Computational and Structural Biotechnology Journal 27 (2025) 346–359

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Review article

Demystifying the black box: A survey on explainable artificial intelligence (XAI) in bioinformatics

Aishwarya Budhkar [a], Qianqian Song [b,*], Jing Su [c,*], Xuhong Zhang [a,*]

[a] Department of Computer Science, Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington, 700 N Woodlawn Ave, Bloomington, IN 47408, USA
[b] Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, 1889 Museum Rd, Suite 7000, Gainesville, FL 32611, USA
[c] Department of Biostatistics and Health Data Science, School of Medicine, Indiana University, Indianapolis, HITS 3000 BSAT, Indianapolis, IN 46202, USA

ARTICLE INFO

Keywords:
Explainable AI (XAI)
Bioinformatics
Omics
Biomedical imaging

ABSTRACT

The widespread adoption of Artificial Intelligence (AI) and machine learning (ML) tools across various domains has showcased their remarkable capabilities and performance. Black-box AI models raise concerns about decision transparency and user confidence. Therefore, explainable AI (XAI) and explainability techniques have rapidly emerged in recent years. This paper aims to review existing works on explainability techniques in bioinformatics, with a particular focus on omics and imaging. We seek to analyze the growing demand for XAI in bioinformatics, identify current XAI approaches, and highlight their limitations. Our survey emphasizes the specific needs of both bioinformatics applications and users when developing XAI methods and we particularly focus on omics and imaging data. Our analysis reveals a significant demand for XAI in bioinformatics, driven by the need for transparency and user confidence in decision-making processes. At the end of the survey, we provided practical guidelines for system developers.

1. Introduction

In the biomedical field, particularly in omics and imaging, AI has proven to be highly effective in a wide range of applications, such as gene expression analysis [1], protein structure prediction [2], disease diagnosis through medical imaging [3], personalized treatment planning [4,5], and integrative multi-omics analysis [6]. The analysis of omics data aims to understand the molecular mechanisms and microenvironment of the diseases, and provide targeted treatments [7,8]. However, the high-dimensional nature of omics data makes visual analysis challenging, necessitating the use of complex deep learning models to extract insights. Techniques such as autoencoder models help reduce data dimensionality, allowing the learning of low-dimensional feature representations to study different biological phenomena [9]. In addition to making predictions based on omics, researchers are interested in uncovering underlying biological mechanisms and processes for knowledge discovery [10]. However, the black-box nature of complex models, such as models based on neural networks with multiple layers, makes it challenging to understand the reasons behind the model decisions. This lack of interpretability raises concerns about trust, accountability, and security, limiting the application of deep learning models in bioinformatics area [8,11].

Several AI models have been developed to gain novel insights from biological datasets [7]. These models are used to uncover novel molecular pathways or biomarkers, which often requires additional references from the biology literature, molecular databases, pathway analysis tools, or inputs from physicians, clinicians, biologists, and other professionals. However, in bioinformatics applications, it is essential to consider the needs of various stakeholders involved, including AI experts, biologists, and bioinformaticians [7,12,13]. AI experts aim to develop models that minimize prediction errors and ensure high precision, while bioinformaticians and biologists require detailed explanations of the underlying rationales [7,12,14,15]. Therefore, the diverse needs of the bioinformatics field must be given special consideration in the development of AI models. Employing XAI techniques to elucidate model decisions can help understand model limitations and ensure satisfaction of various user and application needs.

There are some common open problems in the field of XAI for the biomedical field. First, there is a lack of consensus on the definitions of interpretability, transparency, and explainability in the literature, leading to confusion and hindering clear communication in the field [14]. Second, different domains have different interpretability

* Corresponding authors.
E-mail addresses: qsong1@ufl.edu (Q. Song), su1@iu.edu (J. Su), zhangxuh@iu.edu (X. Zhang).

https://doi.org/10.1016/j.csbj.2024.12.027
Received 13 November 2024; Received in revised form 21 December 2024; Accepted 23 December 2024
Available online 10 January 2025

Table A.1
Summary of differences of existing works from our work.

Paper	Summary
Gerlings, *et al.* (2020) [23]	The article conducts a systematic survey analyzing the major debates within XAI and advocates for a holistic and stakeholder-driven approach for XAI. The work highlights the need of XAI for diverse AI applications without any specific focus. Inspired by their work, we provide application and stakeholder-driven categorization of methods, and present guidelines to choose suitable XAI methods with a focus on bioinformatics applications.
Tjoa, *et al.* (2020) [14]	The survey provides a broad categorization of XAI methods for the medical field. Our work has a narrow focus on bioinformatics to develop categorization considering the unique needs of bioinformatics.
Talukder, *et al.* (2020) [20]	The paper summarizes the DNN interpretation methods in recent studies on genomics and epigenomics, focusing on current data and computing-intensive topics such as sequence motif identification, genetic variations, gene expression, chromatin interactions and non-coding RNAs.
Mohseni, *et al.* (2021) [12]	The survey provides a hierarchical framework for the design and evaluation of XAI systems. Similarly, our work aims to provide a granular view of XAI system design considering the distinctive needs of bioinformatics applications. Our work also presents detailed guidelines for selecting interpretable models.
Antoniadi, *et al.* (2021) [7]	The paper provides a comprehensive review of existing XAI systems in Clinical Decision Support Systems (CDSS) along with their benefits and limitations. The focus of the work is explainable ML-based CDSS. Our categorization of the XAI works is done in consideration of the unique needs of bioinformatics field.
Markus, *et al.* (2021) [24]	The work provides general guidelines for users to choose XAI methods and evaluation metrics. Unlike ours, it is directed towards a broader audience lacking consideration of specific requirements of bioinformatics applications and stakeholders.
Sidak, *et al.* (2022) [10]	The paper presents a detailed review of interpretability methods for omics data. In contrast to ours, the review is targeted towards biologists and not system developers. Our survey focuses on challenges for XAI techniques from software and algorithmic perspective while accounting for unique needs of bioinformatics.
Toussaint et al. (2024) [19]	The paper provides insights into the isolated advances in the omics sciences. Our survey also includes how XAI advances in bioimaging field.
Novakovsky et al. (2023) [25]	In this survey, the authors present an organized overview of the key interpretation approaches with the intention of empowering researchers working across topics in genomics to incorporate xAI into their studies, and their focus on the task of post-hoc interpretation. Our review contains more broader topics in biomedical field and examined different XAI models, including post-hoc ones.
Karim et al. (2023) [26]	This review provides a brief overview of interpretable ML tools and libraries for diverse data types (e.g. from tabular data, texts and knowledge graphs (KGs) to bioimaging). It also shows, through several case studies in bioimaging, cancer genomics and biomedical text mining, how bioinformatics research can benefit from XAI methods and improve decision fairness. Our survey is more focused on different data formats and types like multi-omics.
Zhou, *et al.* (2023) [27]	The paper provides a review of some model-agnostic and model-specific methods used in different bioinformatics areas. Our survey extends their review by including additional XAI methods and works in bioinformatics. In addition, we provide guidelines for the efficient choice of XAI methods highlighting the need to consider system goals and collaboration between different stakeholders.

requirements, and current work on model development and evaluation criteria often fails to accommodate these differences [12]. Third, users may have varying levels of understanding about the system and may require explanations tailored to their specific needs and knowledge levels [12]. But still, it is both possible and valuable to establish working definitions of these terms for the purpose of this review, thereby promoting clearer communication and a more cohesive understanding within the field. In addition, we provide general guidelines for choosing XAI algorithms and evaluation methods tailored to address the diverse needs of end users, and develop algorithm categorization considering the unique requirements of users and applications.

In this survey, we reviewed existing XAI methods in various subdomains, such as genomic, proteomics, particularly transcriptomics, and pathology imaging data, providing guidelines for system developers, ensuring a broad coverage of the needs of various stakeholders, including system developers with ML expertise, biologists, and other professionals. The motivation is to offer a foundational framework that developers can adapt to their particular biomedical niche. Here omics refers to high-throughput technologies and data-driven approaches aimed at comprehensively characterizing biological molecules within cells, tissues, or organisms. Among these, transcriptomics focuses on the complete set of RNA transcripts (the transcriptome) present at a given time, often measured through next-generation sequencing methods such as RNA-seq [16,17], providing insights into gene expression patterns and regulatory mechanisms. In parallel, imaging data such as medical images derived from modalities like MRI, CT, or microscopy capture structural or functional characteristics of tissues and organs, enabling detailed visualization of biological processes and disease states. Together these data types form complementary pillars of biomedical research: omics data reveal molecular-level complexity, while imaging offers anatomical and physiological perspectives, making them both prime candidates for the application of explainable AI techniques.

Although this review is more focused on transcriptomics and imaging, there are many related studies exploring other layers of biological data. For instance, epigenomics research delves into modifications that influence gene regulation without altering the DNA sequence itself, while metabolomics examines the small molecules and metabolites that reflect cellular processes. Additionally, multi-omics integration brings together these various data types-encompassing transcriptomics, epigenomics, metabolomics, etc. and their integration to provide a more holistic view of complex biological systems [18–22]. The differences between our work and the existing review articles, which motivated us for this survey, are summarized in Table A.1, and additionally, we organized the published frameworks at: https://github.com/asbudhkar/XAI-in-Bioinformatics/tree/main.

2. Terminology in explainable AI

XAI encompasses a wide range of methods aimed at building trust in model decisions. Terms such as transparency, interpretability, and explainability are often used interchangeably [28], but there is no widely accepted distinction yet. Among these terms, explainability and interpretability are more often used than others. Antoniadi et al. [7] state that interpretability refers to “how much a model is understood”, transparency provides a “holistic view of how the model works, its training data, methods, and feature explanations”, and explainability helps stakeholders “understand the reasoning behind AI decisions”. Adadi [29] stated that interpretable systems are explainable if their operations can be understood by humans, which shows explainability is closely related to the concept of interpretability. Tjoa et al. [14] used the terms “interpretability” and “explainability” interchangeably, considering research related to explainability if it attempts to “explain model decisions”, “explain the model’s workings”, or “enhance user trust in the model”. Markus et al. [24] provided a formal definition stating that an “AI system is explainable if the model is inherently interpretable” or “an interpretable model is provided additionally to explain the model outcomes”. Gilpin et al. [30] stated that interpretability and fidelity are both necessary components for explainability. They argued that a good explanation should be understandable to humans (interpretability) and accurately describe model behavior in the entire feature space (fidelity) [24]. From this sense, interpretability and fidelity are deemed necessary for explainability, with an explanation being interpretable faithful if it is unambiguous and not overly complex and faithful and if it is correct and

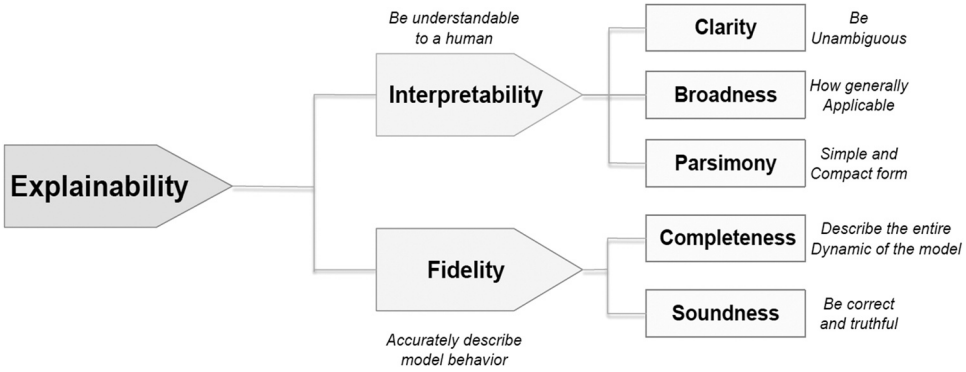

Fig. B.1. Definition of machine learning (ML) explainability and related properties. Concepts adopted from [24]: a good explanation should be understandable to humans (interpretability) and accurately describe model behavior in the entire feature space (fidelity). Interpretability can have properties of clarity and parsimony. Clarity implies that the explanation is unambiguous, while parsimony means that the explanation is presented in a simple and compact form. Fidelity has properties of completeness and soundness, and completeness implies that the explanation describes the entire dynamic of the ML model, while soundness concerns how correct and truthful the explanation is.

sufficient to compute output from input. In this work, we use the terms explainable, interpretable, and transparent interchangeably, considering a model to be explainable if any attempt is made to provide insights into how the model arrived at its decision. Fig. B.1 demonstrates the related concepts for explainability and interpretability.

Another concept is trustworthy AI, which is closely related to XAI, yet they address different aspects of how artificial intelligence systems are developed, evaluated, and accepted by users: XAI focuses primarily on making the inner workings of AI models more transparent, interpretable, and understandable to humans. Trustworthy AI, on the other hand, encompasses a broader set of attributes that ensure the AI system's overall reliability and ethical alignment. While explainability may be one part of it, trustworthiness also includes dimensions such as fairness, safety, privacy, accountability, robustness, and compliance with regulations, according to the EU regulations on AI usage (Regulation (EU) 2024/1689 or AI Act [31]).

2.1. Key dimensions of XAI methodologies

The algorithms are categorized from three different perspectives: post-hoc vs. ante hoc; model agnostic or specific, and their interpretability level.

2.1.1. Ante hoc versus post hoc

In the context of XAI, ante-hoc and post-hoc methods represent two distinct approaches to achieve interpretability. Ante-hoc methods involve building models that are inherently transparent and interpretable from the start, thus reducing the need for separate explanatory tools. In contrast, post-hoc approaches are applied after a non-transparent model (such as deep neural network) has been trained. These explanations typically use techniques like feature importance measures, surrogate models, or counter-factual examples to clarify how the model arrives at its predictions. While ante-hoc models offer interpretability by design, they are not explanations in themselves; rather, their makes it easier for human users to understand and reason about the model's behavior without relying on additional explanation methods.

Ante hoc explanation models are inherently explainable, meaning they are simple enough to understand and able to model relationships between input and output data. Examples of these models include linear regression, decision tree models, k-nearest neighbors, and rule-based learners, which are also referred to as white box models [7,15]. A trained neural network can be complemented with an interpretable model or analyzed post-training to gain insights into its decisions through post hoc explanation [32]. For instance, Local Interpretable Model-Agnostic Explanations (LIME) [33] provides feature-level explanations by training an interpretable model to approximate the behavior of a complex model. Gradient-based methods like Grad-CAM [34] estimate feature importance by analyzing the gradients of a trained neural network to provide explanations [7,15]. In [21], SHAP (SHapley Additive exPlanations) explainers are used to calculate feature contributions for each prediction on genomics and epigenomics data for insulin resistance early diagnosis.

2.1.2. Model agnostic versus model specific

Model-agnostic explanations are independent of the type of ML model used and can be applied to any ML algorithm [7,15]. They operate by understanding the change in input that influence the output to identify the regions or features in the input that most significantly impact the model's decision. These explanations do not impose strict requirements on the model's structure or design. In contrast, model-specific explanations can be applied to only a set of ML algorithms due to their reliance on structure or attributes of the model. For example, Grad-CAM, which requires a neural network with differentiable layers to provide explanations [7,15].

2.1.3. Global versus local

Global explanation offers a comprehensive understanding of a model's overall behavior, providing insights at the dataset level [7,15]. Local explanation focuses on the reasoning behind the prediction of a single instance in the dataset, providing instance-level explanations. For example, a local explanation might clarify why a loan application was rejected citing specific reasons for the individual applicant, such as low income or being a defaulter.

3. Study selection

For this review, we opted for a broad literature survey for works published after 2017, rather than conducting a fully systematic review. We consulted a diverse range of sources encompassing both computer science and bioinformatics, including arXiv, bioRxiv, Nature, Briefings in Bioinformatics, the IEEE Digital Library, the ACM Digital Library, ScienceDirect, Google Scholar, Frontiers, and MDPI. In particular, arXiv and bioRxiv facilitated timely access to emerging research prior to formal publication. We employed a variety of search terms—such as "explainable AI", "interpretable", "transparent", "bioinformatics", "pathology imaging", "histopathology", "genomics", "transcriptomics", "proteomics", "omics" and "XAI" to identify relevant studies. Articles were initially screened based on their titles and abstracts, and those deemed relevant were then examined in full. Selection criteria included thematic relevance, methodological rigor, clarity, and the use of appropriate interpretability techniques. We also incorporated additional references identified through the citations of key papers. Ultimately, this

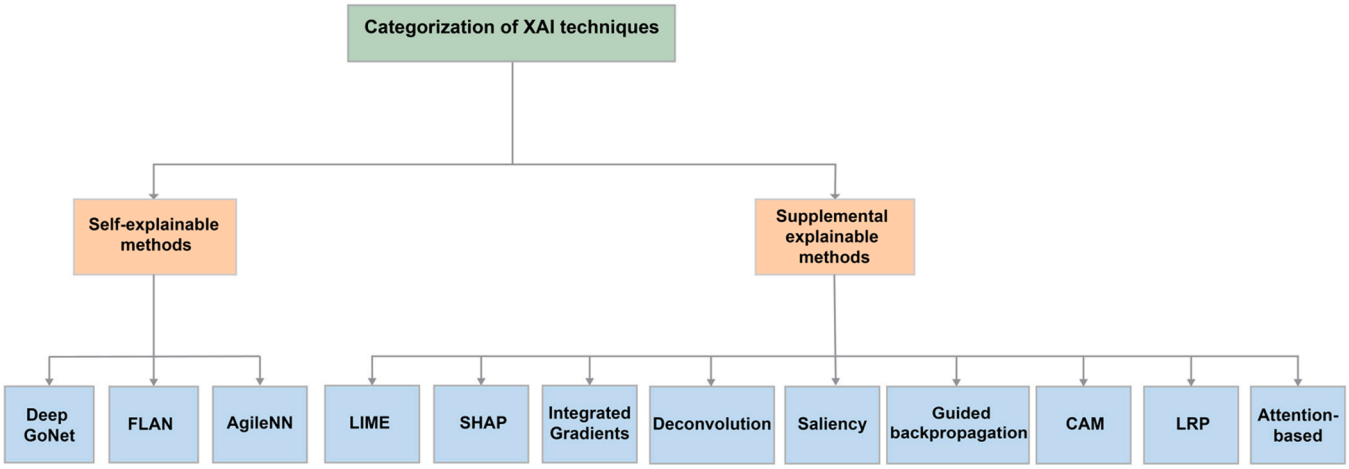

Fig. B.2. A conceptual overview categorizing XAI models into self-explainable models and supplemental explainable models. Self-explainable models offer inherent transparency and do not require additional explanation techniques. In contrast, supplemental explainable models depend on external methods after training such as LIME, SHAP, or saliency maps to reveal their decision-making processes. Representative examples within each category are listed to illustrate this fundamental distinction in model interpretability.

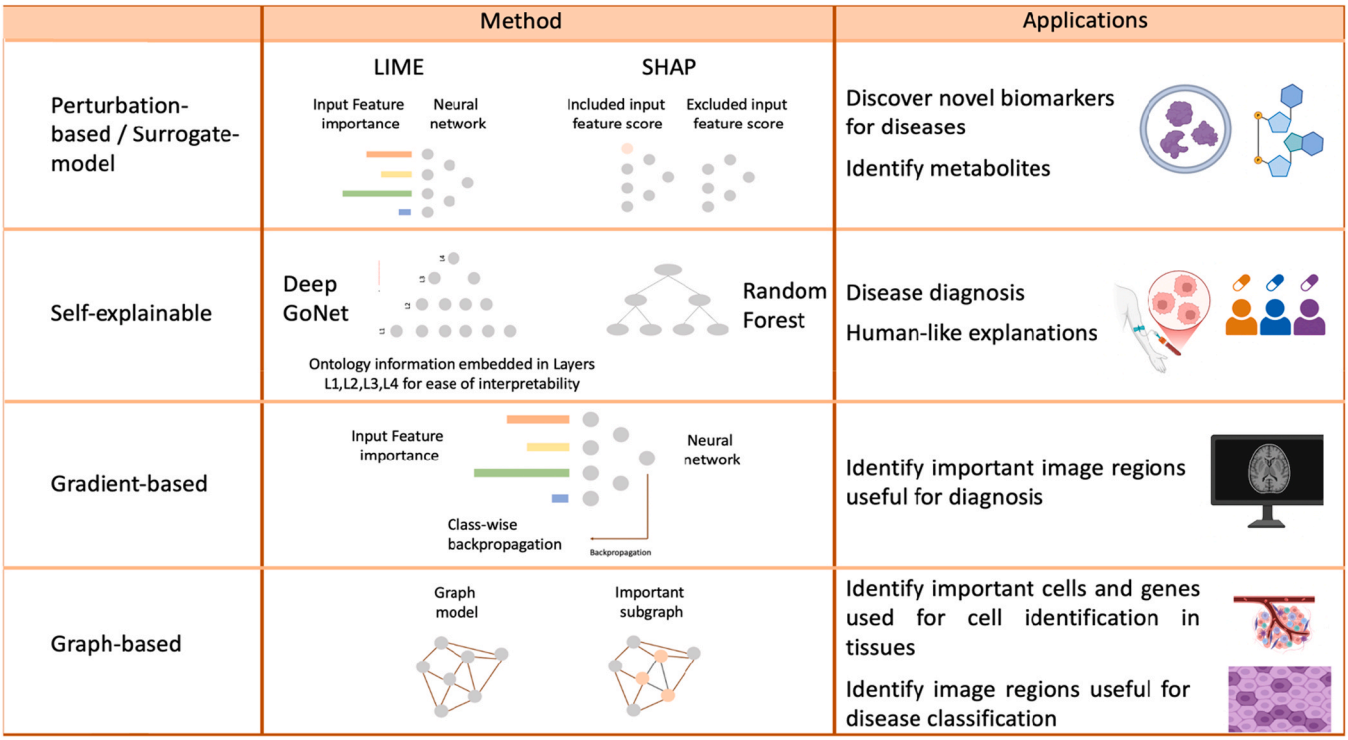

Fig. B.3. A visual summary of common XAI techniques and their applications within the bioinformatics domain. The figure highlights a range of methods and demonstrates how they can be applied to different types of biological data and analytical tasks.

process yielded a set of 55 core articles on explainable methods in bioinformatics, supplemented by further literature to address our specific research questions, with topics cover imaging (33), transcriptomics (14), proteomics (3), metabolomics (1) and genomics (4); post-hoc (48), ante-hoc (7), global (14), local (45), model specific (38), model agnostic (17).

4. Categorization of XAI models in bioinformatics

Based on the review of XAI models, we categorize the works into self-explainable and supplemental explainable models which can help researchers select the most suitable XAI model for their research. Here self-explainable models refer to inherently transparent or interpretable models, where the reasoning behind their predictions is clear from the model's structure and parameters. The working of a self-explainable model is easy to understand with well-defined model design, mathematical formulae, constraints, or hypotheses. On the other hand, a supplemental explainable model are non-transparent models that do not naturally reveal how their outputs are generated. For complex models whose functioning is difficult to explain, a supplemental explainable model can be provided which uses a comparatively simpler model to explain the complex model's decisions. It can be done by providing local explanations which are less complex and easy to understand or by using complementary measures to elucidate the decision-making of the complex model. As a result, these complex models rely on additional

Table A.2
Comparison of supplemental XAI methods.

Method	Complexity and efficiency	Applicability	Limitations
Saliency, Simonyan *et al.* (2013)[57]	Computationally efficient since it involves gradient computations	Efficient performance for imaging applications	Can be prone to noisy gradients. Requires in-depth evaluations to ensure faithful explanations
Deconvolution, Zeiler and Fergus *et al.* (2014)[55]	Computationally challenging for deep networks	Traditionally used in signal processing	Demonstrated high performance on images but may lead to noisy explanations due to approximations involved
Guided backpropagation, Springenberg *et al.* (2014)[60]	Computationally efficient since it involves gradient computations	Efficient performance for imaging applications	Can lead to noisy explanations. Requires in-depth evaluations to ensure faithful explanations
LIME, Ribeiro *et al.* (2016)[33]	Relatively efficient due to local interpretability. Complex input perturbations can make it inefficient for certain problems	Model agnostic, simple, easy to implement	Wrong assumptions about data and underlying model may lead to poor performance. Choice of surrogate model and neighborhood function if incorrect may lead to wrong explanations
Guided Grad-CAM, Selvaraju *et al.* (2016)[66]	Computationally efficient since it involves gradient computations	Provides detailed high-resolution and class contrastive explanations compared to Grad-CAM	May lead to noisy results. Requires in-depth evaluations to ensure faithful explanations
LRP, Binder *et al.* (2016)[37]	Computationally expensive for deep neural networks	Provides detailed feature attributions at each layer in network to explain how input features contribute to model's output	Computationally expensive and unsuitable for models with several layers
CAM, Zhou *et al.* (2016)[63]	Computationally expensive for complex networks since several linear models need to be trained	Efficient performance for imaging applications. Provides class contrastive explanation	Model specific and applicable to a certain set of model architectures
SHAP, Lundberg and Lee (2017) [46]	Computationally expensive due to the requirement of multiple subset computations for feature importance. Approximation measures may be required for efficient processing for high-dimensional data	Model agnostic. Provides both local and global explanations. Captures interactions among different features	Computationally inefficient for large datasets. Approximation measures to speed up performance may not be applicable to all datasets
Integrated gradients, Sundararajan *et al.* (2017)[51]	Computationally moderate since it performs integral along a direct path	Capture interactions among features. Applicable to a large set of models	Performance is influenced by baseline chosen and may lead to incorrect explanations for complex models due to approximations
Grad-CAM, Selvaraju *et al.* (2017)[34]	Computationally efficient since it involves gradient computations	Provides fine-grained visual explanations. Improved localization accuracy compared to CAM. Faster than CAM due to reliance on gradients	May provide low-resolution explanations. Requires in-depth evaluations to ensure faithful explanations
Attention-based methods, Saihood (2023)[78], Raghavan (2023) [79]	Complexity depends on the employed attention-mechanism	Utilize attention-mechanism to improve explanations for certain applications	Applicable to a limited set of models having attention mechanisms. Complex to understand due to the varied attention mechanisms used

explanation techniques applied after training to shed light on their decision-making processes. Fig. B.2 illustrates the categorization scheme.

The categorization into self-explainable and supplemental explainable models serves as a conceptual overlay to the key dimensions of XAI discussed previously (e.g., Ante Hoc vs. Post Hoc, Model Agnostic vs. Model Specific, and Global vs. Local). Self-explainable models generally align with ante-hoc approaches, as their transparency is built directly into the model's structure, often making them more globally interpretable. In contrast, supplemental explainable models typically require post-hoc techniques that can be either model-agnostic or model-specific, and may provide either global or local insights.

Note that since our work is focused on bioinformatics, there are foundational efforts in XAI that have been explored in much broader contexts. Historically, the field of XAI has seen a surge in methods and frameworks designed to make complex models more interpretable, from computer vision and natural language processing to high-stakes domains like healthcare and finance. A more comprehensive overview of the broader landscape of XAI can be found in seminal works such as [32,35]. We also summarize the different XAI techniques and their applications in bioinformatics in Fig. B.3.

4.1. *Self-explainable models*

These models are designed to ensure transparent explanations for end-users using methods such as mathematical formulas, hypotheses about inputs or data modeling, or constraints that make decisions easy to interpret. For example, linear models, decision trees, and rule-based systems are inherently interpretable [7]. However, these models may not always deliver the desired performance, prompting developers to use more complex models to achieve better outcomes.

Some models integrate interpretability directly into their design or leverage domain knowledge to enhance interpretability. For instance, Deep GoNet [36], a self-interpretable model incorporates ontology information into its layers to simplify interpretation. This model, demonstrated for cancer diagnosis, provides explanations at the disease, subdisease, and patient levels. Each neuron is associated with a biological concept and explanations are generated using the Layer-wise

Relevance Propagation (LRP) technique [37] and domain-specific biological knowledge. FLAN [38] is a structurally constrained deep neural network model designed to explain the relevance of input features to its decisions. Like general additive models, FLAN processes each input feature separately generating distinct latent representations, which are then summed to provide the model output. By predicting on individual features, their importance to the decision can be computed similarly to linear models. The authors have demonstrated that FLAN achieves high performance in several biological tasks while maintaining self-interpretability. Patrício et al. [39] proposed a self-interpretable model for diagnosing skin lesions which provides both natural language explanations and visual explanations. The model uses concept vectors and segmentation masks along with a coherent loss to capture different concepts of skin lesions. Detected features are encoded by the GloVe model [40] to generate vector representations that facilitate human-like explanations. The authors have shown that the model's performance is comparable to existing black-box approaches, however, it can be challenging to obtain concept annotations for its broader application.

4.2. Supplemental explainable model

Complex algorithms are often favored for their superior performance over self-explainable models. However, they often lack transparency in their decision-making process due to their intricate nature, making it challenging to explain their behavior using formulae or system constraints. To address this, post-hoc techniques are employed, which utilize simpler interpretable models to explain the behavior of the complex ones. These techniques can approximate the model's performance providing local explanations with reduced complexity or quantify the impact of different inputs on the output. However, it is crucial to strike a balance between interpretability and performance. Table A.2 compares the commonly used supplemental models.

4.2.1. LIME

LIME (Local Interpretable Model-Agnostic Explanations) [33] offers local explanations for complex models by approximating their predictions using interpretable models. Let f be the complex model to be explained, and $g \in G$ be a potentially interpretable model, such as linear or decision tree models. The complexity of the interpretable model's explanation is denoted by $\Omega(g)$, which could be, for example, the depth for decision trees or the number of non-zero weights for linear models. The first step is to select an instance x and then consider a proximity measure π_x to compute the similarity between x and its n neighbors. Then a local interpretable model is trained to minimize the loss based on locality denoted as $L(f,g,\pi_x)$. The interpretable model's coefficients are utilized to explain the decisions made by the complex model. LIME generates the explanations using the following equation:

$$\xi(x) = argmin_{g \epsilon G} L(f, g, \pi_x) + \Omega(g) \tag{1}$$

For bioinformatics studies, Yagin et al. [41] used LIME to identify biomarkers associated with COVID-19 to provide valuable information for clinicians for combating the disease. Similarly, Yilmaz et al. [42] used LIME to explore metabolites for the identification of acute myocardial infarction using metabolomics data. LIME is used to identify important genes for the prognosis and treatment of bladder cancer using gene expression data [43]. Park et al. [44] utilized LIME to highlight key genomic factors in predicting drug responses based on cancer gene expression and mutation maps data.

The advantages of this model include its model-agnostic nature and its ability to explain decisions for different modalities, such as text and image. Slack et al. [45] demonstrated several drawbacks of the method including sensitivity to the choice of proximity measure, high computational demands due to input perturbation, and vulnerability to adversarial attacks. The memory usage and speed of LIME depends on factors like the number of neighbors, the chosen proximity measure and the interpretable model used.

4.2.2. SHAP

SHAP (SHapley Additive exPlanations) [46] provides both local and global model-agnostic explanations by calculating the contribution of each feature, known as Shapley values. These values are computed by evaluating the difference in model outputs with and without each feature included, across all potential subsets. For a complex model with a prediction function $f(x)$ and total features T, the Shapley values can be obtained using the following equation:

$$\varnothing_i = \sum_{S \subseteq T \setminus \{i\}} \frac{|S|!(|T| - |S| - 1)!}{|T|!} \left(f_{S \cup i}\left(x_{S \cup \{i\}}\right) - f_s(x_s)\right) \tag{2}$$

where S represents any subset which does not include the i^{th} feature. Computing Shapley values exactly is challenging due to the vast number of feature subsets involved, leading to the development of approximation methods. These methods often assume feature independence and linearity to simplify computation. Global explanations are obtained by averaging the Shapley values in all instances.

Ramírez-Mena et al. [47] applied SHAP to elucidate predictions of a random forest model for prostate cancer detection. SHAP provides global explanations to identify crucial genes for patient screening and local explanations for personalized treatments. Sobhan et al. [48] used variants of SHAP, including gradient SHAP and tree explainer SHAP to develop novel therapies for lung cancer by identifying biomarker genes for individual patients. Additionally, the XAI-CNVMarker [49] framework leveraged SHAP to discover potential biomarkers for breast cancer, which aids in drug development. Dwivedi et al. [50] employed gradient SHAP to identify relevant biomarkers for non-small cell lung cancer that aid in the development of targeted therapy.

SHAP has several advantages such as easy integration with any model and the ability to provide both local and global explanations. However, it has been shown to be vulnerable to adversarial attacks [45]. The speed and memory requirement of SHAP depends on the complexity of the model, the approximation methods used, and the number of features. Exact computation can be resource intensive potentially requiring exponential time $O(2^T)$ for the computation and storage of multiple subsets.

4.2.3. Integrated gradients

Integrated Gradients (IG) offer both global and local explanations by computing gradients along a path from a chosen baseline to the original input. It is based on two axioms, namely sensitivity and implementation irrelevance. Based on [51], considering a complex model function as f with input x and reference x', a straight-line path is assumed between x' and x, and gradients are computed along the path. The integrated gradient along dimension i is:

$$IG_i(x) = \left(x_i - x_i'\right) \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \quad d\alpha \tag{3}$$

The integral is approximated with a summation:

$$IG_i^{approx}(x) = \left(x_i - x_i'\right) \sum_{k=1}^{m} \frac{\partial F\left(x' + \quad \frac{k}{m} \times (x - x')\right)}{\partial x_i} \times \frac{1}{m} \tag{4}$$

where m represents the number of steps for the approximation, typically ranging between 20 and 300. Gao et al. [52] used IG to identify important amino acids for the prediction of transcription factors in gene regulation. Li et al. [13] utilized IG to attribute importance to genomic features for cancer diagnosis using multimodal data such as histopathology images and genomic sequences. Dwivedi et al. [50] employed IG to identify relevant biomarkers for non-small cell lung cancer identification to aid for development of targeted treatments. Li et al. [53] used IG to reveal importance of radiological features used by their model for

Table A.3
Comparison of CAM-based methods.

Method	Applicability	Comparison with CAM
iCAM, Cai *et al.* (2019)[64]	Uses activation maps of CNN layers at both high and low level. Generates fine-grained class discriminatory explanations by using top and bottom layers	Improved region localization without performance degradation. Computationally expensive due to the requirement of learning weights at multiple layers.
BroadCAM, Lin *et al.* (2023)[65]	Broad learning system is utilized to improve performance on small-scale data	Robust for small-scale data. Computationally expensive due to overhead of additional using broad learning system.
Grad-CAM, Selvaraju *et al.* (2017)[34]	Generalization of CAM. Instead of training linear models to learn class discriminative feature map weights, gradients are used	Computationally efficient. Applicable to a wider set of models.
Guided Grad-CAM, Selvaraju *et al.* (2016)[66]	Improvement over Grad-CAM. Utilizes guided backpropagation to improve explanation quality	Provides class discriminative as well as detailed explanations.

survival prediction of cancer patients.

IG can explain any differentiable model without specific architecture requirements, but the choice of reference baseline can influence feature attributions. Although it can be applied to any model with differentiable functions, its reliability has been questioned in certain contexts [54]. The speed depends on the number of integration steps, the number of dimensions, and the complexity of explanation generation. Memory requirement is influenced by both the input and reference dimensions.

4.2.4. Deconvolution

Zeiler et al. [55] introduced deconvolutional networks to visualize feature activity in intermediate layers by reversing convolutional network operations. This method approximates inverse pooling operations and applies transposed convolutional filters to understand the workings of each layer. By reconstructing original inputs, important regions are revealed. Gao et al. [56] used deconvolution in signal processing to visualize learned features for decoding brain activity, yielding relevant explanations. The explanations were validated using known brain region knowledge. The speed and memory requirements of deconvolution method depends on the complex model used and how much temporary data needs to be stored.

4.2.5. Saliency

Simonyan et al. [57] proposed Saliency method that visualizes Convolutional Neural Network (CNN) models by using gradients to highlight input pixel relevance to the predicted class. For example, de Souza Jr et al. [58] used the saliency method to emphasize relevant image regions for esophagus cancer classification using trained CNN models. Comparison with annotations from different XAI techniques demonstrated promising results, with saliency performing the best and closely matching expert annotations.

Saliency is easy to compute but it can produce noisy or vanishing gradients with deep networks. Its speed depends on the size of complex model, requiring only a single backpropagation pass for gradient computation. The requirement to store feature maps dominates memory usage of the method. Shrikumar et al. [59] demonstrated that multiplying gradients with the input improved saliency results.

4.2.6. Guided backpropagation

Springenberg et al. [60] presented guided backpropagation method which improves deconvolution results by backpropagating only positive gradients, resulting in sharper saliency maps compared to deconvolution. Badea et al. [61] adopted it to visualize histology features with significant gene correlations, offering biological interpretability. Similarly, Wickstrøm et al. [62] used it for colorectal polyp segmentation, enhancing trust among users and facilitating comparison of different models.

While it provides detailed explanations, guided backpropagation may suffer from vanishing gradients, requiring careful evaluation for accuracy. Its speed is constrained by forward and backward passes through the network, and its memory requirement depends on the storage of feature maps and intermediate results.

4.2.7. CAM

Class Activation Mapping (CAM) [63] is a model-specific technique that identifies important parts of an image for a particular class. CAM utilizes the global average pooling layer (GAP) after the last convolutional layer in CNN models to generate explanations. For a feature map k in the last convolutional layer, denoted $f_k(x,y)$ for any image at spatial location x,y, the global average pooling is performed as follows:

$$F_k = \sum_{x,y} f_k(x,y) \tag{5}$$

For a class c, the input to softmax function is given as

$$\sum_k w_k^c F_k \tag{6}$$

where w_k^c indicates the importance of F_k for class c. Therefore, the important regions for a particular class in the image are represented as a linear combination of weighted feature maps. CAM calculates the global average of each feature map, resulting in n scalars if the last convolutional layer has n feature maps. Subsequently, a linear model is learned for each class. The combination of different weights learned for different classes is then applied to weight the pixels in the image, generating saliency maps that highlight the spatial regions of the image used by the model to make class predictions. Following CAM, there are several variant versions proposed, including iCAM [64], Broad-CAM [65], Grad-CAM [34], and Guided Grad-CAM [66].

Civit-Masot et al. [67] provided human-like explanations by generating a report containing relevant image region generated by Grad-CAM along with the model's confidence in classification, to enhance trust in the model for both doctors and patients. The authors also included accuracy information along with the top few features used by the model for prediction and their relevance scores computed using SHAP. Yin et al. [68] demonstrated that Grad-CAM highlights essential genes for survival prediction. Shovon et al. [69] used Grad-CAM to reveal the image regions that the trained neural network model relies on for breast cancer classification using histology images. Altini et al. [70] utilized Grad-CAM along with Mask-RCNN for instance segmentation, enhancing nuclei detection by distinguishing individual nuclei instances. The authors claim that by integrating Grad-CAM with Mask-RCNN they achieved state-of-the-art performance in nuclei detection. Kallipolitis et al. [71] used Guided Grad-CAM to explain ensemble models predictions by highlighting the key regions of histopathology images used by the model for cancer classification. Similarly, Korbar et al. [72] used Guided Grad-CAM to assign contribution scores to image pixels, highlighting areas that the trained ResNet model used to detect different colorectal polyps in histopathology images. Grad-CAM and Guided Grad-CAM are popular for elucidating imaging models, and attention-based models offer detailed visual explanations [68,73,74]. Cai et al. [64] introduced improved CAM (iCAM), a variant of CAM, to identify critical regions in images for predicting muscular dystrophy. iCAM combines activation maps from both high and low-level CNN layers leveraging the semantic information from top and bottom layers to generate fine-grained class discriminatory explanations. Lin et al. [65] developed Broad-CAM to enhance performance on small-scale data using broad learning system (BLS) demonstrating its effectiveness in breast cancer image classification. Several studies have used diverse datasets together for XAI techniques to elucidate model operations. For

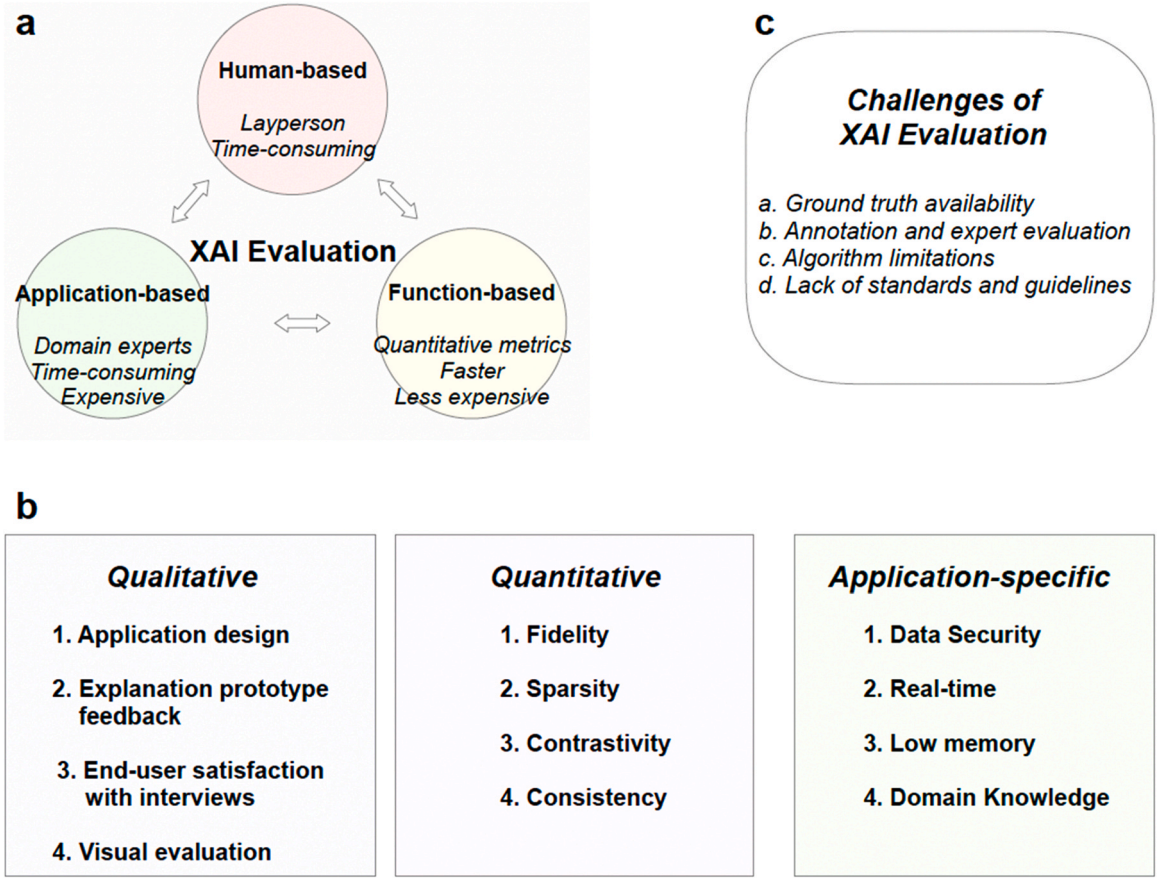

Fig. B.4. Overview of evaluation techniques for XAI and the associated challenges. The subFig. a illustrates various evaluation strategies—ranging from human-centered assessments, application-based and function-based. SubFig. b shows factors need to be consider for qualitative, quantitative and application-specific evaluations. Subfigure c shows some common challenges. Together, these elements underscore the need for careful, context-dependent evaluation frameworks that ensure XAI methods meet their intended goals in practice.

example, Li et al. [13] performed cancer diagnosis and prognosis using both gene expression and image data. Similarly, Kayser et al. [75] used Grad-CAM and Integrated Gradients to reveal important genes and provide visual explanations simultaneously. The authors used radiology reports along with imaging data to provide human-like explanations [75]. Li et al. [53] used Grad-CAM to reveal the significant regions in histological images used by the model for survival prediction of cancer patients. Aryal et al. [76] used whole slide images and graph modeling for cancer classification. They used Graph Grad-CAM to highlight important image areas for classification, showing that these areas closely matched pathologists' annotations.

The disadvantages of using CAM include its dependency on specific architectures with a global average pooling layer, as well as the overhead of training linear models over the GAP outputs to generate explanations. Furthermore, CAM may not provide reliable explanations with weakly labeled data due to unstable training from insufficient data, as noted by Lin et al. [65]. The speed is primarily determined by the forward pass and the computations required to learn linear models for generating saliency maps. Storage of feature maps from the last convolutional layers and any intermediate outputs dominate the memory usage. See Table A.3 for more details.

4.2.8. *Layer-wise relevance propagation*

Layer-wise relevance propagation (LRP) [37] is a model-agnostic technique that explains the contribution of input features to the output by assigning a relevance score for each feature through decomposition techniques. It propagates relevance scores from the output layer back to the input layer, layer by layer. While several versions of the algorithm exist, the general strategy of LRP is explained in Binder et al. [37] as follows: Consider a complex model function f with multiple neural network layers. Each neuron in layer l contributes to the activation of neuron j in layer $l+1$. The algorithm ensures that the total relevance is conserved in each layer, assuming that there is a known relevance in the output layer $l+1$. The relevance score R of each neuron is propagated backward through the network maintaining the total relevance across all layers. Therefore,

$$\sum_i R^{i\to j}{}_{l,l+1} = \quad R^j_{l+1} \tag{7}$$

The relevance of I is calculated using the β rule as follows:

$$R^{i\to j}{}_{l,l+1} = \left((1+\beta)\frac{z^+_{ij}}{z^+_j} - \beta\frac{z^-_{ij}}{z^-_j}\right)R^j_{l+1} \tag{8}$$

where $z_{ij} = (w_{ij}x_i)^p$ such that p is the pixel and output of neuron j is given as $x_j = g\left(\sum_i w_{ij} + b\right)$ with g is non-linear activation function and b is bias term. Here, z^+_{ij} and z^-_{ij} denote the positive and negative contribution of neuron I to the activation of neuron j. The relevance score is conserved from layer $l+1$ to l by dividing the individual contributions by total contribution of all nodes. B is used to control the relevance distribution so that a larger β will provide sharper heatmaps. LRP can be applied to any differentiable model, but the initialization of relevance at the output layer will affect the interpretation of relevance scores.

Bourgeais et al. [36] used LRP to compare the explanations of their Deep Go-Net model. Springenberg et al. [77] combined LRP with attention maps to highlight important image areas for cancer

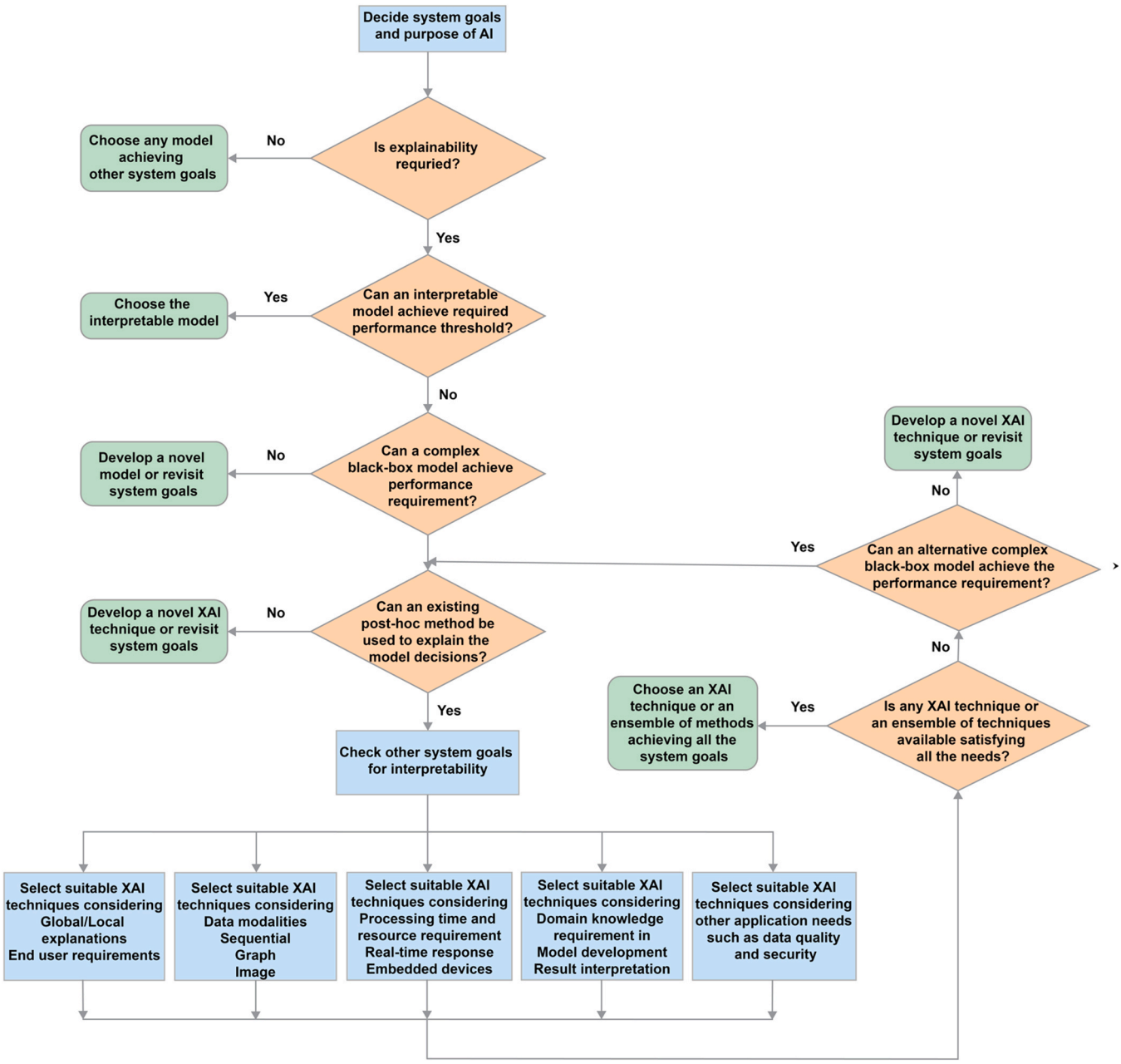

Fig. B.5. A decision-making flow chart for implementing XAI solutions in practice. The process consists multiple steps guide practitioners through selecting the appropriate XAI techniques—such as choosing between self-explainable models or post-hoc explanatory methods—based on data characteristics, desired interpretability, and evaluation criteria. The chart culminates in an iterative refinement stage, ensuring that chosen XAI strategies align with stakeholder needs, ethical considerations, and real-world constraints.

classification. The speed for LRP depends on the network parameters, including the number of layers and operations performed at each layer. The size of complex model and the parameters to be stored at each layer for relevance propagation need to be considered to calculate the memory usage of the method.

4.2.9. Attention-based

Recently, attention-based methods have been employed to enhance model transparency. A multi-orientation-based guided attention module (MOGAM) framework is proposed by Saihood et al. [78]. The framework aims to leverage the varying texture distribution in lung nodules from CT scan images for lung cancer classification. MOGAM fuses several texture feature descriptors (TFDs) such as contrast, dissimilarity, and homogeneity and inputs using CNN models to generate attention maps. Saihood et al. [78] demonstrated enhanced interpretability on CT scan images using Grad-CAM and Guided Grad-CAM.

Raghavan et al. [74] presented attention-guided Grad-CAM for identifying important regions in infrared breast cancer images. Channel and spatial attention are computed to assign weights to channels and image regions, respectively, using CNN features. These attentions are combined to focus on critical regions, resulting in a more robust and interpretable heatmap. Causal attention is utilized by Chen et al. [73] for classifying lymphoma ultrasound images, employing both channel and spatial attention aiding model generalization. The authors also used a counterfactual explanation method using image perturbation for interpretability. While attention-based methods offer promising advancements for interpretability, their choice, effectiveness, and complexity depend on the dataset and problem domain.

Table A.2 provides a comparison among different supplemental XAI methods.

5. Opportunities and challenges of XAI in bioinformatics

5.1. Resource-constrained environments

Resource constraints, such as storage, and computational time, are crucial factors in selecting XAI techniques. Some techniques are resource-intensive, making them impractical for embedded devices with limited memory. Research has focused on developing suitable solutions for embedded and mobile devices with limitations on space and computation and that require real-time responses. For example, Chen et al. [80] presented an AI system for detecting monkeypox from smartphone-captured skin images utilizing CNN models VGG-16 [81] and MobileNetV3 [82], trained on ImageNet [83]. By employing quantization techniques to reduce operation precision from 32-bit to 16-bit floats, MobileNetV3 optimizes model performance reducing its size by 4 $\times$. The best-performing model uses Grad-CAM to identify class-discriminative features providing responses within seconds that are suitable for mobile devices. Another study [84] presents a light-weight CNN model consisting of only three layers for retinal disease diagnosis, delivering comparable performance to complex models with several layers. Predictions are explained using visualization maps generated by gradient activations. The authors claim that their model provides accurate predictions for embedded devices, mobile systems, and IoT devices. AgileNN [85] designed for real-time performance on embedded devices employ light-weight neural networks and utilize backpropagation-based techniques, which require a single pass to compute input feature importance, resulting in real-time decisions. Experiments on popular datasets demonstrated a 6 $\times$ reduction in output latency. Self-interpretable models, that integrate explanations directly into their design, can be computationally expensive depending on the model's complexity [86]. Techniques that use interpretable models to approximate the workings of complex models through different approximation techniques and input perturbations may also require substantial computational resources. Backpropagation-based techniques, which require a single pass to compute input feature importance, are comparatively faster [15]. Example-based techniques, which rely on clustering and similarity measures, have resource requirements that vary based on their specific implementation.

5.2. Data challenges and security in XAI

When applying XAI techniques in bioinformatics, proper data preprocessing and rigorous quality control procedures are essential for ensuring that subsequent interpretations and insights are both reliable and meaningful. Factors, such as normalization, outlier handling, and batch effect removal can significantly influence the outcomes of model training and the generation of explanations. Equally important is the recognition that different types of biological data be it gene expression profiles, epigenetic marks, metabolite concentrations, or imaging data, each have their own characteristics and noise patterns, necessitating tailored preprocessing strategies. For example, epigenomic and metabolomic technologies are highly sensitive to batch effects and outliers, necessitating careful normalization. In the context of genomics, particularly for complex diseases, relying on single nucleotide polymorphisms may not sufficiently predict outcomes, prompting the use of polygenic risk scores as a more suitable approach. Furthermore, the inherent heterogeneity among both healthy individuals and within diseased populations can complicate the identification of meaningful patterns. By carefully considering the nature of the data and applying appropriate quality control measures, researchers can better ensure that the explanations produced by XAI methods faithfully represent the underlying biological processes rather than technical artifacts [87–89].

Weakly labeled or unlabeled data pose significant challenges when applying existing XAI methods. Data preprocessing may be necessary to address these challenges. For instance, BroadCAM [65] enhances performance with weakly labeled data using a broad learning system. The authors utilize a broad learning system to improve CAM performance since CAM's performance suffer from unstable training with weakly-labeled data.

To comply with AI regulations and safeguard user data privacy and security, it is crucial to moderate the techniques used. Perturbation-based explainability techniques, such as LIME and SHAP, can conceal biases in the training data [45]. Adversarial classifiers can successfully hide biases by manipulating input perturbations during training. To address these concerns and identify disparities in explanations metrics like fidelity, stability, sparsity, and consistency are introduced [54]. Fidelity assesses explanation accuracy, stability measures consistency across similar data, sparsity gauges explanation complexity, and consistency evaluates repeatability of explanations.

5.3. Recent advances in generative AI for XAI

Generative large language models are making significant strides across various fields due to their superior performance on a range of tasks [86]. ThyGPT [86], a computer-aided diagnosis model for thyroid nodules, leverages annotated ultrasound images and diagnostic reports to generate human-interpretable diagnosis explanations. However, obtaining annotated multimodal datasets can be challenging. CodonBERT [90] predicts gene expressions using codon patterns. It uses fine-tuned pre-trained BERT models and SHAP to explain predictions, thereby elucidating the model's function. HistoGPT [91] generates human-like histopathology reports from whole slide images, enabling user interaction. The model explains its decisions using saliency maps, indicating image regions that lead to specific findings. Keyword overlap and semantic similarity metrics are used for evaluation. While this research found limited examples of generative AI models explaining bioinformatics works, they hold great potential for advancing explainable solutions in the field and represent a promising avenue for future research.

6. The limitations of current XAI techniques in bioinformatics

Recent evaluations have uncovered limitations of existing XAI techniques through experiments. Graziani et al. [92] assessed the explanations of CAM, Grad-CAM, GradCAM+ +, and LIME on breast tissue datasets. They found that the explanations generated by XAI algorithms were similar for both a trained CNN model and a randomly initialized one. Particularly, LIME provided inconsistent and non-repeatable explanations across different hyperparameters. They suggested the development of quantitative evaluation methods instead of relying solely on visualizations. Rudin et al. [93] advocate for the use of self-interpretable models, citing issues like incompleteness and low accuracy of post-hoc explanations, their complexity and mismatch between model design and domain knowledge. For instance, FLAN [38], a constrained neural network model is introduced for ease of interpretation, demonstrated good performance on various datasets containing gene expression and imaging data. Chanda et al. [94] presented an interpretable model which utilizes dermoscopic images to predict ontology-based concepts using ResNet and Grad-CAM. Patrício et al. [39] proposed a self-explanatory model involving segmentation and feature extraction, trained to predict concept-based explanations using ground truth annotations. Similarly, Kayser et al. [75] developed an interpretable model that uses annotations from radiology reports capable of explaining predictions using radiology concepts.

To achieve optimal performance and reliable explanations, researchers advocate for developing algorithms tailored for the unique needs of applications and datasets. Current XAI methods often fail to provide satisfactory explanations due to their lack of consideration of the specific requirements of different bioinformatics data [27]. Sidak

et al. [10] state that the classification of different XAI techniques is complex due to vague terminology and suggest that customization of techniques for different scientific problems may be necessary due to the varied requirements of the system and the characteristics of the data. The review by Li et al. [95] demonstrates how XAI models developed specifically considering the graph structure for GNN models provide better explanations for some applications, emphasizing the importance of considering the model architecture in developing better explainability techniques. Furthermore, Patrício et al. [9] mention that the application requirements dictate the choice of algorithms and explainability techniques. They highlight the limitation of Recurrent Neural Network (RNN) models in generating long reports, which has led to the adoption of transformer-based models for textual explanations.

Studies have emphasized the need to investigate data for biases or misrepresentations and to ensure there are no security loopholes compromising user privacy. For example, Tjoa et al. [14] stress the importance of developing robust algorithms, as explanations can be manipulated by input changes or noise, potentially leading to erroneous decisions and interpretation. Markus et al. [24] recommend publicly reporting data quality to provide transparency about system limitations, and developing standards and regulations to safeguard user information and ensure safe data usage. The authors address challenges associated with medical data, such as expensive and time-consuming collection, highly imbalanced and noisy data sets and note that labeling of data is expensive due to the need for experts and conflicting opinions among them. Thus, it is imperative to develop XAI algorithms that consider both the characteristics and limitations of the data.

In a more general perspective, counterfactual and pro-hoc explanations provide two complementary avenues for enhancing the effectiveness of XAI [96]. Counterfactual explanations focus on illustrating how changing certain input features might alter the model's outcome, thereby illuminating decision boundaries and making it easier to identify and address underlying biases. Pro-hoc explanations, on the other hand, offer a set of alternative explanations for each possible outcome, rather than a single, post-hoc rationale. By leveraging example-based reasoning, this pro-hoc approach fosters what has been termed "frictional AI", a deliberate introduction of interpretive friction that encourages users to consider multiple plausible scenarios. In doing so, it can mitigate cognitive pitfalls like automation bias and over-reliance, ultimately promoting a more thoughtful engagement with AI recommendations. Together, counterfactual and pro-hoc explanations not only deepen our understanding of the relationships between inputs and outputs but also contribute to more trustworthy, user-aligned, and context-sensitive decision support systems [97].

Current algorithms may require modification to accommodate application resource limitations. Real-time systems, for example, require quick responses, while some mobile operating systems have very limited memory and processing capabilities, restricting them to run only resource-constrained applications. Thus, XAI algorithms should be developed with these computational constraints in mind. It is suggested that resource-intensive explanation methods should be used carefully, and alternative techniques should be considered if the application cannot afford high computational capabilities [98]. The use of Large AI (LAI) models to process large datasets has increased, showing significant performance improvements for several tasks [99]. While they demonstrate considerable potential for applications in several bioinformatics use cases but require further development to address challenges like limited interpretability and high computational demands. Therefore, the development of XAI algorithms that align with application and stakeholder requirements, consider unique data characteristics, and incorporate domain knowledge for model design or result validation represents promising future directions for further progress in the bioinformatics field.

We also notice that it would be valuable to include a reflection on the potential misuse of XAI techniques, acknowledging that even well-intentioned attempts to make AI decisions more interpretable can be misapplied or misunderstood. For example, post-hoc XAI methods may rely on simplified surrogate models or heuristic explanations that do not fully capture the underlying complexity of the decision process, potentially introducing inaccuracies or misleading conclusions. In such scenarios, stakeholders could be misled into believing they fully understand the model's reasoning, when in fact the provided explanations might only scratch the surface. To mitigate these risks, one strategy is to consider self-explainable models i.e., models inherently designed with transparency in mind so that interpretability is not an afterthought but a fundamental characteristic of the approach. This shift can reduce the danger of oversimplified explanations and offer more robust, faithful insights into the model's true decision-making criteria. Ultimately, careful consideration of when and how to apply XAI, along with ongoing dialog about its limitations, is critical to ensuring that interpretability efforts do not inadvertently compromise the integrity or trustworthiness of AI-driven decisions [100].

7. Validation and evaluation for XAI algorithms in bioinformatics

Several evaluation metrics for XAI have been proposed in the literature; however, no standard evaluation measures have yet been universally adopted across applications [12,15,24].

7.1. Human-based evaluation

Humans are involved in analyzing the quality of the explanation. Humans are hired to provide generic evaluation of explanations through crowdsourcing [14,15]. Although involving lay people is cost effective compared to expert evaluations, the evaluation of explanations may be less reliable due to their limited knowledge [15]. It can also be time-consuming due to the dependence on human evaluators [24]. Mental model, user satisfaction, user trust and human-AI task performance are some of the measures to evaluate explanations for AI novices [12].

Experts with domain-specific experience are involved in analyzing the quality of explanations. For instance, doctors with significant experience in the field are hired to assess the quality of AI-provided explanations [14,15]. Using expert evaluations is costly and time-consuming but more trustworthy than laypersons owing to the application specific experience of experts [15]. Although it can be more reliable than non-experts, it is not feasible for all scenarios due to resource-constraints.

7.2. Function-based evaluation

In function-based evaluation, proxy models are used to measure the quality of the explanation [14] with existing annotations are used to evaluate the explanations [15]. Correctness, completeness, and consistency of explanations are computational measures for evaluating XAI techniques [12]. Fidelity, contrastivity, stability, consistency, sparsity are some existing functional evaluation metrics [54,95]. The evaluation of XAI algorithms for bioinformatics applications involve several approaches. One common method is to validate explanations using the known literature on gene biomarkers for specific diseases or using pathway discovery tools together with the existing literature on the identified pathways [36,41–43,47–49,68,101–103]. However, validation of novel biomarker discoveries that are not widely recognized is often lacking [68]. Although some studies provide visual explanations by highlighting relevant regions of pathological images used for model predictions, these works lack quantitative measures [53,69,71,73,78,104–106]. Domain experts are involved in the validation process in several studies like [107–109]. Their expertise can help ensure the accuracy and relevance of the explanations provided by XAI algorithms. Some studies have introduced quantitative measures for evaluating explanations. For example, Lombardi et al. [110] involve experts at

different stages of model development, providing accuracy information and top features used for prediction based on user feedback. Krzyziński et al. introduced the Brier score [111,112] was introduced as a measure to assess the quality of the explanations. Korbar et al. [72] used the intersection over union between important regions of the image provided by the XAI algorithm and the reference regions as an evaluation measure, although such references may not be available in all datasets. The most relevant first curve (MoRF) and Area over Perturbation Curve (AOPC) were used as quantitative evaluation measures by Kallipolitis et al. [79]. The authors assumed that randomly distorting the important regions should lead to a greater decrease in classification performance compared to non-important regions for design of evaluation measures. Macedo et al. [113] developed a novel metric for the evaluation of explanations to measure the tumor nuclei present in the important regions identified by Grad-CAM for the classification of breast cancer. The authors emphasized that the choice of evaluation metric is largely dependent on the problem requirements and suggested that users prioritize models with higher values for both classification accuracy and explanation evaluation metrics. Civit-Masot et al. [67] provide a report consisting of classification accuracy along with the pathology image, with highlighted regions identified by the algorithm as most important for prediction to aid doctors in explaining the results to patients. Springenberg et al. [77] used correlation with segmentation masks as a quantitative metric while Lamy et al. [108] used quantitative similarity between images as the evaluation metric.

Moreover, while lot of applications have focused on applications and evaluations primarily in supervised learning contexts, unsupervised scenarios which are integral to many biomedical research endeavors, pose distinct challenges for XAI evaluation, such as clustering, pattern recognition and association rules [114–116]. Without ground-truth labels, conventional metrics such as accuracy or F1-score are not directly applicable. Instead, evaluation often relies on human-centered methods, where domain experts assess the utility and interpretability of explanations, as well as concept-based approaches that examine whether inferred clusters or latent factors align with known biomedical concepts or structures. Surrogate models and prototype-based explanations can be used to approximate the underlying decision boundaries, enabling researchers to assess factors like fidelity and stability. Ultimately, evaluating XAI in unsupervised learning settings requires a more context-specific blend of expert input, conceptual relevance, and model fidelity to ensure that explanations are both meaningful and credible.

The choice of evaluation methods largely depends on the specific requirements of the application. Researchers commonly advocate for the development of standards for the validation of XAI explanations to ensure a comprehensive evaluation. Moreover, ground truth availability largely limits the choice of evaluation methods since annotations are not always available in datasets and involvement of experts to provide annotations can be expensive. Additionally, the AI model used for the application and the XAI technique employed to explain the model functioning also play a crucial role in determining the appropriate evaluation metric. Fig. B.4 provides an overview of evaluation techniques for XAI applications and the associated challenges for ensuring holistic evaluation. Through our analysis we recognize that understanding the purpose of the application, system and privacy constrains, and end-user requirements is critical to choose appropriate XAI model. We propose a flowchart for developers to facilitate the choice of XAI techniques in Fig. B.5.

8. Conclusion

In this work, we reviewed existing explainability works to gain deeper insights into XAI techniques in bioinformatics, with a specific focus on omics and imaging data. We discussed the current XAI methods used in bioinformatics along with their limitations and evaluation techniques from the perspective of distinct needs of the bioinformatics field. Through the review, we discovered it is crucial to consider the needs of the application and its users when developing XAI techniques. It is also important to use domain knowledge to validate explanations. Since there are no clear guidelines for choosing the appropriate XAI method, based on the review we propose guidelines that can aid users to select the appropriate XAI method.

Funding

J.S. and A.B. were supported by the National Library of Medicine of the National Institutes of Health (R01LM013771). J.S. was also supported by the National Institute on Alcohol Abuse and Alcoholism (R21AA031370 and U24AA026969), the National Institute of Health Office of the Director (OT2OD031919), the Indiana University Melvin and Bren Simon Comprehensive Cancer Center Support Grant from the National Cancer Institute (P30CA 082709), and the Indiana University Precision Health Initiative. Q.S. is supported by the National Institute of General Medical Sciences of the National Institutes of Health (R35GM151089).

CRediT authorship contribution statement

Jing Su: Writing – review & editing, Validation, Conceptualization. **Qianqian Song:** Writing – review & editing, Validation, Conceptualization. **Xuhong Zhang:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Aishwarya Budhkar:** Writing – original draft, Methodology, Conceptualization.

Declaration of Competing Interest

No competing interest is declared.

Acknowledgments

Not applicable.

Author Agreement Statement

We declare that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed.

We further confirm that the order of authors listed in the manuscript has been approved by all of us. We understand that the Corresponding Author is the sole contact for the Editorial process. He/she is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs.

References

[1] Misra BB, Langefeld C, Olivier M, Cox LA. Integrated omics: tools, advances and future approaches. J Mol Endocrinol 2019;62(1):R21–45.
[2] Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature 2020; 577(7792):706–10.
[3] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.
[4] Ahmed Z. Multi-omics strategies for personalized and predictive medicine: past, current, and future translational opportunities. Emerg Top life Sci 2022;6(2): 215–25.
[5] Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018;2(10):719–31.
[6] Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. Bioinforma Biol Insights 2020;14. 1177932219899051.
[7] Antoniadi AM, Du Y, Guendouz Y, Wei L, Mazo C, Becker BA, et al. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. Appl Sci 2021;11(11). Available from: ⟨https://mdpi-res.com/d_attachment/applsci/applsci-11-05088/article_deploy/applsci-11-05088.pdf?version= 1622439587⟩.

[8] Hartman E, Scott AM, Karlsson C, Mohanty T, Vaara ST, Linder A, et al. Interpreting biologically informed neural networks for enhanced proteomic biomarker discovery and pathway analysis. Nat Commun 2023;14(1):5359.
[9] Patrício C, Neves JC, Teixeira LF. Explainable deep learning methods in medical image classification: a survey. ACM Comput Surv 2023;56(4). Article 85.
[10] Sidak D, Schwarzerová J, Weckwerth W, Waldherr S. Interpretable machine learning methods for predictions in systems biology from omics data. Front Mol Biosci 2022;9:926623.
[11] Janizek JD, Dincer AB, Celik S, Chen H, Chen W, Naxerova K, et al. Uncovering expression signatures of synergistic drug responses via ensembles of explainable machine-learning models. Nat Biomed Eng 2023;7(6):811–29.
[12] Mohseni S, Zarei N, Ragan ED. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Trans Inter Intell Syst 2021;11(3–4). Article 24.
[13] Li L, Pan H, Liang Y, Shao M, Xie S, Lu S, et al. PMFN-SSL: Self-supervised learning-based progressive multimodal fusion network for cancer diagnosis and prognosis. Knowl-Based Syst 2024:111502.
[14] Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learn Syst 2021;32(11):4793–813.
[15] van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal 2022;79:102470.
[16] Deshpande D, Chhugani K, Chang Y, Karlsberg A, Loeffler C, Zhang J, et al. RNA-seq data science: from raw data to effective interpretation. Front Genet 2023;14:997383.
[17] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol 2016;17:1–19.
[18] Fang X, Miao R, Wei J, Wu H, Tian J. Advances in multi-omics study of biomarkers of glycolipid metabolism disorder. Comput Struct Biotechnol J 2022;20:5935–51.
[19] Toussaint PA, Leiser F, Thiebes S, Schlesner M, Brors B, Sunyaev A. Explainable artificial intelligence for omics data: a systematic mapping study. Brief Bioinforma 2024;25(1):bbad453.
[20] Talukder A, Barham C, Li X, Hu H. Interpretation of deep learning in genomics and epigenomics. Brief Bioinforma 2021;22(3):bbaa177.
[21] Torres-Martos Á, Anguita-Ruiz A, Bustos-Aibar M, Ramírez-Mena A, Arteaga M, Bueno G, et al. Multiomics and eXplainable artificial intelligence for decision support in insulin resistance early diagnosis: a pediatric population-based longitudinal study. Artif Intell Med 2024;156:102962.
[22] Picard M, Scott-Boyer M-P, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J 2021;19:3735–46.
[23] Gerlings J., Shollo A., Constantiou I. Reviewing the need for explainable artificial intelligence (xAI). arXiv preprint arXiv:201201007. 2020.
[24] Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform 2021;113:103655.
[25] Novakovsky G, Dexter N, Libbrecht MW, Wasserman WW, Mostafavi S. Obtaining genetics insights from deep learning via explainable artificial intelligence. Nat Rev Genet 2023;24(2):125–37.
[26] Karim MR, Islam T, Shajalal M, Beyan O, Lange C, Cochez M, et al. Explainable ai for bioinformatics: methods, tools and applications. Brief Bioinforma 2023;24(5):bbad236.
[27] Zhou Z., Hu M., Salcedo M., Gravel N., Yeung W., Venkat A., et al. XAI meets Biology: a comprehensive review of explainable AI in bioinformatics applications. arXiv preprint arXiv:231206082. 2023.
[28] Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artif Intell 2019;267:1–38.
[29] Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 2018;6:52138–60.
[30] Gilpin L.H., Bau D., Yuan B.Z., Bajwa A., Specter M., Kagal L., editors. Explaining explanations: An overview of interpretability of machine learning. In: Proceedings of the 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA); 2018: IEEE.
[31] Artificial Intelligence Act: MEPs adopt landmark law. 2024 [Available from: ⟨https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law⟩.
[32] Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 2020;58:82–115.
[33] Ribeiro M.T., Singh S., Guestrin C. "Why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco, California, USA: Association for Computing Machinery; 2016. p. 1135–1144.
[34] Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D., editors. Grad-CAM: visual explanations from deep networks via gradient-based localization. 2017 In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2017 22-29 Oct. 2017.
[35] Longo L, Brcic M, Cabitza F, Choi J, Confalonieri R, Del Ser J, et al. Explainable Artificial Intelligence (XAI) 2.0: a manifesto of open challenges and interdisciplinary research directions. Inf Fusion 2024;106:102301.
[36] Bourgeais V, Zehraoui F, Ben Hamdoune M, Hanczar B. Deep GONet: self-explainable deep neural network based on gene ontology for phenotype prediction from gene expression data. BMC Bioinforma 2021;22(10):1–25.
[37] Binder A., Montavon G., Lapuschkin S., Müller K.-R., Samek W., editors. Layer-wise relevance propagation for neural networks with local renormalization layers. Artificial neural networks and machine learning–ICANN 2016: In: Proceedings of the 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25; 2016: Springer.
[38] Nguyen A-P, Vasilaki S, Martínez MR. FLAN: feature-wise latent additive neural models for biological applications. Brief Bioinforma 2023;24(3):bbad056.
[39] Patrício C., Neves J.C., Teixeira L.F., editors. Coherent concept-based explanations in medical image and its application to skin lesion diagnosis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023.
[40] Pennington J., Socher R., Manning C.D., editors. Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014.
[41] Yagin, Cicek FH, İB, Alkhateeb A, Yagin B, Colak C, et al. Explainable artificial intelligence model for identifying COVID-19 gene biomarkers. Comput Biol Med 2023;154:106619.
[42] Yilmaz R. Investigation of potential biomarkers in prediction of acute myocardial infarction via explainable artificial intelligence. Med Sci 2023;12(2).
[43] Kırboğa KK. Bladder cancer gene expression prediction with explainable algorithms. Neural Comput Appl 2024;36(4):1585–97.
[44] Park A, Lee Y, Nam S. A performance evaluation of drug response prediction models for individual drugs. Sci Rep 2023;13(1):11911.
[45] Slack D., Hilgard S., Jia E., Singh S., Lakkaraju H., editors. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society; 2020.
[46] Lundberg S.M., Lee S.-I. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems; Long Beach, California, USA: Curran Associates Inc.; 2017. p. 4768–4777.
[47] Ramírez-Mena A, Andrés-León E, Alvarez-Cubero MJ, Anguita-Ruiz A, Martinez-Gonzalez LJ, Alcala-Fdez J. Explainable artificial intelligence to predict and identify prostate cancer tissue by gene expression. Comput Methods Prog Biomed 2023;240:107719.
[48] Sobhan M., Mondal A.M., editors. Explainable machine learning to identify patient-specific biomarkers for lung cancer. 2022 In: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2022: IEEE.
[49] Rajpal S, Rajpal A, Agarwal M, Kumar V, Abraham A, Khanna D, et al. XAI-CNVMarker: Explainable AI-based copy number variant biomarker discovery for breast cancer subtypes. Biomed Signal Process Control 2023;84:104979.
[50] Dwivedi K, Rajpal A, Rajpal S, Agarwal M, Kumar V, Kumar N. An explainable AI-driven biomarker discovery framework for non-small cell lung cancer classification. Comput Biol Med 2023;153:106544.
[51] Sundararajan M., Taly A., Yan Q., editors. Axiomatic attribution for deep networks. In: Proceedings of the International Conference on Machine Learning; 2017: PMLR.
[52] Gao L., Shu K., Zhang J., Sheng V.S., editors. Explainable Transcription Factor Prediction with Protein Language Models. 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2023: IEEE.
[53] Li Z, Jiang Y, Lu M, Li R, Xia Y. Survival prediction via hierarchical multimodal co-attention transformer: a computational histology-radiology solution. IEEE Trans Med Imaging 2023.
[54] Dai J., Upadhyay S., Aivodji U., Bach S.H., Lakkaraju H., editors. Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society; 2022.
[55] Zeiler M.D., Fergus R., editors. Visualizing and understanding convolutional networks. Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13; 2014: Springer.
[56] Gao Y, Zhang Y, Wang H, Guo X, Zhang J. Decoding behavior tasks from brain activity using deep transfer learning. IEEE Access 2019;7:43222–32.
[57] Simonyan K., Vedaldi A., Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv pre-print server. 2014.
[58] de Souza JrLA, Mendel R, Strasser S, Ebigbo A, Probst A, Messmann H, et al. Convolutional neural networks for the evaluation of cancer in barrett's esophagus: explainable AI to lighten up the black-box. Comput Biol Med 2021;135:104578.
[59] Shrikumar A., Greenside P., Shcherbina A., Kundaje A. Not just a black box: learning important features through propagating activation differences. arXiv pre-print server. 2017.
[60] Springenberg J.T., Dosovitskiy A., Brox T., Riedmiller M. Striving for simplicity: The all convolutional net. arXiv preprint arXiv:14126806. 2014.
[61] Badea L, Stănescu E. Identifying transcriptomic correlates of histology using deep learning. PloS One 2020;15(11):e0242858.
[62] Wickstrøm K., Kampffmeyer M., Jenssen R., editors. Uncertainty modeling and interpretability in convolutional neural networks for polyp segmentation. 2018 In: Proceedings of the 28th International Workshop on Machine Learning for Signal Processing (mlsp); 2018: IEEE.
[63] Zhou B., Khosla A., Lapedriza A., Oliva A., Torralba A., editors. Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016.

[64] Cai J, Xing F, Batra A, Liu F, Walter GA, Vandenborne K, et al. Texture analysis for muscular dystrophy classification in MRI with improved class activation mapping. Pattern Recognit 2019;86:368–75.
[65] Lin J., Han G., Xu X., Liang C., Wong T.-T., Chen C., et al. BroadCAM: outcome-agnostic class activation mapping for small-scale weakly supervised applications. arXiv preprint arXiv:230903509. 2023.
[66] Selvaraju R.R., Das A., Vedantam R., Cogswell M., Parikh D., Batra D. Grad-CAM: Why did you say that? arXiv preprint arXiv:161107450. 2016.
[67] Civit-Masot J, Bañuls-Beaterio A, Domínguez-Morales M, Rivas-Pérez M, Muñoz-Saavedra L, Corral JMR. Non-small cell lung cancer diagnosis aid with histopathological images using explainable deep learning techniques. Comput Methods Prog Biomed 2022;226:107108.
[68] Yin Q., Chen W., Wu R., Wei Z., editors. Cox-ResNet: a survival analysis model based on residual neural networks for gene expression data. 2022 In: Proceedings of the International Conference on Networking, Sensing and Control (ICNSC); 2022: IEEE.
[69] Shovon MSH, Mridha M, Hasib KM, Alfarhood S, Safran M, Che D. Addressing uncertainty in imbalanced histopathology image classification of HER2 breast cancer: an interpretable ensemble approach with threshold filtered single instance evaluation (SIE). IEEE Access 2023;11:122238–51.
[70] Altini N, Brunetti A, Puro E, Taccogna MG, Saponaro C, Zito FA, et al. NDG-CAM: nuclei detection in histopathology images with semantic segmentation networks and grad-CAM. Bioengineering 2022;9(9):475.
[71] Kallipolitis A, Revelos K, Maglogiannis I. Ensembling EfficientNets for the classification and interpretation of histopathology images. Algorithms 2021;14 (10):278.
[72] Korbar B., Olofson A.M., Miraflor A.P., Nicka C.M., Suriawinata M.A., Torresani L., et al., editors. Looking under the hood: Deep neural network visualization to interpret whole-slide image analysis outcomes for colorectal polyps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2017.
[73] Chen D., Han Y., Dong Y., Zhou X., editors. Lymphoma ultrasound image classification with causal attention and feature fusion. 2022 IEEE Smartworld, Ubiquitous Intelligence & Computing, Scalable Computing & Communications, Digital Twin, Privacy Computing, Metaverse, Autonomous & Trusted Vehicles (SmartWorld/UIC/ScalCom/DigitalTwin/PriComp/Meta); 2022: IEEE.
[74] Raghavan K. Attention guided grad-CAM: an improved explainable artificial intelligence model for infrared breast cancer detection. Multimed Tools Appl 2023:1–28.
[75] Kayser M., Emde C., Camburu O.-M., Parsons G., Papiez B., Lukasiewicz T., editors. Explaining chest x-ray pathologies in natural language. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention; 2022: Springer.
[76] Aryal M., Yahyasoltani N. Explainable and position-aware learning in digital pathology. arXiv preprint arXiv:230608198. 2023.
[77] Springenberg M, Frommholz A, Wenzel M, Weicken E, Ma J, Strodthoff N. From modern CNNs to vision transformers: assessing the performance, robustness, and classification strategies of deep learning models in histopathology. Med Image Anal 2023;87:102809.
[78] Saihood A, Karshenas H, Naghsh-Nilchi AR. Multi-orientation local texture features for guided attention-based fusion in lung nodule classification. IEEE Access 2023;11:17555–68.
[79] Kallipolitis A, Yfantis P, Maglogiannis I. Improving explainability results of convolutional neural networks in microscopy images. Neural Comput Appl 2023: 1–19.
[80] Campana MG, Colussi M, Delmastro F, Mascetti S, Pagani E. A transfer learning and explainable solution to detect mpox from smartphones images. Pervasive Mob Comput 2024;98:101874.
[81] Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556. 2014.
[82] Sandler M., Howard A., Zhu M., Zhmoginov A., Chen L.-C., editors. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018.
[83] Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L., editors. Imagenet: A large-scale hierarchical image database. 2009 In: Proceedings of the Conference on Computer Vision and Pattern Recognition; 2009: IEEE.
[84] Altan G. DeepOCT: an explainable deep learning architecture to analyze macular edema on OCT images. Eng Sci Technol, Int J 2022;34:101091.
[85] Huang K., Gao W., editors. Real-time neural network inference on extremely weak devices: agile offloading with explainable AI. In: Proceedings of the 28th Annual International Conference on Mobile Computing And Networking; 2022.
[86] Yao J., Wang Y., Lei Z., Wang K., Li X., Zhou J., et al. AI-generated content enhanced computer-aided diagnosis model for thyroid nodules: a ChatGPT-style assistant. arXiv preprint arXiv:240202401. 2024.
[87] Torres-Martos Á, Bustos-Aibar M, Ramírez-Mena A, Cámara-Sánchez S, Anguita-Ruiz A, Alcalá R, et al. Omics data preprocessing for machine learning: a case study in childhood obesity. Genes 2023;14(2):248.
[88] Whalen S, Schreiber J, Noble WS, Pollard KS. Navigating the pitfalls of applying machine learning in genomics. Nat Rev Genet 2022;23(3):169–81.
[89] Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol 2022;23(1):40–55.
[90] Babjac A.N., Lu Z., Emrich S.J., editors. CodonBERT: using BERT for sentiment analysis to better predict genes with low expression. In: Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics; 2023.
[91] Tran M., Schmidle P., Wagner S.J., Koch V., Lupperger V., Feuchtinger A., et al. Generating highly accurate pathology reports from gigapixel whole slide images with HistoGPT. medRxiv. 2024:2024.03. 15.24304211.
[92] Graziani M., Lompech T., Müller H., Andrearczyk V., editors. Evaluation and comparison of CNN visual explanations for histopathology. In: Proceedings of the AAAI Conference on Artificial Intelligence Workshops (XAI-AAAI-21), Virtual Event; 2021.
[93] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1(5): 206–15.
[94] Chanda T, Hauser K, Hobelsberger S, Bucher T-C, Garcia CN, Wies C, et al. Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma. Nat Commun 2024;15(1):524.
[95] Li Y., Zhou J., Verma S., Chen F. A survey of explainable graph neural networks: taxonomy and evaluation metrics 2022.
[96] Cabitza F, Natali C, Famiglini L, Campagner A, Caccavella V, Gallazzi E. Never tell me the odds: Investigating pro-hoc explanations in medical decision making. Artif Intell Med 2024;150:102819.
[97] Del Ser J, Barredo-Arrieta A, Díaz-Rodríguez N, Herrera F, Saranti A, Holzinger A. On generating trustworthy counterfactual explanations. Inf Sci 2024;655:119898.
[98] Huang W, Suominen H, Liu T, Rice G, Salomon C, Barnard AS. Explainable discovery of disease biomarkers: the case of ovarian cancer to illustrate the best practice in machine learning and Shapley analysis. J Biomed Inform 2023;141: 104365.
[99] Qiu J, Li L, Sun J, Peng J, Shi P, Zhang R, et al. Large AI models in health informatics: applications, challenges, and the future. IEEE J Biomed Health Inform 2023;27(12):6074–87.
[100] Mohamed YA, Khoo BE, Asaari MSM, Aziz ME, Ghazali FR. Decoding the black box: Explainable AI (XAI) for cancer diagnosis, prognosis, and treatment planning-A state-of-the art systematic review. Int J Med Inform 2024:105689.
[101] Withnell E, Zhang X, Sun K, Guo Y. XOmiVAE: an interpretable deep learning model for cancer classification using high-dimensional omics data. Brief Bioinforma 2021;22(6):bbab315.
[102] Ye T, Li S, Zhang Y. Genomic pan-cancer classification using image-based deep learning. Comput Struct Biotechnol J 2021;19:835–46.
[103] Bichindaritz I., Liu G., editors. An explainable deep network framework with case-based reasoning strategies for survival analysis in oncology. In: Proceedings of the International KES Conference on Innovation in Medicine and Healthcare; 2023: Springer.
[104] Maouche I, Terrissa LS, Benmohammed K, Zerhouni N. An explainable AI approach for breast cancer metastasis prediction based on clinicopathological data. IEEE Trans Biomed Eng 2023;70(12):3321–9.
[105] Malhi A., Kampik T., Pannu H., Madhikermi M., Främling K., editors. Explaining machine learning-based classifications of in-vivo gastral images. 2019 Digital Image Computing: Techniques and Applications (DICTA); 2019 2-4 Dec. 2019.
[106] Foersch S, Glasner C, Woerl A-C, Eckstein M, Wagner D-C, Schulz S, et al. Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer. Nat Med 2023;29(2):430–9.
[107] Shamai G, Livne A, Polónia A, Sabo E, Cretu A, Bar-Sela G, et al. Deep learning-based image analysis predicts PD-L1 status from H&E-stained histopathology images in breast cancer. Nat Commun 2022;13(1):6753.
[108] Lamy J-B, Sekar B, Guezennec G, Bouaud J, Séroussi B. Explainable artificial intelligence for breast cancer: a visual case-based reasoning approach. Artif Intell Med 2019;94:42–53.
[109] DeGrave AJ, Cai ZR, Janizek JD, Daneshjou R, Lee S-I. Auditing the inference processes of medical-image classifiers by leveraging generative AI and the expertise of physicians. Nat Biomed Eng 2023:1–13.
[110] Lombardi A, Arezzo F, Di Sciascio E, Ardito C, Mongelli M, Di Lillo N, et al. A human-interpretable machine learning pipeline based on ultrasound to support leiomyosarcoma diagnosis. Artif Intell Med 2023;146:102697.
[111] Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev 1950;78(1):1–3.
[112] Krzyziński M, Spytek M, Baniecki H, Biecek P. SurvSHAP (t): time-dependent explanations of machine learning survival models. Knowledge-Based Syst 2023; 262:110234.
[113] Macedo D.C., De Lima John W., Santos V.D., LO M.T., MP N.F., Arrais N., et al., editors. Evaluating interpretability in deep learning using breast cancer histopathological images. 2022 In: Proceedings of the 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI); 2022: IEEE.
[114] Li Y, Stanojevic S, Garmire LX. Emerging artificial intelligence applications in spatial transcriptomics analysis. Comput Struct Biotechnol J 2022;20:2895–908.
[115] Anguita-Ruiz A, Segura-Delgado A, Alcalá R, Aguilera CM, Alcalá-Fdez J. eXplainable Artificial Intelligence (XAI) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research. PLOS Comput Biol 2020;16(4):e1007792.
[116] Zhang C, Wang L, Shi Q. Computational modeling for deciphering tissue microenvironment heterogeneity from spatially resolved transcriptomics. Comput Struct Biotechnol J 2024.