Building Tools for Machine Learning and Artificial Intelligence in Cancer Research: Best Practices and a Case Study with the PathML Toolkit for Computational Pathology

Jacob Rosenthal[1,2], Ryan Carelli[1,2], Mohamed Omar[2], David Brundage[1,2], Ella Halbert[1,3], Jackson Nyman[1], Surya N. Hari[1], Eliezer M. Van Allen[1,4], Luigi Marchionni[2], Renato Umeton[1,2,5,6], and Massimo Loda[1,2]

ABSTRACT

Imaging datasets in cancer research are growing exponentially in both quantity and information density. These massive datasets may enable derivation of insights for cancer research and clinical care, but only if researchers are equipped with the tools to leverage advanced computational analysis approaches such as machine learning and artificial intelligence. In this work, we highlight three themes to guide development of such computational tools: scalability, standardization, and ease of use. We then apply these principles to develop PathML, a general-purpose research toolkit for computational pathology. We describe the design of the PathML framework and demonstrate applications in diverse use cases. PathML is publicly available at www.pathml.com.

Big Data, Image Analysis, and Machine Learning in Cancer Research

Imaging has long been a cornerstone of cancer research and clinical care, providing insight into tissue morphology and spatial intercellular dynamics. Technological advances in recent years have enabled microscopy at a larger scale than ever before, leading to exponential growth in the size of commonly available datasets—a trend that is likely to continue to accelerate in coming years.

"Big Data" in biomedical imaging can be conceptualized along two orthogonal axes: sample size and data dimensionality (**Fig. 1**). The first axis (n) can be measured by simply counting the number of cases in a dataset. Scaling in this dimension has been chiefly driven by advances in high-throughput imaging technologies. A notable example can be seen in the field of pathology, where increasing adoption of digital workflows results in slide scanning being routinely incorporated into pathologists' workflows, consequently creating large databases of whole slide images (WSI). Early adopters of digital pathology workflows are scanning more than 1 million slides per year (1)—several orders of magnitude larger than current benchmark datasets such as The Cancer Genome Atlas, and an indication of the potential volume of data that large academic tertiary care hospitals can expect to routinely generate as workflows are increasingly digitized.

At the same time, data are also growing in the amount of information captured in each image, which we refer to as data dimensionality (d). This is chiefly driven by emerging technologies in spatial omics (i.e., spatial quantification of molecular markers such as proteins or RNA) and highly multiplexed imaging (reviewed in ref. 2). In contrast to brightfield images with three channels (red, green, and blue), each of these high-dimensional images may have upward of 10,000 channels, each representing a specific target. Volumetric imaging further increases information content in each specimen by adding a depth dimension, enabling the capture of three-dimensional tissue morphology. Thus, dataset sizes can grow even while the number of cases remains constant.

This rapid proliferation of imaging data has significant implications for cancer research, especially in conjunction with accompanying metadata such as genomics and outcomes. Large sample sizes provide sufficient power for discovery and quantification of histologic patterns associated with clinically and biologically relevant features, with recent work demonstrating the potential of these methods to improve clinical and diagnostic workflows (3–5) and discover image-based biomarkers that recapitulate molecular features (6, 7). Similarly, the rich contextual information captured in high-dimensional imaging data lays the groundwork for interrogation of tumor microenvironment at unprecedented resolution (8, 9). The ubiquity of brightfield microscopy makes it an especially attractive candidate for image-based biomarker development, as digital workflows are increasingly deployed in a wider variety of clinical contexts.

However, while increasing scale of imaging datasets presents new opportunities and avenues of investigation, it also presents major challenges. Namely, these advances are only possible by leveraging computational image analysis methods, particularly deep learning. Deep learning models are flexible and powerful and have demonstrated remarkable success at identifying patterns in large datasets. As cancer research enters the age of "Big Data," machine learning and is therefore poised to become an increasingly essential tool in the

[1]Dana-Farber Cancer Institute, Boston, Massachusetts. [2]Weill Cornell Medicine, New York, New York. [3]Oberlin College, Oberlin, Ohio. [4]Broad Institute of MIT and Harvard, Cambridge, Massachusetts. [5]Harvard T.H. Chan School of Public Health, Boston, Massachusetts. [6]Massachusetts Institute of Technology, Cambridge, Massachusetts.

Note: Supplementary data for this article are available at Molecular Cancer Research Online (http://mcr.aacrjournals.org/).

Corresponding Authors: Renato Umeton, Department of Informatics and Analytics, Dana-Farber Cancer Institute, Boston, MA 02215. Phone: 617-632-3000; E-mail: renato_umeton@dfci.harvard.edu; and Massimo Loda, New York Presbyterian-Weill Cornell Campus, 1300 York Avenue, Room C-302 New York, NY 10065. E-mail: mloda@med.cornell.edu

Mol Cancer Res 2022;20:202–6

doi: 10.1158/1541-7786.MCR-21-0665

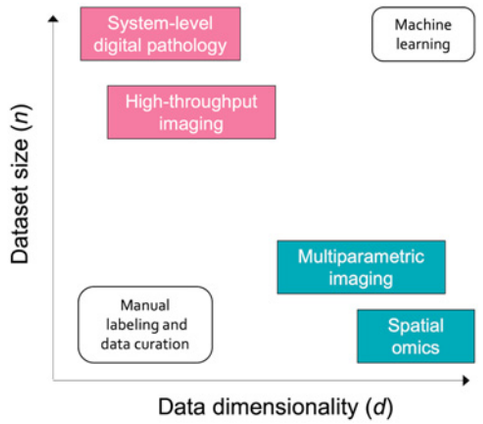

Figure 1.
"Big Data" in biomedical imaging scales along two orthogonal axes: dataset size (*n*), which captures the number of data points (i.e., cases) in each dataset, and data dimensionality (*d*), which refers to the amount of data captured in each data point.

researcher's toolkit, necessary for making use of massive datasets to study impactful questions in cancer biology and clinical care.

To enable this transition, software tools must lower the barrier to entry for computational image analysis, providing a bridge between the worlds of cancer research and machine learning. In this work, we discuss general features necessary for successful software tools, and present PathML: an open-source toolkit for computational pathology which we have built to address this outstanding need.

Guiding Principles for Building Software Tools to Accelerate Research

To effectively leverage the wealth of imaging data, researchers must be equipped with the tools to easily incorporate powerful computational image analysis methods into their research. We identified three key elements that should guide design and development of software tools in this domain: scalability, standardization, and ease of use.

Scalability

As datasets grow, analysis tools must be carefully designed to meet the technical challenges presented by scaling up in both *n* and *d*. Algorithms should be parallelized wherever possible, reducing computation time by running tasks concurrently. To enable efficient computation at the massive scale of tomorrow's datasets, tools should embrace distributed processing and provide easy integration with commonly used open source big data solutions for on-premise, cloud, and hybrid infrastructures (e.g., Kubernetes, Hadoop YARN, Slurm, etc.). Support for hardware accelerators such as graphics processing units and tensor processing units is a requirement for computationally intensive tasks such as training machine learning models. Finally, tools should enable users to work with data that are larger than available memory—an important feature for accommodating larger data and supporting exploratory research on consumer-grade computers.

Standardization

Another crucial consideration is standardization. No single tool can or should do everything; rather, by embracing standardized file formats, data structures, and application programming interfaces (API), individual tools can focus on specialized tasks while still providing cross-compatibility with other tools. For example, researchers may need to implement domain-specific algorithms for working with specific data types of interest [e.g., stain deconvolution for hematoxylin and eosin (H&E) images], but should interface with industry-standard machine learning frameworks [e.g., PyTorch (10) and TensorFlow (11)] rather than implementing basic machine learning functionalities from scratch. In addition to providing consistency for users, this approach also promotes emergence of a cohesive ecosystem of tools, such as those built around the AnnData (12) standard in single-cell omics.

Ease of use

A tool may be scalable and standardized, but it can only have an impact on accelerating research if users adopt it into their workflows. Therefore, software should be designed from the ground up with the intended audience in mind, and tools should be accessible with only minimal prior training in programming. This can be facilitated by building applications around well-defined APIs, which reduce the learning curve by providing consistency and by abstracting away some technical details from end users. All source code should be fully documented, with reproducible worked examples and detailed reference materials for all APIs. On the flip side of the coin, researchers will stand to benefit the most from advances in computational approaches if they are comfortable with the basics of coding in commonly used languages such as Python or R.

PathML: A Toolkit for Computational Pathology

We applied these guiding principles to development of PathML, an open-source toolkit designed for digital pathology research.

There are several existing tools that serve various needs in computational pathology. Some tools provide implementations of specific workflows or workflow components but are not designed as fully customizable, general-purpose libraries (13–17). HistomicsTK (18) offers a Python API for running aspects of analysis workflows, but is built around a specific data management platform (Girder) rather than being platform agnostic. QuPath (19) is an open-source tool for viewing and analyzing WSIs which has a scripting language for programmatic analysis but does not natively support Python. SquidPy (20) is primarily focused on spatial omics rather than general-purpose image analysis. There are also commercial tools available for digital pathology, some with support for machine learning analysis; however, these proprietary tools are not always ideal for researchers due to their cost and reduced flexibility in development relative to open source tools which enable full transparency into the underlying source code. Notably, there are no currently available open source tools which support the following requirements: the ability to load images from a wide array of file formats, including proprietary formats and standard formats such as TIFF and DICOM, under a common API; a standardized API for building custom preprocessing pipelines from modular components; support for running preprocessing at scale on commonly used high-performance computing solutions; integration with industry-standard machine learning frameworks in Python; and uniting analysis of brightfield and fluorescence images under a common framework.

To fill this unmet need, we developed PathML as a general-purpose toolkit for computational pathology, designed to be both highly performant and easy to use for researchers without requiring extensive training in programming or data science. PathML provides a general framework for creating and running preprocessing

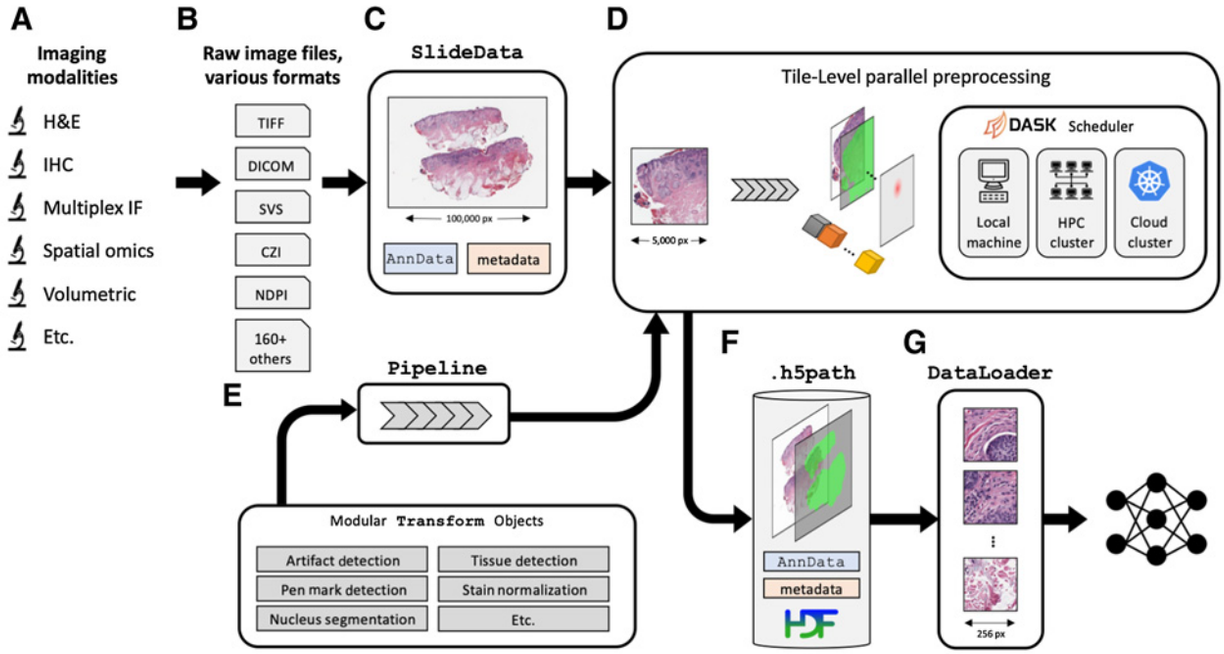

Figure 2.
Overview of the PathML preprocessing framework. **A,** A wide range of imaging platforms and modalities are supported via **B,** support for loading a comprehensive set of more than 165 file formats, including proprietary formats from vendors (see full list of supported file formats in Supplementary Table S1). **C,** Raw image files are loaded into SlideData objects, which encapsulate the image as well as associated metadata. **D,** To enable efficient processing of gigapixel-scale scans, images are divided into tiles and preprocessing pipelines are applied independently to each tile. Tiles can thus be processed in parallel, using the Dask scheduler to orchestrate distributed computation on large clusters, with support for both cloud and on-premise computing. Smaller images are processed in this framework as a single tile containing the entire image. **E,** A preprocessing pipeline is defined as a set of transformations applied sequentially. Transformations are modular, so can be mix-and-matched to rapidly build custom pipelines. **F,** Processed tiles are aggregated into an h5path file on disk, along with associated metadata such as labels, masks, and counts matrix. HDF5 is used to enable efficient slicing and indexing of the resulting file without needing to load the entire file into memory. **G,** DataLoaders from frameworks such as PyTorch then interact with the h5path file to efficiently feed images from the processed image into downstream machine learning models.

pipelines, unifying analysis of varying file formats (e.g., TIFF, DICOM, proprietary file formats from vendors, etc.), imaging modalities (e.g., H&E, IHC, Vectra Opal, CODEX, etc.), and dataset scales (from individual images to millions of images) under a single object-oriented API, with data structures and design choices specifically tailored to digital pathology (**Fig. 2**). The PathML library is written in Python 3 to promote ease of use and integration with the broader ecosystem of standard tools for data science and machine learning; however, we leverage libraries such as NumPy (21) and PyTorch which are written in low-level languages such as C, C++, and CUDA to handle computationally intensive operations more efficiently. An extensive suite of unit testing and integration testing helps minimize bugs and ensure that all code in PathML is working as expected. By developing PathML as an open source tool, we hope to build a community of users and collaborators to collectively accelerate the pace of innovation in digital pathology research.

The first step in a PathML workflow is loading the raw image file to create a SlideData object, which is the central data class representing an image and associated metadata. To accommodate the wide array of file formats commonly used in digital pathology, we provide three separate backends for reading image files, each supporting a complementary set of file formats (Supplementary Table S1). Each backend adheres to a standardized API, enabling users to manipulate images with a consistent interface regardless of file format or imaging modality (Supplementary Vignette S1).

The next step is to create a preprocessing pipeline, which we define as the sequential application of independent building blocks, or transformations. Each transformation applies a specific operation which may include modifying an input image, creating or modifying pixel-level metadata (i.e., masks), or creating or modifying image-level metadata (e.g., image quality metrics or an AnnData counts matrix). Transformations are general and flexible, providing a standardized interface to compose preprocessing pipelines. We provide in PathML a set of commonly used transformations, both domain-specific (e.g., H&E stain deconvolution, tissue detection, WSI artifact detection) and general-purpose (e.g., blurring, binary thresholding; Supplementary Fig. S1). Users may also implement custom transformations, and we provide an API to enable integration of custom transformations alongside prebuilt transformations. Multiple transformations can be composed into a single compound transformation. Transformations therefore provide the building blocks for formalizing the design and implementation of arbitrary preprocessing pipelines. This API allows researchers to write scalable, end-to-end preprocessing pipelines in only a few lines of code, using the same syntax and building blocks across different file formats and imaging modalities (Supplementary Vignettes S2 and S3).

One of the most common technical challenges in computational pathology is presented by extremely large file sizes, with high-resolution WSIs routinely exceeding the capacity of available memory. We therefore designed PathML based on a paradigm of independent processing of tiles. To run a preprocessing pipeline, subregions of the

image (i.e., tiles) are extracted and passed to the preprocessing pipeline independently. Smaller images are processed in this framework as a single tile containing the entire image. All processed tiles are then aggregated together into an on-disk array optimized for storing and manipulating large imaging datasets. This design allows for efficient preprocessing of large datasets of gigapixel images, as the data parallelism approach can efficiently scale up to make use of additional computational resources (e.g., cores in a multi-core computing unit, computing nodes in a cluster, etc.). We use the dask.distributed (22) scheduler on the backend, which allows for distributed preprocessing on many common high-performance computing platforms, including support for both on-premise and cloud computing environments. Importantly, tile extraction and distributed processing are handled automatically by PathML, enabling users to leverage these features to run analyses at scale with no change to the rest of their code. One limitation of this tile-centric approach is that artifacts may arise when tiles are aggregated back together, such as discontinuities at tile edges or "patchwork" effects. However, processing is inherently limited by the number of pixels that can be stored and manipulated in memory at once, so there is always a tradeoff between processing few low-resolution tiles and many high-resolution tiles. Users have complete control over tile extraction parameters, including the ability to use overlapping tiles which, in conjunction with stitching algorithms such as (23) can mitigate such artifacts.

As tiles are processed, they are aggregated together and written to disk. We define a file specification (h5path) which leverages the Hierarchical Data Format (HDF5) to enable efficient read/write access to regions of the processed image without loading the entire image into memory. Along with the processed images and masks, each h5path file contains associated slide-level and tile-level metadata. Each SlideData object is backed by a corresponding h5path file on disk, allowing for intuitive object-oriented workflows scalable to larger-than-memory images.

After a preprocessing pipeline has been run, we provide utilities to load the processed images into machine learning frameworks for downstream tasks (e.g., PyTorch DataLoaders). Preprocessing pipelines may themselves include transformations which encapsulate machine learning algorithms, for example using a model to perform nucleus detection and/or classification on each tile. PathML further provides PyTorch implementations of commonly used models such as U-Net (24) and HoVer-Net (25). Finally, we provide streamlined access to domain-specific datasets including PanNuke (26), PESO (27), and DeepFocus (28) for use in model training and benchmarking for various tasks. With support from open-source contributors, we hope that the inventory of available datasets and machine learning models will continue to expand.

In sum, PathML provides comprehensive support for each step in the computational pathology research workflow. We define a framework for preprocessing images and metadata which is streamlined and flexible for a wide variety of file formats and imaging modalities, implemented in an efficient, open source, fully tested and thoroughly documented Python package. We have already applied PathML to enable published (29) and currently ongoing computational pathology research at our institutions; by releasing it as an open source standard toolkit to bridge the gap between digital pathology and the broader machine learning and artificial intelligence (AI) ecosystem, we aim to lower the barrier to entry and accelerate progress in digital pathology research, thus benefiting the entire research community and moving one step closer to implementation of computational methods in the clinic.

Conclusion

With biomedical imaging datasets growing exponentially in both number of samples and dimensionality (i.e., data within each sample), machine learning is emerging as an increasingly essential tool for cancer researchers. To support these efforts, software tools must be designed with emphasis on scalability, standardization, and ease of use. Here we introduce PathML, a framework built with these best practices in mind that aims at lowering the barrier of entry to digital pathology, and show how a number of heterogeneous computational pathology use cases can be readily implemented in very few lines of code. With comprehensive support for all aspects of computational pathology research, from loading a wide variety of imaging modalities and file formats, to building modular and completely customizable preprocessing pipelines, to parallel-computing provisions, and integrations with other tools in the machine learning, AI, and single-cell analysis ecosystems, PathML can be employed to tackle a variety of biologically relevant problems. We anticipate that the real impactful part of this work will be around the applications of this technology, which we made open source and therefore available to all researchers. PathML is publicly available at www.pathml.com.

Authors' Disclosures

E.M. Van Allen reports personal fees from Tango Therapeutics, Genome Medical, Invitae, Monte Rosa Therapeutics, Manifold Bio, Illumina, Enara Bio, and Janssen; grants from Novartis and BMS outside the submitted work; in addition, E.M. Van Allen has a patent for institutional patents filed on Chromatin Mutations and Immunotherapy Response, and Methods for Clinical Interpretation pending and issued. The Editor-in-Chief of *Molecular Cancer Research* is an author on this article. In keeping with AACR editorial policy, a senior member of the *Molecular Cancer Research* editorial team managed the consideration process for this submission and independently rendered the final decision concerning acceptability. No disclosures were reported by the other authors.

Acknowledgments

The authors would like to thank Jason Johnson, Jerri Zhang, the Artificial Intelligence Operations and Data Science Services group, and many collaborators in the Department of Informatics and Analytics at Dana-Farber Cancer Institute and in the Department of Pathology and Laboratory Medicine at Weill Cornell Medicine for their continuous support and critical feedback that improved this work. We thank Angeles Duran, Jorge Moscat, and Maria T. Diaz-Meco for providing the images used in Supplementary Fig. S1, panels F and G. J. Nyman's work is supported by NIH F31 Predoctoral Individual National Research Service Award F31CA250136. M. Loda's work is supported by NCI P50CA211024, DoD PC160357, DoD PC180582, and the Prostate Cancer Foundation.

Received August 13, 2021; revised October 25, 2021; accepted December 1, 2021; published first December 8, 2021.

References

1. Schüffler PJ, Geneslaw L, Yarlagadda DVK, Hanna MG, Samboy J, Stamelos E, et al. Integrated digital pathology at scale: a solution for clinical diagnostics and cancer research at a large academic medical center. J Am Med Inform Assoc 2021;28:1874–84.
2. Lewis SM, Asselin-Labat ML, Nguyen Q, Berthelet J, Tan X, Wimmer VC, et al. Spatial omics and multiplexed imaging to explore cancer biology. Nat Methods 2021;18:997–1012.
3. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;25:1301–9.
4. Chatrian A, Colling RT, Browning L, Alham NK, Sirinukunwattana K, Malacrino S, et al. Artificial intelligence for advance requesting of immunohistochemistry in diagnostically uncertain prostate biopsies. Mod Pathol 2021;34:1780–94.

5. Lu MY, Chen TY, Williamson DFK, Zhao M, Shady M, Lipkova J, et al. AI-based pathology predicts origins for cancers of unknown primary. Nature 2021;594: 106–10.
6. Kather JN, Heij LR, Grabsch HI, Loeffler C, Echle A, Muti HS, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. Nat Cancer 2020;1:789–99.
7. Kather JN, Pearson AT, Halama N, Jäger D, Krause J, Loosen SH, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat Med 2019;25:1054–6.
8. Berry S, Giraldo NA, Green BF, Cottrell TR, Stein JE, Engle EL, et al. Analysis of multispectral imaging with the AstroPath platform informs efficacy of PD-1 blockade. Science 2021;372:eaba2609.
9. Schürch CM, Bhate SS, Barlow GL, Phillips DJ, Noti L, Zlobec I, et al. Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. Cell 2020;183:838.
10. Paszke A, et al. PyTorch: an imperative style, high-performance deep learning library; 2019.
11. Abadi M, et al. Tensorflow: a system for large-scale machine learning. In: Proceedings of 12th USENIX symposium on operating systems design and implementation (OSDI 16); 2016.
12. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 2018;19:15.
13. Lee S, et al. HistomicsML2. 0: fast interactive machine learning for whole slide imaging data; 2020.
14. Janowczyk A, Zuo R, Gilmore H, Feldman M, Madabhushi A. HistoQC: an open-source quality control tool for digital pathology slides. JCO Clin Cancer Inform 2019;3:1–7.
15. Byfield P. Peter554/StainTools, Zenodo; 2019.
16. Berman AG, et al. PathML: a unified framework for whole-slide image analysis with deep learning. medRxiv; 2021.
17. Jaume G, et al. HistoCartography: A toolkit for graph analytics in digital pathology. In: Proceedings of the MICCAI Workshop on Computational Pathology; 2021; PMLR.
18. Gutman DA, Khalilia M, Lee S, Nalisnik M, Mullen Z, Beezley J, et al. The digital slide archive: a software platform for management, integration, and analysis of histology for cancer research. Cancer Res 2017;77:e75–8.
19. Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, et al. QuPath: open source software for digital pathology image analysis. Sci Rep 2017;7:16878.
20. Palla G, et al. Squidpy: a scalable framework for spatial single cell analysis. bioRxiv; 2021.
21. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature 2020;585:357–62.
22. Rocklin M. Dask: parallel computation with blocked algorithms and task scheduling. In: Proceedings of the 14th python in science conference; 2015; Citeseer.
23. Preibisch S, Saalfeld S, Tomancak P. Globally optimal stitching of tiled 3D microscopic image acquisitions. Bioinformatics 2009;25:1463–5.
24. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer; 2015.
25. Graham S, Vu QD, Raza SEA, Azam A, Tsang YW, Kwak JT, et al. Hover-net: simultaneous segmentation and classification of nuclei in multi-tissue histology images. Med Image Anal 2019;58:101563.
26. Gamper J, et al. Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In: European Congress on Digital Pathology. Springer; 2019.
27. Bulten W, Bándi P, Hoven J, van de Loo R, Lotz J, Weiss N, et al. Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard. Sci Rep 2019;9:864.
28. Senaras C, Niazi MKK, Lozanski G, Gurcan MN. DeepFocus: detection of out-of-focus regions in whole slide digital images using deep learning. PLoS One 2018; 13:e0205387.
29. Linares JF, Zhang X, Martinez-Ordoñez A, Duran A, Kinoshita H, Kasashima H, et al. PKCλ/ι inhibition activates an ULK2-mediated interferon response to repress tumorigenesis. Mol Cell 2021;81:4509–26.