on image presentation, data and code sharing, and the use of AI tools in research and manuscript preparation. I then discussed how these policies apply to cutting-edge bioimaging research and some of the challenges editors and our authors commonly face during the peer-review and publication process. Notably, I made the case that authors should think critically about the openness, ethics and transparency of AI models and training datasets used in their research, and should keep in mind that reliance on closed-source commercial models could impact the transparency and publishability of their work. I also highlighted some of the dangers of poorly-designed AI detection tools, and argued that while we must be vigilant against AI-enabled fraud, our main focus as a community should be on positively promoting and rewarding innovative, rigorous and open research. A selection of papers mentioned in my talk are included below.

References

- 1 Bagheri, N., et al (2023) The new era of quantitative cell imaging—challenges and opportunities. *Mol. Cell* 82, 241-247. <https://doi.org/10.1016/j.molcel.2021.12.024>
- 2 Gu, J., et al (2022) AI-enabled image fraud in scientific publications. *Patterns* 3, 100511. <https://doi.org/10.1016/j.patter.2022.100511>
- 3 Liang, W., et al (2023) GPT detectors are biased against non-native English writers. *Patterns* 4, 100779. <https://doi.org/10.1016/j.patter.2023.100779>
- 4 Wang, W., et al. (2023) On the transparency of large AI models. *Patterns* 4, 100797. <https://doi.org/10.1016/j.patter.2023.100797>

3.7 Discovering interpretable models of scientific image data with deep learning

Alan Lowe (*The Alan Turing Institute – London, GB*)

License © Creative Commons BY 4.0 International license
© Alan Lowe

Joint work of Christopher Soelistyo, Alan Lowe

Main reference Christopher J. Soelistyo, Alan R. Lowe: “Discovering interpretable models of scientific image data with deep learning”, CoRR, Vol. abs/2402.03115, 2024.

URL <https://doi.org/10.48550/ARXIV.2402.03115>

Deep learning (DL) is now a powerful tool in microscopy data analysis, routinely used for image processing applications such as segmentation and denoising. However, it is rarely used to directly learn scientific models of a biological system, owing to the complexity of the internal representations. Here, we present our recent attempts to learn interpretable DL-based models of complex cell biological phenomena directly from a large corpus of time-lapse imaging data. In particular, we implement disentangled representation learning, causal time series models, network sparsity and symbolic methods, and assess their usefulness in forming interpretable models of complex data. We find that such methods can produce highly parsimonious models that achieve $\sim 98\%$ of the accuracy of black-box benchmark models, with a tiny fraction of the complexity. We explore the utility of such interpretable models in producing scientific explanations of the underlying biological phenomenon.

References

- 1 Soelistyo, Christopher and Vallardi, Giulia and Charras, Guillaume and Lowe, Alan. (2022) *Learning biophysical determinants of cell fate with deep neural networks*. Nature Machine Intelligence
- 2 Soelistyo, Christopher and Charras, Guillaume and Lowe, Alan. (2023) *Virtual perturbations to assess explainability of deep-learning based cell fate predictors*. In Proceedings of the IEEE/CVF International Conference on Computer Vision
- 3 Soelistyo, Christopher and Lowe, Alan. (2024) *Discovering interpretable models of scientific image data with deep learning*. arXiv preprint arXiv:2402.03115

3.8 Improving trustworthiness of ML in bioimaging through experimentally testable explanations

Axel Mosig (*Ruhr-Universität Bochum, DE*)

License © Creative Commons BY 4.0 International license
© Axel Mosig

Main reference David Schuhmacher, Stephanie Schörner, Claus Küpper, Frederik Großerüschkamp, Carlo Sternemann, Celine Lugnier, Anna-Lena Kraeft, Hendrik Jütte, Andrea Tannapfel, Anke Reinacher-Schick, Klaus Gerwert, Axel Mosig: “A framework for falsifiable explanations of machine learning models with an application in computational pathology”, *Medical Image Anal.*, Vol. 82, p. 102594, 2022.

URL <https://doi.org/10.1016/J.MEDIA.2022.102594>

The black box nature of neural networks is commonly regarded as the main source why predictions obtained from deep neural networks, despite their often unprecedented predictive accuracy, are often considered untrustworthy. In this contribution, I argue that the lack of trustworthiness of machine learning in general is due to its inductive nature: Machine learning models are obtained from inductive inferences, where specific observations in the form of training data are used to infer a general model that can classify data points beyond

the training data. From this perspective, machine learning is subject to the problem of induction, which has been brought to the point by the no-free-lunch theorem: Since there is no justification to assume that future events will resemble the past, all machine learning algorithms perform equal in terms of their out-of-training error.

Our further reasoning follows two interpretations, a global and a local interpretation, of the no-free-lunch theorem, which have been formulated recently by Sterkenburg and Grünwald. The global interpretation is in a sense the pessimistic interpretation, stating that no universal learning algorithm exists, since across the domain of all possible learning problems, all classifiers are identical in terms of their out-of-training error. The local interpretation is more constructive towards applied machine learning: When dealing with one specific problem, some learning algorithms do perform better on this specific task than other learning algorithms. This can be understood in terms of the inductive bias of different learning algorithms: As a direct implication of the no-free-lunch theorem, each learning algorithms must involve an either implicit or an explicit set of assumptions about how to generalize to data points beyond the training data. This set of assumptions is referred to as the inductive bias of a learning algorithm. From the perspective of one specific learning task, one can now ask what learning algorithm has an inductive bias that matches the underlying learning problem. This local interpretation of the no-free-lunch theorem essentially leads to considering machine learning as an inductive bias modeling problem.

The question that follows the local interpretation of the no-free-lunch theorem is how to justify inductive bias. I argue that our recently proposed framework for falsifiable explanations of artificial intelligence, or FXAI framework for short, addresses this question: The FXAI framework builds on the concept of explainability methods for neural networks, which usually provide an explainable output along with the classification of an input item. In the case of image classification, for example, the interpretable output is often a heat map that indicates which input variables have been relevant for obtaining the classification result of a specific image. In the FXAI framework, this explainable extension of the output is referred to as the interpretable space, or I-space for short. It is important to realize that an I-space, while being interpretable, can usually not be considered an interpretation in itself. The role of an interpretation (or, synonymously, an explanation) is rather assigned to a hypothesis that, in the sense of a scientific hypothesis, is required to be experimentally testable. The latter criterion is of crucial importance: Now, the explanation – and along with it, the I-space and hence the machine learning model – can be tested experimentally.

Experimental testability has relevant consequences: First of all, since the experiment that tests the explaining hypothesis is a different experiment than the experiment that yielded the data that were input to the machine learning model, the FXAI framework yields an experimental, deductive path to validate a machine learning model that is fully independent of cross validation. Second, the testable hypothesis suggests what should guide the inductive bias of a learning algorithm: namely an experimentally testable hypothesis.

We can now finally argue why experimentally testable explanations improve the trustworthiness of machine learning models. My argument lies in the nature of scientific hypotheses, which usually do not refer to one specific experiment. Rather, a strong hypothesis will usually suggest a wide range of different experiments through which the hypothesis can be tested. The more experiments a hypothesis invites for it to be tested, the more vulnerable the hypothesis becomes, because each experiment potentially falsifies the hypothesis. If, on the other hand, the hypothesis withstands all experimental attempts to falsify it, then the trustworthiness of the hypothesis and with it the associated machine learning model is undermined.

References

- 1 David Schuhmacher, Stephanie Schörner, Claus Küpper, Frederik Großerueschkamp, Carlo Sternemann, Celine Lugnier, Anna-Lena Kraeft, Hendrik Jütte, Andrea Tannapfel, Anke Reinacher-Schick, et al. A framework for falsifiable explanations of machine learning models with an application in computational pathology. *Medical Image Analysis*, 82:102594, 2022.
- 2 Tom F Sterkenburg and Peter D Grünwald. The no-free-lunch theorems of supervised learning. *Synthese*, 199(3-4):9979–10015, 2021.

3.9 Segment Anything for Microscopy

Constantin Pape (Universität Göttingen, DE)

License © Creative Commons BY 4.0 International license
© Constantin Pape

Main reference Anwai Archit, Sushmita Nair, Nabeel Khalid, Paul Hilt, Vikas Rajashekar, Marei Freitag, Sagnik Gupta, Andreas Dengel, Sheraz Ahmed, Constantin Pape: “Segment Anything for Microscopy”, bioRxiv, Cold Spring Harbor Laboratory, 2023.

URL <https://doi.org/10.1101/2023.08.21.554208>

The segmentation of cells in light microscopy or organelles in electron microscopy is one of the fundamental tasks in microscopy image analysis. While deep learning based approaches have improved segmentation qualities for a wide array of tasks, these solutions require specialized architectures and, unless very similar training data is publicly available, a significant amount of manual annotation. Recently versatile models that can be applied to a wider set of vision tasks – commonly referred to as foundation models – have been introduced. These models promise to bridge this gap and enable readily available solutions for many vision tasks. The foundation model “Segment Anything” developed by Meta implements this paradigm for segmentation tasks and can be applied for interactive and automatic segmentation in a large variety of image modalities. Our work builds on Segment Anything and evaluates and improves it for microscopy data. In particular, we implement a fine-tuning methodology that significantly improves the quality for microscopy and a software plugin for fast interactive data annotation., showing the promise of vision foundation models for microscopy image analysis.

3.10 High-volume, label-free imaging for quantifying single-cell dynamics in iPSC colonies

Anne Plant (NIST – Gaithersburg, US)

License © Creative Commons BY 4.0 International license
© Anne Plant

Joint work of Anthony Asmar, Zackery Benson, Adele Peskin, Mylene Simon, Michael Halter
Main reference Anthony Asmar, Zack Benson, Adele P. Peskin, Joe Chalfoun, Mylene Simon, Michael Halter, Anne Plant: “High-volume, label-free imaging for quantifying single-cell dynamics in induced pluripotent stem cell colonies”, bioRxiv, Cold Spring Harbor Laboratory, 2023.


URL <https://doi.org/10.1101/2023.09.29.558451>

To facilitate the characterization of unlabeled induced pluripotent stem cells (iPSCs) during culture and expansion, and to be able to address gene expression in individual living cells over time, we developed an AI pipeline for nuclear segmentation and mitosis detection from phase contrast images of individual cells within iPSC colonies. The analysis uses a 2D convolutional neural network (U-Net) plus a 3D U-Net applied on time lapse images to detect and segment

nuclei, mitotic events, and daughter nuclei to enable tracking of hundreds of thousands of individual cells over long times in culture. The analysis uses fluorescence data to train models for segmenting nuclei in phase contrast images. The use of classical image processing routines to segment fluorescent nuclei precludes the need for manual annotation and provides hundreds of thousands of cell objects for training. We explored reproducibility and generalizability of the pipeline, and how pipeline parameters influenced metrics of accuracy. The model is generalizable in that it performs well on different datasets with an average F1 score of 0.94, on cells at different densities, and on cells from different pluripotent cell lines. The method allows us to assess, in a non-invasive manner, rates of mitosis and cell division which serve as indicators of cell state and cell health. We assess these parameters in culture for more than 36 hours, at different locations in the colonies, and as a function of excitation light exposure.

3.11 Publishing microscopy and AI in Nature Methods


Rita Strack (Nature Publishing Group, US)

License  Creative Commons BY 4.0 International license
© Rita Strack

Reporting microscopy data and metadata are critical for data reproducibility, sharing, and reuse, and journals can have a key role in improving reporting standards. This talk discussed a methodological reporting crisis in microscopy, published works seeking to address this issue, and standards that are being implemented at Nature Methods. It also discussed the unique challenges associated with publishing reproducible AI for use in bioimage analysis and why this is crucial for the field moving forward. The goal was to inspire researchers to develop and implement best practice to promote reproducibility and growth within the field.

3.12 Frequency shortcuts learning and generalization in computer vision

Nicola Strisciuglio (University of Twente – Enschede, NL)

License  Creative Commons BY 4.0 International license
© Nicola Strisciuglio

Main reference Shunxin Wang, Raymond N. J. Veldhuis, Christoph Brune, Nicola Strisciuglio: “What do neural networks learn in image classification? A frequency shortcut perspective”, in Proc. of the IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023, pp. 1433–1442, IEEE, 2023.

URL <https://doi.org/10.1109/ICCV51070.2023.00138>

Neural networks trained through optimization techniques based on variants of stochastic gradient descent (SGD) present a spectral bias. Model training dynamics are biased towards learning features related to low-frequency components of the input data at early stages of training, and subsequently focusing on high-frequency features. Another important phenomenon in training dynamics is the emergence of shortcut learning, that is learning spurious correlations between the input data and prediction target. This results from the tendency of SGD-based training to find solutions that simplify the minimization of a loss function used as target of the training optimization problem. Shortcuts harm the generalization abilities of neural networks, especially in out-of-distribution (OOD) settings, and thus require particular attention during model validation.

We investigate the relationship between spectral bias and shortcut learning in image classification and expose the existence of shortcuts learned by vision models in the Fourier domain, which we call frequency shortcuts. We propose a method to detect possible frequency shortcuts, based on the importance that single frequency components have in the classification task, and construct dominant frequency maps (DFM). We demonstrate that frequency shortcuts can be learned at low or high-frequency and potentially harm the generalization capabilities in out-of-distribution settings, showing that shortcuts presents in OOD data can cause an illusion of strong generalization. In order to mitigate their impact on model performance, we also investigate the use of DFMs in a negative data augmentation strategy that improves adversarial robustness. However, extensive analysis of shortcuts learned by vision models is necessary and requires substantial attention to validate model performance and transferability to real-world tasks.

3.13 Visual interpretability of deep learning models in cell imaging

Assaf Zaritsky (Ben Gurion University – Beer Sheva, IL)

License © Creative Commons BY 4.0 International license
© Assaf Zaritsky

Joint work of Oded Rotem, Tamar Schwartz, Ron Maor, Yishay Tauber, Maya Tsarfati Shapiro, Marcos Meseguer, Daniella Gilboa, Daniel S. Seidman, Assaf Zaritsky

Main reference Oded Rotem, Tamar Schwartz, Ron Maor, Yishay Tauber, Maya Tsarfati Shapiro, Marcos Meseguer, Daniella Gilboa, Daniel S. Seidman, Assaf Zaritsky: “Visual interpretability of image-based classification models by generative latent space disentanglement applied to in vitro fertilization”, bioRxiv, Cold Spring Harbor Laboratory, 2023.

URL <https://doi.org/10.1101/2023.11.15.566968>

With the rapid growing volume and complexity of modern biomedical visual data, we can no longer rely on humans’ amazing capacity to identify visual patterns in biomedical images. Deep learning has emerged as a powerful technique to identify hidden patterns that exceed human intuition in complex cell imaging data. Extracting a deeper biological understanding, such as mechanistic description of complex phenotypes, require human interpretable explanation of the deep learning model’s decision process, however, the non-linear entanglement of image features makes deep learning models a “black box” that lacks straightforward explanations of which biologically meaningful image properties are important for the models’ decision. In my talk I presented a new generalized method toward systematic visual interpretability of deep learning image-based classification models that relies on counterfactual visual explanations using a disentangled latent representation. This method enables visually intuitive traversal of the latent space and we applied it to decipher blastocysts morphological quality properties in the context of in vitro fertilization.

3.14 Foundation models for biomedical image analysis

Shanghang Zhang (Peking University, CN)

License © Creative Commons BY 4.0 International license
© Shanghang Zhang

In this presentation, we delve into the potential of Foundation Models (FMs), which encompass Large Language Models (LLMs), Large Vision Models (VLMs), and Multimodal Large Language Models (MLLM). These models have demonstrated promising outcomes in various

scenarios. However, integrating FM capabilities into professional domains remains an unresolved inquiry. We present some recent relevant research endeavours to address emergent challenges during this transition. The initial query pertains to efficiently aligning data from specialized domains with non-specific FMs. Parameter Efficient Fine-tuning (PEFT) offers a viable approach, and to adapt PEFT more suitably for medical data in our case, we have devised a tree-like structured adapter that hierarchically incorporates medical knowledge into the Segment Anything (SAM) model. Secondly, we illustrate how quantization techniques can accelerate FM performance for biological tasks. Subsequently, we demonstrate the fine-tuning process of an MLLM using medical data to generate medical reports and accomplish vision-based question-answering tasks. This process leverages methods such as in-context learning to align training data across different modalities with retrieval augmentative generation to support our model giving a more comprehensive report.

3.15 Algorithmic Fairness, Robust Generalization and Trustworthy Machine Learning

Han Zhao (University of Illinois – Urbana-Champaign, US)

License  Creative Commons BY 4.0 International license
© Han Zhao

Joint work of Han Zhao, Haoxiang Wang, Haozhe Si, Gargi Balasubramaniam, Bo Li

In this talk, I will discuss two important aspects of machine learning: algorithmic fairness and robust generalization under the common framework of invariant causal prediction. I will first provide some motivating examples of these two problems in the context of biomedical and healthcare applications. I will then introduce our recent work [1, 2] on invariant feature recovery to address the above two problems. I will conclude the talk with a discussion of some open problems and future research directions.

References

- 1 Wang, Haoxiang and Si, Haozhe and Li, Bo and Zhao, Han. *Provable domain generalization via invariant-feature subspace recovery*. In Proceedings of the 39th International Conference on Machine Learning (ICML 2022)
- 2 Wang, Haoxiang and Balasubramaniam, Gargi and Si, Haozhe and Li, Bo and Zhao, Han. *Invariant-Feature Subspace Recovery: A New Class of Provable Domain Generalization Algorithms*. arXiv preprint arXiv:2311.00966

Participants

- Chao Chen
Stony Brook University, US
- Jianxu Chen
ISAS – Dortmund, DE
- Evangelia Christodoulou
DKFZ – Heidelberg, DE
- Beth Cimini
Broad Institute of MIT & Harvard – Cambridge, US
- Gaole Dai
Peking University, CN
- Meghan Driscoll
University of Minnesota – Minneapolis, US
- Edward Evans III
University of Wisconsin – Madison, US
- Matthias Gunzer
Universität Duisburg-Essen, DE & ISAS e.V. – Dortmund, DE
- Andrew Hufton
Patterns, Cell Press – Würzburg, DE
- Florian Jug
Human Technopole – Milano, IT
- Anna Kreshuk
EMBL – Heidelberg, DE
- Thomas Lemberger
EMBO – Heidelberg, DE
- Alan Lowe
The Alan Turing Institute – London, GB
- Shalin Mehta
Chan Zuckerberg Biohub – Stanford, US
- Axel Mosig
Ruhr-Universität Bochum, DE
- Matheus Palhares Viana
Allen Institute for Cell Science – Seattle, US
- Constantin Pape
Universität Göttingen, DE
- Anne Plant
NIST – Gaithersburg, US
- Susanne Rafelski
Allen Institute for Cell Science – Seattle, US
- Ananya Rastogi
Springer Nature – New York, US
- Albert Sickmann
ISAS – Dortmund, DE
- Rita Strack
Nature Publishing Group, US
- Nicola Strisciuglio
University of Twente – Enschede, NL
- Aubrey Weigel
Howard Hughes Medical Institute – Ashburn, US
- Assaf Zaritsky
Ben Gurion University – Beer Sheva, IL
- Shanghang Zhang
Peking University, CN
- Han Zhao
University of Illinois – Urbana-Champaign, US

