Briefings in Bioinformatics, 2024, bbae449
https://doi.org/10.1093/bib/bbae449
Review

Designing interpretable deep learning applications for functional genomics: a quantitative analysis

Arno van Hilten[1,‡], Sonja Katz[1,2,‡,*], Edoardo Saccenti[2], Wiro J. Niessen[3], Gennady V. Roshchupkin[1,4]

[1]Department of Radiology and Nuclear Medicine, Erasmus MC, 3015 GD Rotterdam, The Netherlands
[2]Laboratory of Systems and Synthetic Biology, Wageningen University & Research, 6700 HB Wageningen WE, The Netherlands
[3]Department of Imaging Physics, Delft University of Technology, 2628 CD Delft, The Netherlands
[4]Department of Epidemiology, Erasmus MC, 3015 GD Rotterdam, The Netherlands

[*]Corresponding author. Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands. E-mail: sonja.katz@wur.nl
[‡]Arno van Hilten and Sonja Katz, that they contributed equally.

Abstract

Deep learning applications have had a profound impact on many scientific fields, including functional genomics. Deep learning models can learn complex interactions between and within omics data; however, interpreting and explaining these models can be challenging. Interpretability is essential not only to help progress our understanding of the biological mechanisms underlying traits and diseases but also for establishing trust in these model's efficacy for healthcare applications. Recognizing this importance, recent years have seen the development of numerous diverse interpretability strategies, making it increasingly difficult to navigate the field. In this review, we present a quantitative analysis of the challenges arising when designing interpretable deep learning solutions in functional genomics. We explore design choices related to the characteristics of genomics data, the neural network architectures applied, and strategies for interpretation. By quantifying the current state of the field with a predefined set of criteria, we find the most frequent solutions, highlight exceptional examples, and identify unexplored opportunities for developing interpretable deep learning models in genomics.

Keywords: interpretable; explainable; visible; deep learning; omics; genomics

Introduction

The overarching goal of functional genomics research is to understand and intervene in the underlying biological processes between genotype, environment, and phenotype [1]. To probe these underlying biological processes, a wide variety of omics are gathered. Genetic data (DNA sequences, single nucleotide polymorphisms (SNPs), and copy number variants (CNV)) serves as the stable foundation from which many biological processes start. Transcriptomics data provides the expression of genes by measuring messenger RNA (mRNA), which is dynamically transcribed from DNA and can be regulated by various epigenetic mechanisms. These epigenetic mechanisms include DNA methylation, RNA methylation, regulatory non-coding RNA (ncRNA), histone modifications, and chromatin accessibility. Altogether, these omic types provide insight in the biological processes that link heritable and environmental factors to observable characteristics in a cell or individual [1]. The underlying biological processes can be very complex [2] and massive amounts of data are acquired to study these processes. The massive amount of data and the complexity of the biological processes allow and justify using more complex models, such as deep learning models, to help unravel the intricacies between genotype, environment, and phenotype.

Deep learning, a subset of machine learning, consists of a group of methods that can capture complex interactions and non-linear relationships given a sufficiently large number of examples. Deep learning solutions have been successful in a wide variety of applications; a few examples include biomedical image segmentation [3], natural language modelling [4], and protein modelling [5]. Despite their widespread popularity and convincing performances of deep-learning models there are drawbacks in using neural networks; deep learning methods offer no explanation for their decision-making process.

Explaining the decision-making process of AI-driven technologies is essential, given that these technologies impact various facets of our daily lives, encompassing critical domains such as healthcare, governance, and legal systems [6]. Recent European data and privacy legislation, particularly the General Data Protection Regulation (GDPR), has put '*explainability*' as a top priority in machine learning research [7]. In the special case of decisions reached using automated processing, the rights of data subjects (e.g. patients) were phrased as to '*obtain an explanation of the decision reached*' [8], encoding the right to explanations for data subjects within European law. Interpretability is central to inspiring trust in neural networks. Understanding *why* and *how* a model makes decisions, as opposed to blindly trusting that they are correct,

Received: April 5, 2024. **Revised:** August 7, 2024. **Accepted:** August 28, 2024

contributes to its trustworthiness [6]. The ability to understand the reasoning behind a model's prediction (the *why*) is often termed 'explainability', whereas the ability to fully understand the inner workings of the model (the *how*) is referred to as its 'interpretability'. According to this definition, all interpretable AI is inherently explainable but not all explainable AI systems are interpretable. Generally, rule-based models and simpler machine learning models such as linear and logistic regression and decision trees are intrinsically interpretable. In more complex models, like large random forests and dense neural networks, predictions can often only be approximated through explainability methods [9].

Designing an interpretable deep-learning application is an inherently creative process without a set path to a successful application. Researchers require not only in-depth knowledge of the data they are working with, but also need to be aware of the different types of model architectures that can be applied, their limitations such as required sample size, how these architectures can be made interpretable, and, finally, how to combine these components in a way that yields insight into the biological question of interest. Fortunately, there are numerous successful applications of interpretable deep learning in genomics that can guide development and inspire novel approaches. In this study, we provide an in-depth analysis of the current state of the interpretable genomics field by discussing the most prevalent solutions, visualise common combinations of solutions, and identify interesting opportunities for designing interpretable deep-learning applications in genomics. While there exist numerous comprehensive reviews outlining different interpretability approaches [10–14], we provide a more practical guide for researchers that want to bring interpretability into their models. We dissect the challenges associated with the three key considerations of every interpretable deep-learning application in genomics: (i) the characteristics of the utilised data, (ii) the model architecture, and (iii) the interpretation strategy selected. We assessed each surveyed paper according to predefined criteria relating to these key considerations (Section 6), resulting in an overview table that provides a quantitative overview of all approaches and applications of interpretable deep learning in genomics (Section 6). From this comprehensive overview table, we extract general statistics that provide insight into the prevalent research questions addressed and the popularity of both models and interpretability methods. Furthermore, we provide supporting graphics to outline (dis-) advantages for combinations of data types, models, and interpretation strategies, with the goal of identifying inspirational examples, challenges, and unexplored opportunities in developing an interpretable deep-learning model.

In the first section **Considerations for designing an interpretable model**, we discuss the major decisions that need to be taken during model development and we provide a short intuition on the most commonly used data types, models, and interpretability strategies. After establishing the fundamentals, we analyse and visualise trends and statistics of current solutions in section **The current state of the field: a quantification**. Finally, in **Opportunities and perspectives**, we highlight unexplored opportunities, best practices, and considerations for designing interpretable deep-learning applications in genomics.

Considerations for designing an interpretable model

We focus on three key aspects that mark major decisions during developing an interpretable deep-learning model for omics data: (i) the type and characteristics of data used, (ii) the model architecture of choice, and (iii) the interpretability strategy deployed. As all these three aspects are intertwined and cannot be discussed separately, we aim to clarify dependencies by gradually stacking information. First, we expand and motivate the criteria for the quantitative analysis, before moving on to the quantitative analysis itself.

Criteria and considerations used for evaluation:

1. **Characteristic of the input data**
   - (a) *Sequencing type*: was the data generated using single cell or bulk sequencing?
   - (b) *Omic type*: which omics (e.g. SNPs, CNVs, mRNA, CpGs) were used in the study? Single omic or multi-omic?
   - (c) *Data dimensions*: what is the number of examples (e.g. number of patients, cells) in the dataset?
2. **Choosing a model architecture**:
   - (a) *Neural network type*: what kind of neural network architecture was used (convolutional neural network (CNN), visible neural network (VNN), transformer, etc.)
   - (b) *Input dimensions*: what was the dimension of the input for each example to the network? (e.g between 50 and 100, between 1000 and 2000, more than 1 million)
   - (c) *Computational resources*: which computational resources (CPU/GPU memory) was available for model construction and interpretation?
3. **Navigating interpretation strategies**
   - (a) *Biological level of interpretation*: on which (biological) level was interpretability applied (gene-level, pathway, etc.)?
   - (b) *Interpretation taxonomy*: how was model interpretation facilitated (global, local, attribution methods, hidden semantics, etc.)?
   - (c) *Interpretation strategy*: what are the defining characteristics of the interpretation method (use prior knowledge, visualisation, backpropagation method, etc.)?
   - (d) *Prior knowledge*: if prior knowledge was used, which database was used? (KEGG, Reactome, Gene-Onthology, etc.)

Characteristics of the input data

The data are the basis on which the neural networks are built and its characteristics largely influence the choice of model architecture and the interpretation method. We included studies with at least one of the following types of data: genetic (SNPs, CNVs), transcriptomics (mRNA), and epigenetic data such as chromatin accessibility (ATAC-seq, DNase), non-coding RNAs (ncRNAs), and DNA methylation (CpGs). We make a distinction between single-cell and bulk sequencing, as the challenges and characteristics of these sequencing types can be quite different and we categorised the number of examples in the dataset to give an impression of the volume of the data needed to perform the study.

Table 1 highlights some of the characteristics of the omic types included and challenges associated with designing a neural network for the included omic types. These characteristics and challenges, in combination with the research question mainly shape the realm of possible options for neural network architectures. For instance, due to the expansive dimensions of genetic data, utilising sparse models becomes necessary, since fully connected layers could exceed GPU memory capacities. For sequence data, neural networks that scan the sequences for patterns, such as CNNs are typically utilised. However, the atypical out-of-the-box applications are interesting to highlight. These demonstrate that with both conventional and unconventional transformations of

Table 1. Overview of the characteristics and challenges of the main omics types encountered in the surveyed papers

	Genetic	Transcriptomics		Epigenetic	
Category	SNPs, sequences	Bulk gene expression	Single-cell gene expression	Chromatin accessibility	Methylation
Methods	Genotype arrays, whole genome sequencing	RNA-seq, microarrays	scRNA-seq	ATAC-seq, DNase-seq	BeadChip
Number of measurements	>88 million variants	∼24 000 genes	<24 000 genes	Millions of reads	∼450 000 CpGs
Data type	Categorical (nucleotide, dosage)	Positive continuous (gene expression level)	Positive continuous (gene expression level)	Positive continuous (read counts per region)	Fraction (CpG methylation level)
Challenges for ML	Large input size, small effect sizes, non-coding regions	Input order, mixture of cell signals	Identifying cell-types and cell-states, data sparsity	TN5 bias, peak-calling	High dimensionality, cell-type heterogeneity

the data, one can open up new possibilities. For example, the use of *k*-mers to decrease the input size of genetic data, and to increase the depth of the data [15]. Transforming the gene expression data into images to apply CNNs and image-based interpretation strategies [16–18]. ChromBPnet [19] avoids peak-calling by using the raw counts as an input. Additionally, they add an additional network trained on the non-peak regions to correct for a bias in ATAC-seq measurements (TN5 bias). The latter is also a great example of demonstrating that a thorough understanding of the data and preprocessing steps is crucial in identifying steps that can be replaced or benefit from deep learning.

Choosing a model architecture

The choice of model architectures is mainly driven by data, technological innovations, and trends. Most deep learning architectures are variations of neural networks. We categorise each network as one of the main neural network types: multi-layer perceptron (MLP), CNN, VNN, graph neural networks (GNN), autoencoders (AE), and the recently introduced generative pretrained transformers (GPT) (Fig. 1). Historically, CNNs were designed for image data, GNNs for data that can be represented as graphs (e.g. social networks, molecules, and proteins), and transformers had their first successes in text-based natural language tasks. However, neural networks can consist of a mix of multiple types of layers, providing endless opportunities to tailor the neural network to specific problems and data types, such as the various types of omics data. Genomics data generally has a large input dimension and one can choose to modify the network to take all the inputs or choose a subset of the data to feed into the network. This is related to the last criterion: the computational resources. Larger networks that take all the data need more memory train longer.

Multi-layer perceptron (MLP) is the most traditional neural network architecture [21]. Each layer consists of a set of neurons that is fully connected to neurons in previous and subsequent layers (see Fig. 1(a). MLP architectures are abundantly used, especially in common supervised tasks relevant in medicine, such as the prediction of cancer types, estimating disease severity, and molecular subtyping [22]. These feed-forward neural networks do not need to be deep with multiple subsequent layers to model complex functions. It has been shown that a shallow network with a single layer with infinite width, can accurately approximate any function [23, 24].

Convolutional neural networks (CNNs) have a rich history of successful applications, particularly in imaging, where they excel in extracting useful patterns from local correlation structures, such as edges in images [21]. CNNs are not fully connected, instead, in each convolutional layer it optimises a predefined number of filters that slide across the input features (as shown in Fig. 1b). This sliding operation over the inputs makes the network invariant to where the pattern is located. In other words, the network is translation-invariant. Stacking multiple layers results in a fully CNN. In such a network, each subsequent layer can capture more abstract patterns. When used in conjunction with genomic data, CNNs commonly take DNA sequences or gene expression matrices as input. Stacking convolutional layers allows the model to learn increasingly complex gene interaction patterns. For example, the first convolutional layer can detect local clusters of co-expressed genes. The second layer subsequently learns how groups of genes are correlated. Deeper layers may ultimately abstract specific expression signatures characteristic of pathways or even certain diseases or cell types. Stacking convolutional layers also contributes to increasing the receptive field: the region used by the network to create a particular feature. Thus, the receptive field defines the largest distance for which interactions can be learned by the network.In the context of genomics, this can be the length of a DNA sequence or values of reads (e.g. DNase, ATAC-seq data) mapped to a reference sequence.

Visible neural networks (VNN) were introduced in the field of biology to tackle two common problems in deep learning in genomics: efficiently handling large input sizes and addressing the lack of interpretability. These types of neural networks reduce the number of learnable parameters by embedding biological information in the network architecture so that only biologically meaningful connections are retained (see Fig. 1d). Each neuron in a VNN represents a biological entity, for example, a gene or a pathway [25]. In the network illustrated in Fig. 1(d), the highlighted neuron represents a gene. Only genetic variants that have a relation to that gene (according to prior biological knowledge) are used as input. The output of this gene could be connected neurons representing pathways that this gene is involved in. Visible neural networks can be seen as a hybrid between GNNs and MLP. It uses the mechanics of fully-connected feed-forward layers but is shaped like a graph using external sources of biological knowledge.

Generative pre-trained transformers (GPT) are the most recently proposed class of neural networks that have had a major impact in research and society [26]. Transformers were developed for the task of translating natural language texts and are performing best on sequential data, such as DNA sequences. A transformer alternates between feed-forward layers and the

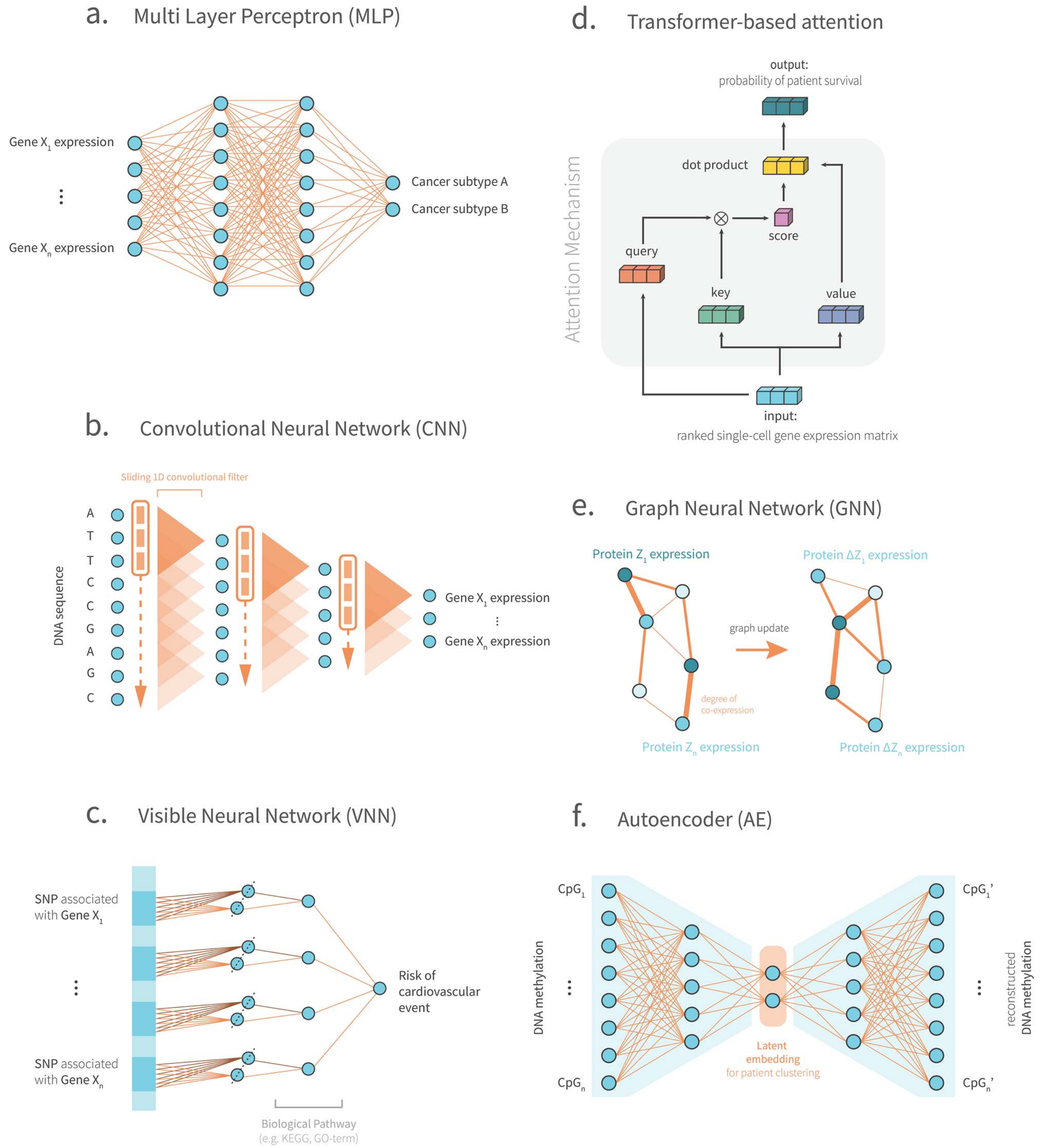
a. Multi Layer Perceptron (MLP)
Gene X_1 expression
⋮
Gene X_n expression
Cancer subtype A
Cancer subtype B
b. Convolutional Neural Network (CNN)
Sliding 1D convolutional filter
DNA sequence
A
T
T
C
C
G
A
G
C
Gene X_1 expression
⋮
Gene X_n expression
c. Visible Neural Network (VNN)
SNP associated with Gene X_1
⋮
SNP associated with Gene X_n
Risk of cardiovascular event
Biological Pathway (e.g. KEGG, GO-term)
d. Transformer-based attention
output: probability of patient survival
Attention Mechanism
dot product
score
query
key
value
input: ranked single-cell gene expression matrix
e. Graph Neural Network (GNN)
Protein Z_1 expression
Protein ΔZ_1 expression
graph update
degree of co-expression
Protein Z_n expression
Protein ΔZ_n expression
f. Autoencoder (AE)
CpG_1
DNA methylation
⋮
CpG_n
Latent embedding for patient clustering
CpG_1'
⋮
CpG_n'
reconstructed DNA methylation

Figure 1. Overview of the most popular neural network architectures. (a) Multilayer perceptron, a fully connected neural network. (b) Convolutional neural network, (c) GNN. (d) GPT; Generative pretrained transformer, here displayed the attention mechanism (inspired by [20]) (e) VNN, and (f) AE.

self-attention mechanism. Self-attention (see Fig. 1e) allows each element in a sequence to dynamically weigh the importance of all other elements. In the original translation task, each word in a sentence can 'attend' to all other words, enabling the model to understand context and relationships within the sequence. In genomics, the equivalents of 'words in a sentence' can be seen as 'genes in a cell' [27]. GPTs are large transformer models that have been trained on massive amounts of data (Supplementary Figure 6b). During training, these models learn to predict the next output in a sequence given the previous inputs. During this self-supervised training procedure, the model gains a deeper understanding of the data and should, therefore, be aware of context. For example, it should be able to infer the gene expression of a masked gene in a cell given the expression of all other genes. These models are also referred to as foundation models, since the pre-trained models, with their better understanding of the general concepts, can be used for various downstream tasks with little fine-tuning [28].

GNN [29] are designed for analysing data structured as graphs, such as molecules, proteins, social networks, or protein interaction networks. A GNN is structured like a graph, using prior information to describe which nodes are connected to which nodes (see Fig. 1c). For example, each protein in a protein interaction network is a node, with edges between proteins that interact. Message passing, aggregating information from neighbouring nodes, enables each node to aggregate and process information from its immediate neighbours. This process occurs in each layer, where layers can be thought of as iterations of the graph with updated weights. This iterative aggregation enables the network to learn node representations that reflect not only the features of the nodes themselves, but also their relationships within the graph. With each successive layer, GNNs integrate information from broader neighbourhoods, capturing more complex and global patterns in the graph structure.

AE (Fig. 1f) are unsupervised neural networks consisting of an encoder and a decoder [21]. The encoder maps the high-dimensional input data into a lower dimensional latent embedding while the decoder reconstructs the original input from this smaller dimensional embedding. This unsupervised encoder-decoder structure, with such an information bottleneck, allows AEs to act as dimensionality reduction tools. For example, an AE can use single-cell gene expression matrices as input and encode them into a UMAP-like structure. The clustering patterns found in this UMAP-like embedding can subsequently be interpreted as different cell cluster and used for e.g. cell type annotation. Variational autoencoders (VAE) [30] use probabilistic resampling to model the output of the encoder as a distribution over the latent space. An extra regularisation term is added in the loss function to encourages the learned distributions to approximate a prior distribution (typically a Gaussian). As a consequence of the probabilistic resampling, each sample can be sampled from a wider area in the latent space as opposed to a single point in regular AEs. This results in a coherent latent space that can be used for generating new data points by sampling from this latent space. For this property VAE have also been used for generating synthetic patient data, such as user-specific cancer types based on DNA methylation data [31].

Navigating interpretation strategies

We classified the interpretation strategies in each paper by the taxonomy defined by Zhang *et al.*, which utilises three dimensions to categorise approaches [12] (Fig. 2).The first dimension divides the interpretation approaches into *active* and *passive* according to whether they require changing the network architecture. Active approaches need a specific configuration of the network to work, for example, embedding biological knowledge in the network architecture. Passive approaches do not have this requirement and can be applied post-hoc to (nearly) any network. The second dimension describes the type of explanation that is obtained by using the method. It differentiates between three major types: *logic rules*, *hidden semantics*, *attributions*, ordered by decreasing explanatory power. Methods that extract logic rules approximate the learned function of the neural network by a set of rules. Hidden semantics includes methods that explain the inner state of a neural network. Attribution, the final option for the second dimension, is subdivided further into *gradient-based*, *permutation*, *perturbation*, *game theory*, *attention*, each describing a different mechanism to attach an importance value to each input. The third and final dimension describes the level of interpretability with regard to the input space, differentiating between *global*, *semi-local*, and *local* approaches. Global interpretability refers to understanding the overall decision logic of a model and its behaviour across all samples, for example, a global overview of the importance of a specific SNP over the whole population. Local explanations provide the interpretation for a single patient, such as a list of genes ranked by their predicted importance to patient survival [32]. As the transition between global and local interpretation is soft, the category of semi-local approaches describes an intermittent state which can be thought of as e.g. biomarkers differentially expressed in a group of similar patients. In addition to the taxonomy, we examined which level of biological information was extracted from the network. This can be inputs (SNPs, genes etc.) or higher level concepts such as gene sets and pathways. Since many of the interpretation methods are novel approaches, we also tagged each interpretation method based on keywords describing methodological characteristics. Finally, if prior biological knowledge is used in the design of the neural network, we tabulate the source of the prior knowledge.

The current state of the field: a quantification

During literature retrieval, we employed a systematic approach and subsequently continuously included newly published articles by evaluating the references and citations of the screened articles (snowballing procedure). In total, our systematic search identified 2008 studies, of which 1146 remained after exclusion of duplicates. During abstract screening, studies were included that (i) use human (multi-) omics measurements, (ii) utilise deep learning architectures, (iii) attempt to facilitate interpretability. Furthermore, to select the most relevant primary studies for our area of interest, we excluded studies which used data featuring a spatial (2D/3D images, spatial-omics) or time component (metabolomics, proteomics; except if used in a multi-omics study), as well as studies with a focus on drug development or predicting genome regulatory elements without a focus on human disease. For the latter, we refer to Koido *et al.* [33], Eraslan *et al.* [34], and Talukder *et al.* [35] for excellent reviews on this topic. Additionally, we included relevant articles identified by the snowballing procedure [36], resulting in a total of 123 research articles for our analysis. A detailed description of the literature retrieval procedure can be found in Supplementary Methods. The table with all the surveyed studies is available online as an interactive table for easy navigation, filtering, and sorting (http://www.roshchupkin.org/xai).

Characteristics of the input data

Interpretable deep learning solutions have been applied to a wide variety of fields and tasks within genomics. Figure 3b shows an overview of fields represented in this study. The majority of interpretable deep learning applications are found in oncology (48%), with neurology (10%) and immunology (10%) following closely behind. This large percentage of oncology studies is likely a result of large amount of publicly available cancer data. For example, TCGA was used in 43 out of the 73 oncology studies. Regarding the types of tasks, supervised learning dominates the field, constituting 78% of interpretable neural network applications. Specifically, supervised classification represents 54%, while regression tasks, encompassing survival prediction, make up 24% of the included applications (see Table 2). Overall, this demonstrates that the utilisation of interpretable models is not limited to specific biological fields or use cases but is widely employed across various disciplines.

Figure 3a graphically summarises the datatypes found in the surveyed papers. Despite the recent advancement of single-cell

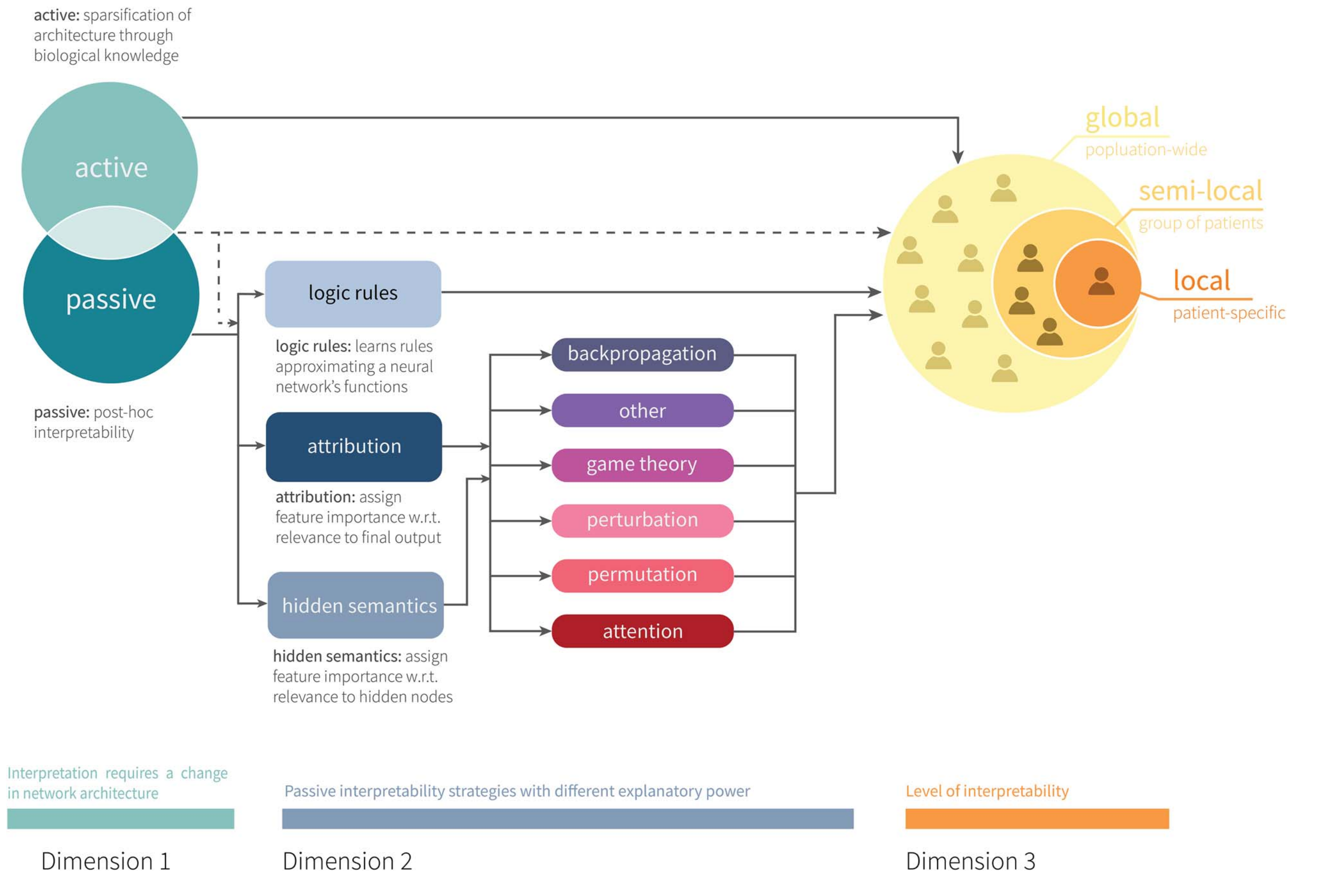
active: sparsification of architecture through biological knowledge
active
passive
passive: post-hoc interpretability
logic rules
logic rules: learns rules approximating a neural network's functions
attribution
attribution: assign feature importance w.r.t. relevance to final output
hidden semantics
hidden semantics: assign feature importance w.r.t. relevance to hidden nodes
backpropagation
other
game theory
perturbation
permutation
attention
global
popluation-wide
semi-local
group of patients
local
patient-specific
Interpretation requires a change in network architecture
Passive interpretability strategies with different explanatory power
Level of interpretability
Dimension 1
Dimension 2
Dimension 3

Figure 2. Overview of the three dimensions defined by Zhang et al. [12] to categorise interpretability strategies. Dimension 1 divides approaches into whether they require to change the network architecture (active) or can be applied post-hoc (passive). Note that active and passive approaches are not mutually exclusive. Dimension 2 further delineates passive interpretability approaches, including methods implementing them. Dimension 3 describes the level of interpretability with regard to the input space, thus differentiating between strategies applicable to whole populations (global), a group of individuals (semi-local), or an individual (local).

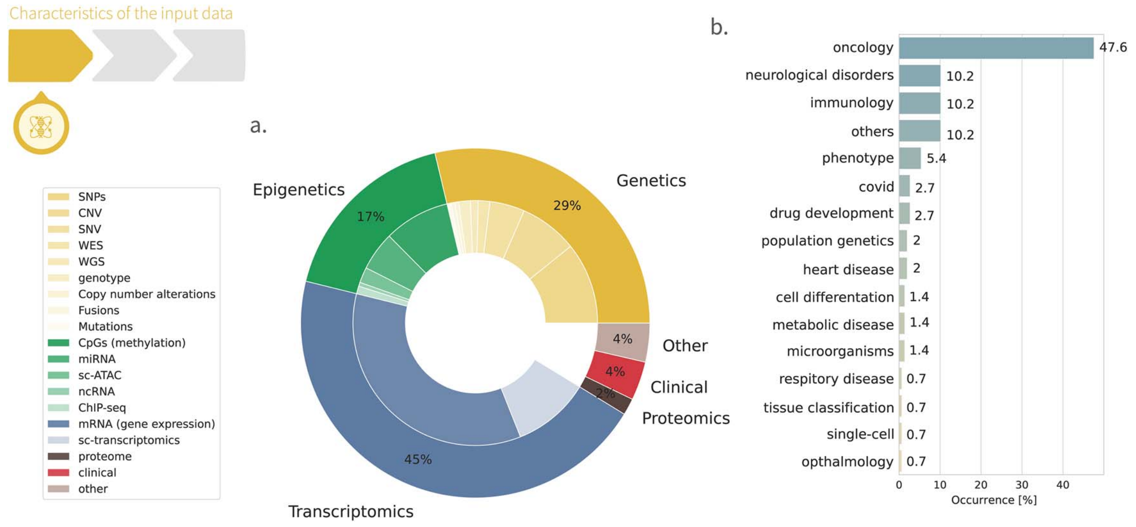

Figure 3. Data types in, and the biological context of, interpretable deep learning studies (123 publications). (a) Summary on the data types usage.Note that the inner circle follows the labels in the legend counter-clockwise. (b) Overview of biological fields.

sequencing, we found that most studies used bulk sequencing (81%). Single-cell sequencing was mostly restricted to single-cell RNA sequencing ($n = 22$) and to a lesser extent single-cell ATAC-seq ($n = 4$).

Genetic data were used as at least one of the input types for 29% of the studies included in the survey (Fig. 3(a). The small effects and the large number of genetic variants have forced the community to bundle genetic datasets to acquire the sample sizes necessary for these studies. This, together with the relatively low cost of genotyping, results in datasets with large numbers of individuals. The largest sample size among interpretable deep-learning applications amounted to 21,105 individuals [37].

Table 2. Overview of the largest categories for each criterion in the main table (see http://www.roshchupkin.org/xai) for the full table. Not all categories sum to the total number of papers ($n = 121$) since some studies fall under multiple categories or use multiple methods or datasets.

General					
Publication per year	2023 (13)*	2022 (45)	2021 (32)	2020 (12)	other (19)
Research Field	oncology (70)	immunology (15)	neurology (15)	heritable traits (8)	other (39)
Task	supervised classification (76)	supervised regression (33)	unsupervised clustering (31)		
Data					
Bulk/single-cell	bulk (97)	single-cell (22)			
Data type	transcriptomics (88)	genetics (56)	epigenetics (34)	clinical (7)	other (10)
Sample size	between 500 and 1000 (33)	>1000 & <5000 (29)	>10 000 & <50 000 (19)	>5000 & <10 000 (12)	other (52)
Model					
Model architecture	multilayer perceptron (37)	autoencoder (36)	visible neural networks (24)	graph neural network (17)	other (27)
Number of features	<10000 (40)	<50000 (31)	<1000 (20)	<100 000 (13)	other (16)
Computational resources	unspecified (88)	GPU <13 GB (23)	GPU >13 GB (16)	CPU (11)	other (5)
Interpretability					
Active/passive (1st dimension)	passive (72)	active (35)	passive & active (14)		
Interpertation strategy (2nd dimension)	attribution (101)	hidden semantics (47)	prior knowledge (37)	connection-weights (32)	other (41)
Interpretation methods (2nd dimension)	SHAP (14)	Integrated Gradients (10)	DeepLIFT (9)	Layerwise Relevance Propagation (7)	other (41)
Granularity of interpretation (3rd dimension)	global (61)	local (49)	semi-local (17)		
Level of interpretation	genes (92)	pathways (35)	SNPs (14)	gene sets (11)	other (24)
Source of prior knowledge (active interpretability)	Gene-Onthology (15)	KEGG (13)	Reactome (11)	StringDB (4)	other (32)

*The systematic literature review was completed before the end of the year.

However, there is a large gap between the sample sizes used in interpretable deep-learning applications and the millions of individuals included in genome-wide association studies (GWAS) (e.g. [38]). Genetic data are sensitive and cannot be readily shared, and gathering large datasets in one place is often infeasible. Distributed learning could be a solution to increase sample sizes [39, 40]. Especially since most common research questions revolve around predicting phenotypes [41–43] and diseases, most commonly cancer [16, 44–49], but also neuro-degenerative diseases [50], psychiatric conditions [51, 52], or hyper-inflammatory conditions [53]. Around these topics, GWAS consortia have formed that have the proper data agreements in place. However, interpretable deep learning with genetic data is not just restricted to these traits and diseases. Examples of other topics for these applications include differentiating populations [54, 55] or detecting gene-gene interactions and epistasis [56–59].

Transcriptomics is the most frequently utilised input type covering 45% of the applications. Effect sizes of genes are generally larger than genetic variants and studies can therefore use smaller sample sizes. Sample sizes in the surveyed papers varied between several hundreds [18, 60–62] to tens of thousands of individuals [16, 63–66]. Similar to studies with genetic data, research questions for studies that use bulk transcriptomics data often revolve around predicting various phenotypes based on gene expression differences. Studies that use single-cell sequencing generally have different research questions, and are focused on clustering and integration using AE architectures (e.g. [67–70]). Recent popular publications on these tasks use foundation models, which are promising as they can perform several tasks such as cell-type annotation and batch correction. These large generative models are trained on millions of cells. Geneformer [71] used 30 million cells, scGPT [27] was trained with a 33 million cells, and scFoundation [28] on 50 million cells.

Epigenetics is an increasingly popular research area, covering 17% of publications. Epigenomics data are commonly included as one of the inputs in a multi-omics framework—only a few publications focus on the sole use of epigenetic data (Supplementary Figure 4). Lemsara *et al.* [72] combined four omic types in a sparse AE based on pathways and Pan *et al.* [73] showcased the combination of up to six different data types using a vanilla AE to improve stratification of breast cancer patients. Epigenetic data, often in combination with other data types, have been used to predict cancer states [46, 49, 74–77], drug response [78], COVID-19 [79], metadata [80], or even other omic types, such as gene expression [61, 81].

Neural network architectures underlying interpretable models

Figure 4a shows the publication date of these papers categorised per network type. From this figure, it is clear that the field is in rapid development, with the vast majority (94%) of papers included in this study being published in 2017 or later. It is also evident that the field has not converged to a single network type, and many types of neural networks are being explored. Figure 4b

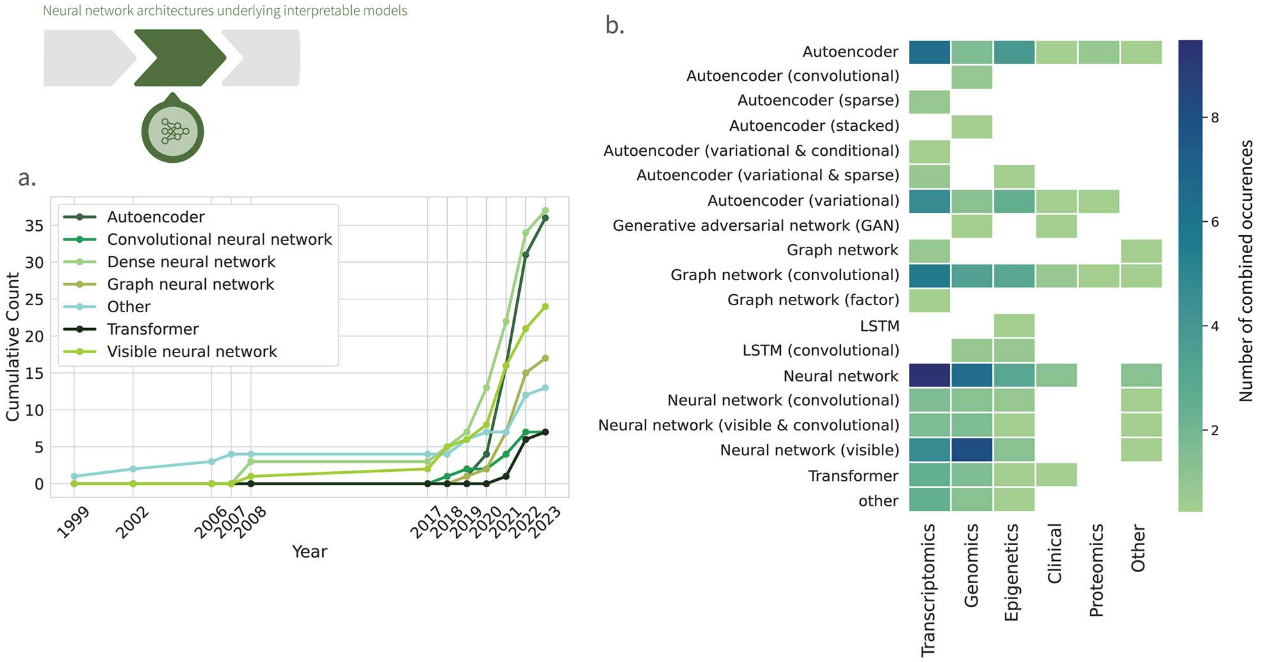

Figure 4. General trends and statistics for model architectures used in the surveyed studies. (a) Number of publications over the years (cumulative) split per neural network type. (b) Heatmap illustrating the variety and quantity of data-model combinations.

provides an overview of the network architectures per data type. Absence of interpretable deep learning architecture for a certain data type can be a consequence of technical limitations, incompatibility, or an opportunity. For example, it will only be a matter of time before interpretable transformer architectures are applied to epigenetics data. At this time, traditional fully connected neural networks (26%) are still the most applied neural network type, closely followed by all the variations of AE (26%), VNNs (17%), and GNNs (12%).

Interpretable deep learning models outperformed standard machine learning baselines in nearly all studies, but this could be inflated by publication bias. For all tasks, the most frequent type of baseline was other neural network architectures (see Supplementary Figure 5). Depending on the task, other popular baselines included statistical frameworks such as Seurat, PCA, and UMAP for clustering tasks, and random forest, linear regression, and support vector machines for classification and regression tasks.

We found that the choice of model architecture is closely linked to data dimensions (Supplementary Figure 6). Regarding the number of input features, such as the number of SNP or gene expression measurements, most surveyed models can handle around 50 000 input features on average (Supplementary Figure 5a). GNNs and in particular VNNs can learn from a larger number of features. Due to the sparsely-connected architecture and reduced computational load, VNNs are ideal candidates in applications with a large number of inputs and limited compute. The largest number of input features, using more than 4 and 6 million genetic variants, were found in studies employing VNNs [52, 82]. GPT-based foundation models, on the other hand, represent the lower end of the spectrum, typically processing between 5 000 and 50 000 input features. This limitation is due to their high computational costs and their exclusive use with single-cell gene expression data, which measures a limited number of genes per cell due to technical constraints.

Standard graph CNNs [83], have been used in all data types (see Fig. 4b). GNNs come in a wide array of variations and can even vary in the way that information from neighbouring nodes are combined. To illustrate, [84] integrated self-attention from transformers in GNNs while [85] used spectral graph convolutions. In terms of the number of samples used in training, we found that all architectures, except for GPT-based foundation models, were typically trained with around 5 000 samples (Supplementary Figure 6b). GNNs have commonly been applied with smaller sample sizes of around 500, making them an attractive choice for researchers with limited data. GPT-based foundation models, however, were only used with at least a million samples. This high sample requirement is feasible only with high-throughput single-cell data, where each cell represents a sample, as it far exceeds the number of patient samples currently available.

All unsupervised learning applications (22.1% of the studies) were clustering tasks, and almost all were applying variations of AEs. Here we find notable differences between the use of bulk data and single-cell data. For single-cell data, there is a data-specific challenge to accurately cluster cell types. Recent articles have proposed to use AEs for this task as AE can provide additional functionalities aside from reducing the dimensionality, for example, remove batch effects, denoise the data, find clusters and integrate multi-omics data [67, 70, 79, 86, 87]. Around 75% surveyed papers using single-cell data use variations of AEs for clustering.

AEs recently adopted the concept of visible networks. This was pioneered by the work of Seninge *et al.* [70] who designed VEGA, a sparse variational AE for sc-transcriptomics supporting user-defined modules, subsequently inspiring numerous other works. Among them was Lotfollahi *et al.* [69], who designed the latent dimensions of an scRNA-seq AE to represent biological modules with their activities being directly interpretable, further proving the versatility of different AE designs. Using biological knowledge to create more sparse AEs, allows these networks to work with

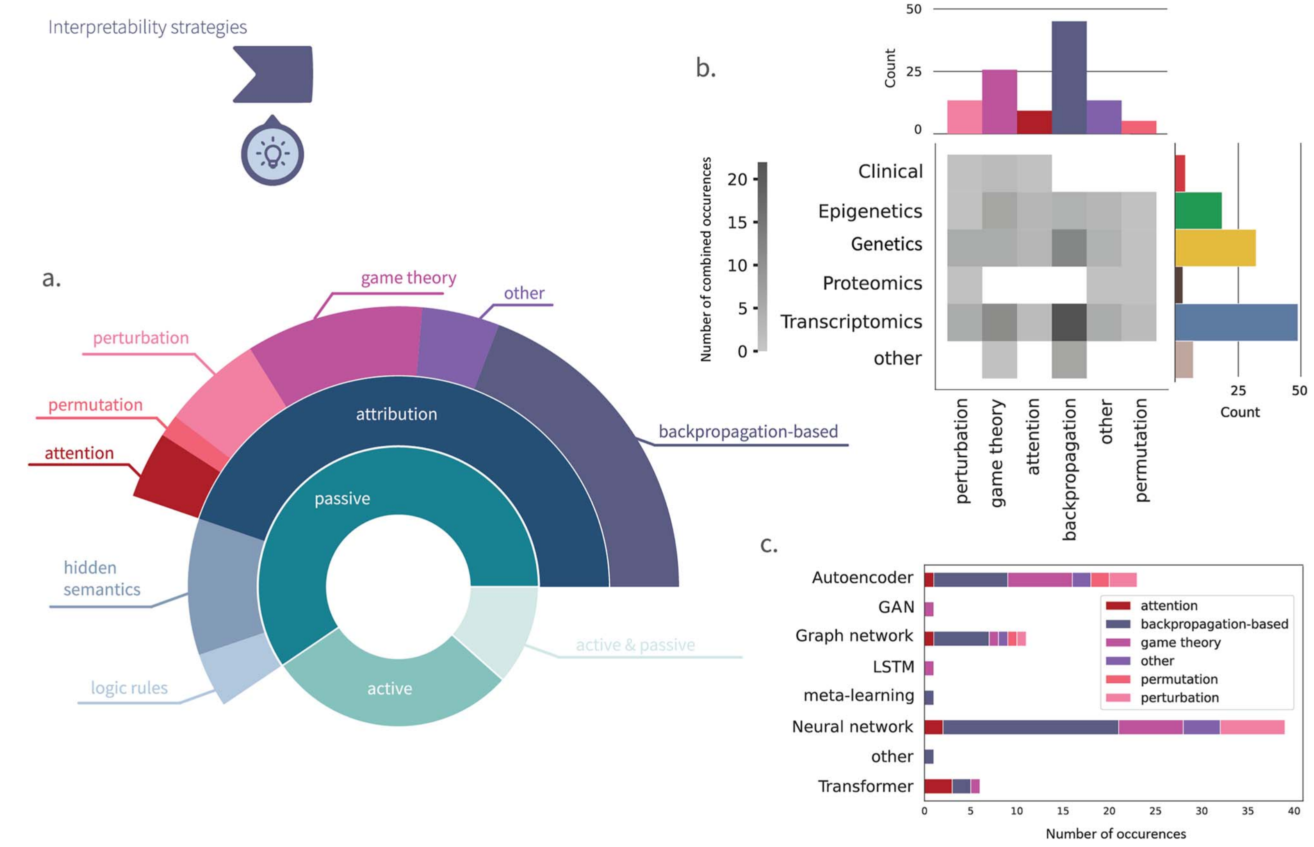

Figure 5. Overview of the interpretability methods and strategies employed. (a) Sunburst plot depicting the strategies employed for the first and second dimensions of interpretability, along with their respective prevalence among 123 surveyed publications. (b) Heatmap illustrating the variety and quantity of data-interpretability combinations. (c) Stacked bar-chart highlighting the identified of model-interpretability combinations.

more input features with a reduced computational cost, something that AEs—with their mirrored design—historically struggled with. It is still an ongoing debate whether the encoder, decoder, or both parts of the model should be sparsified [88]. Finally, the latent embeddings of AEs are known to suffer from the presence of confounding variables [89, 90] and latent features may be entangled, meaning they encode similar information, which hampers direct interpretability.

Interpretability

Dimension 1: active, passive, or both?

Active interpretability methods require architectural changes, often by integrating biological knowledge, in the network structure prior to training. Naturally, the source of prior biological knowledge is largely determined by the application area of the active network. We found a rich diversity of knowledge sources—around 18 different— during our literature search (see Supplementary Figure 3), demonstrating the big advantage of active networks, namely to tune the model with respect to one's scientific interest. While around 50% of publications use one of the three major gene/pathway annotation databases (Gene Ontology (GO) knowledgebase, Reactome, or the Kyoto Encyclopedia of Genes and Genomes (KEGG)), others utilise smaller databases describing e.g. miRNA interactions (miRTarBase) [68], functionally associated genes (GeneMania) [65], curated gene sets (MSigDB) [75, 77, 91], or even own data [61, 76, 92]. Active networks can model gene interactions, reactions, or even whole pathways, but in this they are restricted by the quality of the prior knowledge. Integrating incomplete or subjective data may severely limit model performance and interpretability. Active networks are also hampered in their potential to uncover novel biological connections. None of the active deep-learning applications had the ability to learn new connections or relations after the prior knowledge was introduced, limiting the viability for niche data types, such as microRNA and non-coding RNA, where only a few experimentally validated interactions are recorded in databases.

Passive interpretability approaches can be applied post-hoc, meaning they do not require researchers to change model architectures prior to training. There are a wide variety of post-hoc interpretation algorithms, most of them model-agnostic (Fig. 5a). The latter allows developers to make easy-to-use out-of-the-box interpretation solutions that can work for most neural network architectures. Examples of frameworks for passive interpretation are Captum [93], LIME [94], tf-explain [95], and SHAP [96]. The ease of use makes them the interpretability method of choice in many studies (passive $n = 72$, active & passive $n = 14$, Fig. 5(a). However, while these methods are flexible regarding the type of model used, most of them provide approximations or make strong assumptions regarding data structures. Because passive interpretability strategy is based on different strong assumptions, each method offers a unique perspective on what the network has learned. As a result, the interpretations can vary between passive strategies.

Active and passive interpretability approaches are not mutually exclusive. For example, Elmarakeby *et al.* [74] used VNNs in combination with DeepLIFT [97], a passive approach. Passive analysis methods can complement active methods, such as VNNs, well. Edge weights provide global attribution scores but do not

Table 3. Overview of the most widely used attribution methods with a hand-picked selection of manuscripts applying each method. The full overview table with all entries can be found online (http://www.roshchupkin.org/xai) and in Supplementary Materials

Strategy	Methodology	Citation
Gradient-based	integrated gradients	[64,102–104]
	DeepLIFT	[59,67,74,92,105]
	GradCAM	[18, 85]
	Layerwise relevance propagation (LRP)	[49,106,107]
	SHAP GradientExplainer	[108,109]
	DeepSHAP	[110]
permutation	–	[46,111]
perturbation	LIME	[112]
	modify input	[113,114]
game theory	SHAP	[72,88,115,116]
attention	–	[76,81,117,118]
other	LINA	[58]
	DeepResolve	[119]
	Diet Networks with element-wise input scaling	[48]

provide the important patterns per individual or which features interact.

Dimension 2: how explicit do explanations need to be?

Logic rule sets were found in the earliest applications of interpretability in neural networks [56, 98–100] However, logic rules are not a method of the past; a recent work of Montanez *et al.* [57] showcased how association rule mining can be used to study epistasis. Association rule mining relies on the construction of sets of SNPs that often occur together across different individuals. By describing the relationship between SNPs in sets, easily understandable logic rules can be derived. Another recent publication [101], tackled the problem of explainability in healthcare by developing a framework capable of directly extracting rule sets from neural networks, which can subsequently be inspected and adjusted by clinical experts. Deriving rule sets is indisputably the most explicit way of understanding the decision-making process of networks. In the case of large or complex networks, however, deriving a reasonable number of (understandable) rules might be infeasible.

Attribution methods provide less explanatory power as logic rules but are generally quite feasible to apply, as demonstrated by their popularity. Attribution methods make up 44.8% of the interpretation strategies (Fig. 5(a). As attribution strategies span such a large portion of the interpretation strategies, we subdivided them further based on methodological differences. Table 3 shows the types of attribution strategies accompanied by a selection of representative application examples.

The largest category, **gradient-based methods**, contains many variations that differ on how the gradient is propagated back to obtain the feature importance. All gradient-based methods surveyed, with the exception of GradCAM, need a reference input to compute the attribution score. This gives researchers the flexibility to compute importances with respect to different starting conditions, for example, to distinguish between tumour and normal tissue samples, normal tissues can be utilised as reference points [102]. However, the use of inadequate references holds the possibility of spurious or misleading results, as demonstrated in the framework XOmiVAE [120]. In this study, the authors show that their interpretability results vary substantially depending on the reference chosen. They conclude that the use of random sets of reference samples is inferior to using normal tissue samples as reference, as the latter specifically highlights important cancer pathways.

Permutation and **perturbation** methods shuffle or change the inputs and observe the change in output. While permutation and perturbation of samples represent a simple way of deriving explanations for any model, they are not well suited for studying large input dimensions or interactions between inputs (e.g. epistasis), as this quickly leads to an exploding number of input combinations to perturb. Perturbations and permutations were used in 11 studies, mainly in transcriptomics and in studies that aim to find epistasis [59] (Fig. 5(b)). Sun *et al.* [112] used local interpretable model-agnostic explanations (LIME) [94] to reveal which genetic variants drive the progression of age-related macular degeneration. LIME utilises perturbed samples to build simple surrogate models approximating the predictions of an underlying black box model, thereby revealing important features [46,111,113].

Interpretability methods relying on **game-theory** almost exclusively revolve around Shapley additive explanations (SHAP) [96]. Its root in cooperative game theory makes SHAP model-agnostic and thus universally applicable— from finding which genes attribute to important pathways in VNNs [88], patient-specific feature importance scores for multi-omics cancer data [72], or capturing relevant age-related CpG-CpG interactions with SHAP GradientExplainer, an extension of the integrated gradients method [108]. However, the calculation of SHAP values is complex (NP-hard) [121], so it can be computationally infeasible to deal with high-dimensional data. Additionally, SHAP values are designed to handle continuous data and thus show limited support for categorical features, making it challenging to apply them to genetic data.

Attention-based interpretations are mainly used in transformer architectures, but has also been integrated in GNNs. Through the use of a graph transformer network and separate attention values for nodes and edges, Kaczmarek *et al.* [76] were able to determine important miRNAs and mRNAs (nodes) as well as their interactions (edges) in TCGA cancer samples. The use of attention is popular when translating one omics type into another, for example, when attempting to predict DNA methylation patterns from genetic sequences, as it reveals relevant regions of interest of the input data type, giving further insight into the interplay of genomic mechanisms [81,117].

Although the bulk of attribution methods in genomics are adopted from the fields of image recognition or natural language processing, we wanted to highlight some unique approaches stemming from the biological domain. LINA, a linearising neural network architecture developed by Badre *et al.* [58], is a backpropagation method capable of delivering first-order (individual feature importance), as well as second-order (feature interactions) interpretability. Applied to SNP data with the task of epistasis detection, it outperformed other attribution methods such as DeepLIFT or LIME. *DeepResolve* [119] sets the goal of visualising how genetic features interact and contribute to a final phenotype. The method uses gradient ascent and allows negative values in its feature interaction map, thereby addressing limitations of other gradient-based methods.

Hidden semantics should be considered if the goal is to decipher the inner workings of network rather than focus on what is important for the final result. Exploring which patterns hidden neurons are sensitive to can be easily done when working with active networks. As hidden nodes represent biological entities in sparse networks, such as GO terms [122] or gene modules [69, 70], their activation can be directly interpreted as their activity. In the VEGA framework, authors even propose calculating differential latent variable (gene module) activity, deriving a differential gene expression-like metric. If there is an interest in sequence motifs rather than gene sets, Wang *et al*. showcased that the examination of convolutional filters in CNN, which act as a 'motif detector,' can uncover known Alzheimer disease-associated patterns [42]. In a versatile framework, Märtens *et al*. showed the possibility to simultaneously reduce dimensions and enforce clustering in the latent space of a variational AE [123]. To circumvent the need for additional dimensionality reduction methods like tSNE or PCA, an AE with only two latent dimensions was designed so that the latent space can be directly visualised [54, 55]. For studies not directly observing the activation of hidden nodes or employing AEs with a minimal latent size, our survey revealed that post-hoc attribution methods are often used to infer the meaning of hidden nodes. As an example, Janizek *et al*. utilised integrated gradients in their biologically constrained AE to explain latent feature contribution to reconstruction accuracy, and to find the contribution of genes to pathways [88]. Additionally, they analysed single-cell multi-omics data through sequential perturbation of latent features in an variational AE and observed its downstream differences [87].

Dimension 3: from individual explanations to general patterns

In the surveyed studies, for each data type and each neural network architecture, there was a good mix of local, semi-local, and global interpretation methods applied. We did not find a clear preference for any particular level of interpretation granularity. The choice of granularity mainly depends on the goal of the study. For precision medicine, *local* methods are employed to obtain patient-specific explanations. Popular examples include the investigation of important input features for supervised prediction of a variety of phenotypes, including cancer [16, 64, 82, 124], autism [125], macular degeneration [122], and multiple sclerosis [50], or interpretation of ECG readouts [114]. Besides interrogating the decisions behind predictions for patients, local interpretability can also be used to gain insight into how genes influence the glycophenotype of cells [126], how gene expression and DNA methylation are connected [81], or which genes are best used to approximate the activity of other genes in gene regulatory networks [127].

If the outcome of a group of individual, rather than just one individual, is of interest, *semi-local* approaches should be preferred. Semi-local interpretability is of considerable interest when conducting biomarker discovery or survival analysis, as these research questions have the underlying assumption that groups of individuals exist that can be characterised by a unique genomic pattern. Especially in cancer research, namely NSCLC [47, 128], GBM [129], and BRCA [73,120], we found that by employing semi-local interpretability approaches, the direct characterisation of patient subgroups in terms of important (multi-omics) features was enabled. In single-cell sequencing, semi-local methods can enable the characterisation of cell clusters or aggregates of cell-clusters (pseudo-bulk aggregates), which is a key point when trying to study tissue heterogeneity [51].

The highest level of interpretation is constituted by *global* approaches. They aim at explaining the network as a whole—these research questions revolve around identifying the most predictive variants, genes, and pathways. Linear models, as used in GWAS studies, always provide global interpretations. Neural networks can achieve these interpretations by taking a bottom-up approach of deriving global insight by aggregating all local interpretations when assuming independent and identically distributed random samples [130] or by using global interpretation methods such as inspecting the weights of VNN or extracting a rule set.

Opportunities and perspectives

There is a plethora of tools and strategies to achieve interpretable deep learning. In this review, we have tabulated and analysed 123 studies of interpretable deep learning applications in genomics. We observe an evolving and growing field, rich with a wide variety of strategies and tools. Overall, the most applied neural network architecture is still the traditional fully connected neural network, closely followed by newer network types such as the AE, VNN, and GNN. Post-hoc interpretation methods, in particular attribution methods, from popular frameworks such as DeepLIFT [97], SHAP [96], and Captum [93] make up the majority of interpretability approaches. However, with the rising popularity of GNNs and VNNs, the number of applications with active approaches, in which biological knowledge is used to shape the connections in a neural network, will continue to grow.

A lack of diversity and reproducibility within studies

Most studies validate their interpretations by comparing their findings to existing literature, either by directly comparing the important genes or by conducting enrichment and pathway analysis (e.g. [52,106,113,131]). Although valid when done rigorously, this may be prone to confirmation bias. Holzscheck *et al*. [91] went one step further and validated their interpretations by recapitulating associations from the literature with in-silico gene knockdowns. Other studies, such as Nguyen *et al*. [132], corroborated their findings by integrating other data types. The authors tested the significance of the identified SNP-gene pairs by overlapping these with promoters and enhancer regions derived from Hi-C data. Finally, there are a few studies in this survey that conducted biological experiments to validate their findings. For example, Elmarakeby *et al*. [74] experimentally validated the relevance of *MDM4*, a gene identified by applying DeepLIFT on their VNN. They found that over-expressed *MDM4* was significantly associated with resistance to medication and that depletion of *MDM4* resulted in a significant reduction of proliferation of prostate cancer cells. Experimental validations are expensive and often challenging, but are ultimately necessary to ascertain the causality of results obtained through interpretability methods. To reduce cost and time, a promising advancement in interpreting complex models could be incorporating causality or mechanistic modelling into prediction models [133].

Until causal models are ready, robust computational validation is essential. Unfortunately, we found that most studies (n = 115) utilise only a single interpretation strategy; only six studies used multiple interpretation methods [16,63,78,88,103,114]. With this wide range of different interpretation methods available, and without a consensus on the best methods, it is worthwhile to apply multiple interpretation methods to obtain multiple perspectives. As each interpretation strategy has its own set of strengths and weaknesses, a combination of interpretation methods will paint a clearer and more consistent picture. Especially the use of interpretation methods of different categories may complement each other, as global interpretation strategies might miss individual-level or group-level patterns. Local interpretation strategies, on the other hand, may fail to provide a clear overview. Even the use of multiple interpretation methods from the same dimension may be beneficial, as some methods are particularly designed to find interactions between features while others are designed to find the most important features.

In addition to the observation that most studies apply a single interpretation strategy, we also observe that most studies just apply the interpretation approach once. Neural networks have stochastic elements, and each trained network will inevitably find a different local minimum with different weights. If the goal of a study is to understand the underlying biology, then it is vital to assess the reproducibility of interpretations over multiple network iterations. Studies will need to assess if the set of the most important genes or pathways is consistent over multiple runs. In the surveyed papers, we noticed a general lack of reporting regarding the reproducibility of interpretability results or overlap of results from different passive interpretation strategies. Only a handful of studies have focused on estimating the robustness of their network interpretations [45,66,80,103,134,135].

Finally, it is important to consider that neural networks are non-linear, and that non-linear interactions cannot be captured in a single value. Therefore, extracting a set of rules, although harder, provide more value than the popular attribution methods. Extracted rules can provides insight in the number of interactions, the stability of the prediction model, the behaviour of examples that fall outside of the training set. In this regard, the noticeable absence of probabilistic deep learning methods is noteworthy. A well-calibrated certainty estimate could offer a clear indication of whether a method is applied to a sample unlike any training examples.

Future innovations for interpretable deep learning

The majority of the future innovations in interpretability strategies will likely come from adapting established technologies to genomics data. Novel strategies to explain or interpret deep learning will follow from all scientific fields where deep learning is applied. In fields with inherently more interpretable data, such as image data, the validity of the obtained interpretation can be visually assessed. For example, one can overlay an image and the attribution scores and visually assess the plausibility or, for AEs, one can generate the resulting images while traversing through the latent space. In genomics, intuitive validation is limited. Simulated data can bring relief, and the field has provided useful tools for validating new interpretation strategies (e.g. [136–138]).

The field of genomics in itself also offers unique opportunities for interpretable deep learning. High-quality databases with various types of knowledge (protein-protein interactions, gene and pathway annotations) have been leveraged in various ways to create interpretable neural network architectures. The field has had novel contributions for interpretable deep learning such as DeepLIFT [97] and many specialised neural networks architectures such as VNN architectures [25].

Visible neural networks are promising neural network architectures where all weights are interpretable. Nevertheless, this most likely comes with a cost in performance and in this aspect there is room for improvement. For example, in most implementations, genes and pathways are represented with a single neuron. The number of patterns that a network can learn within a gene or pathway is quite restricted. In contrast, CNNs commonly use between 64 and 512 feature maps. Additionally, the sparsity of the connections limit the number of interactions that these models can capture between genes and pathways. The quality of interpretations is thus strongly dependent on the quality of the biological information embedded. None of the current implementations can compensate well for missing information, and here lies an opportunity to balance between a data-driven approach and a knowledge-guided approach. Learning the gap in the prior knowledge may not be easy, but interpretability, for example, finding the interacting nodes, can be a tool to aid in identifying the missing connections. Networks that can learn missing connections will not only perform better and provide higher quality interpretation, they will also provide opportunities to fill gaps in biological knowledge.

GPT (e.g. [27, 28, 71]) will be a popular tool for at least the short-term future. Ease of use, as shown by the popularity of model-agnostic interpretation methods, is a major factor for adoption. Here, AEs, GNNs, and VNNs have a disadvantage as they require more expert knowledge in the design phase. Counter-intuitively, transformers are easily adopted, as once trained, they can be widely shared and easily applied. Interpretation for these large models is more complex for various reasons. Experiments in the natural language domain have shown that there is often little correlation between important features revealed by gradient-based interpretation methods and attention. Completely different sets of attention weights can result in the same prediction [139], and Bastings *et al.* [140] argues that attention weights reflect the importance of *representations* of inputs rather than the original inputs themselves and that those representations might already have mixed in information from other inputs. Finally, transformers are often used as an extra preprocessing step that transforms the data before applying an additional network for a downstream task, bringing an extra hurdle for interpretation. Novel interpretation strategies may therefore be required to enable transformer architectures to help researchers in understanding the underlying biology in genomics.

In the long term, we expect more large-scale multi-omics datasets. Integrating multi-omics data is difficult as the data combines the complexities of all the omics types used [141]. Deep-learning applications offer unique qualities that are particularly useful for combining omics data. While other machine learning or statistical methods often depend on dimensionality reduction tools, such as PCA, to bring the data to the same dimension, deep-learning models can handle multiple inputs of different sizes and use the appropriate layers for each input. Sequence data can be fed through convolutional filters, whereas expression data can be processed using attention, or fully connected layers. The hierarchical structure of a neural network—each layer leads to a more abstract representation—provides freedom in when and how to connect separate inputs. Similarly, other types of data such as clinical data, imaging data, and patient records can be integrated in a single prediction model [142, 143]. These models will grow in size with the complexity of the data and the complexity of the task, but distributed learning might offer a solution to acquire the sample sizes necessary to train these large models. Finding novel ways to interpret these large models or to combine this data efficiently in smaller, interpretable models will be the challenge.

Concluding remarks

There are many ways to bring interpretability in deep-learning applications in genomics, and many opportunities to develop novel approaches to interpret and explain neural networks. Aside from the concerns and opportunities raised in the previous sections, we quantified and visualised common solutions and combinations of solutions. We observed an exponentially growing field, rich with a wide diversity of methods and strategies, and we believe that this healthy diversity will inspire the next generation of more interpretable and trustworthy interpretable deep-learning applications.

Key Points

- We systematically evaluate and quantify the most frequently used interpretable deep learning algorithms in genomics. Our findings are summarized in a comprehensive and interactive table, which is publicly available.
- By visualising various aspects such as input size, interpretation method, neural network type, and more, we explore common solutions and common combinations of solutions for problems encountered in designing interpretable deep learning applications.
- We discuss and highlight exceptional examples, common approaches, and unexplored opportunities for developing interpretable deep learning models in genomics.
- Our findings indicate that the field is not converging towards a single type of solution; instead, it is exploring a diverse range of approaches.
- We observe a gap in the literature, with few studies employing multiple interpretation techniques to offer a comprehensive understanding of what the deep learning model has learned.

Acknowledgements

The authors wish to thank Dr Maarten (M.F.M.) Engel from the Erasmus MC Medical Library for developing and updating the search strategies.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: The authors declare no conflicts of interest.

Funding

S.K. was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement (860895); E.S. acknowledges the funding received from The Netherlands Organisation for Health Research and Development (ZonMW) through the PERMIT project (Personalized Medicine in Infections: from Systems Biomedicine and Immunometabolism to Precision Diagnosis and Stratification Permitting Individualized Therapies, project number 456008002) under the PerMed Joint Transnational call JTC 2018 (Research projects on personalised medicine—smart combination of pre-clinical and clinical research with data and ICT solutions). G.V.R. is supported by the ZonMw Veni grant (Veni 1936320).

Data and code availability

All tabulated information can be found online under http://www.roshchupkin.org/xai. All code to reproduce all figures can be found on GitHub: https://github.com/sonjakatz/reviewInterpretability_figures.

Author contributions

A.H., S.K. conducted the screening, analysis, wrote the first draft, and prepared the figures. E.S., W.J.N., and G.V.R. provided critical feedback. All authors revised, and approved the paper.

References

1. Caudai C, Galizia A, Geraci F. *et al.* Ai applications in functional genomics. *Comput Struct Biotechnol J* 2021;**19**:5762–90. https://doi.org/10.1016/j.csbj.2021.10.009.
2. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. *Cell* 2017;**169**:1177–86. https://doi.org/10.1016/j.cell.2017.05.038.
3. Litjens G, Kooi T, Bejnordi BE. *et al.* A survey on deep learning in medical image analysis. *Med Image Anal* 2017;**42**:60–88. https://doi.org/10.1016/j.media.2017.07.005.
4. Vaswani A, Shazeer N, Parmar N. *et al.* Attention is all you need. *Adv Neural Inf Process Syst* 2017;**30**:5998–6008. https://doi.org/10.48550/arXiv.1706.03762.
5. Jumper J, Evans R, Pritzel A. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* 2021;**596**:583–9. https://doi.org/10.1038/s41586-021-03819-2.
6. Kaur D, Uslu S, Rittichier KJ. *et al.* Trustworthy artificial intelligence: a review. *ACM Comput Surv* 2022;**55**:39. 1–39:38.
7. Hamon R, Junklewitz H, Sanchez I. *et al.* Bridging the gap between AI and explainability in the GDPR: towards trustworthiness-by-design in automated decision-making. *IEEE Comput Intell Mag* 2022;**17**:72–85. https://doi.org/10.1109/MCI.2021.3129960.
8. Radley-Gardner O, Beale H, Zimmermann R (eds). *Fundamental Texts on European Private Law*. Hart Publishing, Oxford, United Kingdom.
9. Ali S, Abuhmed T, El-Sappagh S. *et al.* Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. *Inf Fusion* 2023;**99**:101805. https://doi.org/10.1016/j.inffus.2023.101805.
10. Novakovsky G, Dexter N, Libbrecht MW. *et al.* Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nat Rev Genet* 2023;**24**:125–37. https://doi.org/10.1038/s41576-022-00532-2.
11. Azodi CB, Tang J, Shiu S-H. Opening the black box: interpretable machine learning for geneticists. *Trends Genet* 2020;**36**:442–55. https://doi.org/10.1016/j.tig.2020.03.005.
12. Zhang Y, Tino P, Leonardis A. *et al.* A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*. 2018;**5**:726–42.

13. Watson DS. Interpretable machine learning for genomics. *Hum Genet* 2022;**141**:1499–513. https://doi.org/10.1007/s00439-021-02387-9.
14. Wysocka M, Wysocki O, Zufferey M. *et al.* A systematic review of biologically-informed deep learning models for cancer: fundamental trends for encoding and interpreting oncology data. *BMC Bioinform* 2023;**24**:1–31. https://doi.org/10.1186/s12859-023-05262-8.
15. Min X, Zeng W, Chen N. *et al.* Chromatin accessibility prediction via convolutional long short-term memory networks with k-mer embedding. *Bioinformatics* 2017;**33**:i92–101. https://doi.org/10.1093/bioinformatics/btx234.
16. Karim MR, Cochez M, Beyan O. *et al.* OncoNetExplainer: explainable predictions of cancer types based on gene expression data. *Proceedings of the 19th IEEE International Conference on Bioinformatics and Bioengineering (BIBE 2019)*. IEEE, 2019, p. 415–22.
17. Karim MR, Rahman A, Jares JB. *et al.* A snapshot neural ensemble method for cancer-type prediction based on copy number variations. *J Comput Biol* 2024;**32**:15281–99.
18. Lombardo E, Hess J, Kurz C. *et al.* DeepClassPathway: molecular pathway aware classification using explainable deep learning. *Eur J Cancer* 2022;**176**:41–9. https://doi.org/10.1016/j.ejca.2022.08.033.
19. Pampari A, Shcherbina A, Nair S. *et al.* Bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints, and regulatory variants. *J Biol Comput* 2023;**1**:1–12.
20. Choi SR, Lee M. Transformer architecture and attention mechanisms in genome data analysis: a comprehensive review. *Biology* 2023;**12**:1033. https://doi.org/10.3390/biology12071033.
21. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;**521**: 436–44. https://doi.org/10.1038/nature14539.
22. Tran KA, Kondrashova O, Bradley A. *et al.* Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med* 2021;**13**:1–17. https://doi.org/10.1186/s13073-021-00968-x.
23. Lu Z, Pu H, Wang F. *et al.* The expressive power of neural networks: a view from the width. *Adv Neural Inf Process Syst* 2017;**30**:1–12.
24. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. *Neural Netw* 1989;**2**:359–66. https://doi.org/10.1016/0893-6080(89)90020-8.
25. Michael KY, Ma J, Fisher J. *et al.* Visible machine learning for biomedicine. *Cell* 2018;**173**:1562–5.
26. Abdullah M, Madain A, Jararweh Y, ChatGPT: fundamentals, applications and social impacts. In: *2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, IEEE, New York, NY, USA, 2022, pp. 1–8.
27. Cui H, Wang C, Maan H. *et al.* scGPT: Towards building a foundation model for single-cell multi-omics using generative AI. *bioRxiv* 2023. https://doi.org/10.1038/s41592-024-02201-0.
28. Hao M, Gong J, Zeng X. *et al.* Large scale foundation model on single-cell transcriptomics. *bioRxiv* 2023. https://doi.org/10.1101/2023.05.29.542705.
29. Sanchez-Lengeling B, Reif E, Pearce A. *et al.* A gentle introduction to graph neural networks. *Distill* 2021;**6**:e33. https://doi.org/10.23915/distill.00033.
30. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv preprint. 2013. https://doi.org/10.48550/arXiv.1312.6114.
31. Choi J, Chae H. methCancer-gen: a DNA methylome dataset generator for user-specified cancer type based on conditional variational autoencoder. *BMC Bioinform* 2020;**21**:1–10. https://doi.org/10.1186/s12859-020-3516-8.
32. Watson DS *Interpretable Machine Learning for Genomics*. New York, NY: Springer, 2022.
33. Koido M, Tomizuka K, Terao C. Fundamentals for predicting transcriptional regulations from dna sequence patterns. *J Hum Genet* 2024;**69**:499–504. https://doi.org/10.1038/s10038-024-01256-3.
34. Eraslan G, Avsec ž, Gagneur J. *et al.* Deep learning: New computational modelling techniques for genomics. *Nat Rev Genet* 2019;**20**:389–403. https://doi.org/10.1038/s41576-019-0122-6.
35. Talukder A, Barham C, Li X. *et al.* Interpretation of deep learning in genomics and epigenomics. *Brief Bioinform* 2021;**22**:bbaa177. https://doi.org/10.1093/bib/bbaa177.
36. Wohlin C. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. New York, NY, USA: ACM, 2014, pp. 1–10.
37. Kassani PH, Lu F, Le Guen Y. *et al.* Deep neural networks with controlled variable selection for the identification of putative causal genetic variants. *Nat Mach Intell* 2022;**4**:761–71. https://doi.org/10.1038/s42256-022-00525-0.
38. Yengo L, Vedantam S, Marouli E. *et al.* A saturated map of common genetic variants associated with human height. *Nature* 2022;**610**:704–12. https://doi.org/10.1038/s41586-022-05275-y.
39. Rieke N, Hancox J, Li W. *et al.* The future of digital health with federated learning. *NPJ Digit Med* 2020;**3**:119. https://doi.org/10.1038/s41746-020-00323-1.
40. Roth HR, Cheng Y, Wen Y. *et al.* Nvidia FLARE: Federated learning from simulation to real-world. arXiv preprint. 2022. https://arxiv.org/abs/2210.13291.
41. Tonner PD, Pressman A, Ross D. Interpretable modeling of genotype-phenotype landscapes with state-of-the-art predictive power. *Proc Natl Acad Sci USA* 2022;**119**:e2114021119.
42. Wang Y, Chen L. DeepPerVar: a multi-modal deep learning framework for functional interpretation of genetic variants in personal genome. *Bioinformatics* 2022;**38**(24):5340–51. https://doi.org/10.1093/bioinformatics/btac696.
43. Demetci P, Cheng W, Darnell G. *et al.* Multi-scale inference of genetic trait architecture using biologically annotated neural networks. *PLoS Genet* 2021;**17**:e1009754.
44. Hu J, Yu W, Dai Y. *et al.* A deep neural network for gastric cancer prognosis prediction based on biological information pathways. *J Oncol* 2022;**2022**:2965166.
45. Feng J, Zhang H, Li F. Investigating the relevance of major signaling pathways in cancer survival using a biologically meaningful deep learning model. *BMC Bioinform* 2021;**22**:47.
46. Li X, Ma J, Leng L. *et al.* MoGCN: a multi-omics integration method based on graph convolutional network for cancer subtype analysis. *Front Genet* 2022;**13**:806842.
47. Kipkogei E, Arango Argoty GA, Kagiampakis I. *et al.* Explainable transformer-based neural network for the prediction of survival outcomes in non-small cell lung cancer (NSCLC). *medRxiv*. 2021. Available from: https://doi.org/10.1101/2021.10.11.21264761.
48. Kobayashi K, Bolatkan A, Shiina S. *et al.* Fully-connected neural networks with reduced parameterization for predicting histological types of lung cancer from somatic mutations. *Biomolecules* 2020;**10**(9):1249. https://doi.org/10.3390/biom10091249.
49. Schulte-Sasse R, Budach S, Hnisz D. *et al.* Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. *Nat Mach Intell* 2021;**3**:513.

50. Ghafouri-Fard S, Taheri M, Omrani MD. *et al*. Application of artificial neural network for prediction of risk of multiple sclerosis based on single nucleotide polymorphism genotypes. *J Mol Neurosci* 2020;**70**:1081–7.
51. Nguyen ND, Huang J, Wang D. A deep manifold-regularized learning model for improving phenotype prediction from multi-modal data. *Nat Comput Sci* 2022;**2**:38–46. https://doi.org/10.1038/s43588-021-00185-x.
52. van Hilten A, Kushner SA, Kayser M. *et al*. GenNet framework: interpretable deep learning for predicting phenotypes from genetic data. *Commun Biol* 2021;**4**:1094. https://doi.org/10.1038/s42003-021-02622-z.
53. Raimondi D, Simm J, Arany A. *et al*. An interpretable low-complexity machine learning framework for robust exome-based in-silico diagnosis of Crohn's disease patients. *NAR Genom Bioinform* 2020;**2**:lqaa011. https://doi.org/10.1093/nargab/lqaa011.
54. Battey CJ, Coffing GC, Kern AD. Visualizing population structure with variational autoencoders. *G3 (Bethesda)* 2021;**11**(1). https://doi.org/10.1093/g3journal/jkaa036.
55. Ausmees K, Nettelblad C. A deep learning framework for characterization of genotype data. *G3 Genes|Genomes|Genetics* 2022;**12**(3). https://doi.org/10.1093/g3journal/jkac020.
56. Motsinger-Reif AA, Reif DM, Fanelli TJ. *et al*. A comparison of analytical methods for genetic association studies. *Genet Epidemiol* 2008;**32**:767–78. https://doi.org/10.1002/gepi.20345.
57. Montanez CAC, Fergus P, Chalmers C. *et al*. SAERMA: Stacked Autoencoder Rule Mining Algorithm for the interpretation of epistatic interactions in GWAS for extreme obesity. *Comput Biol Med* **8**:112379–92.
58. Badre A, Pan C. LINA: a linearizing neural network architecture for accurate first-order and second-order interpretations. *IEEE Access* 2022;**10**:36166–76. https://doi.org/10.1109/ACCESS.2022.3163257.
59. Greenside P, Shimko T, Fordyce P. *et al*. Discovering epistatic feature interactions from neural network models of regulatory DNA sequences. *Bioinformatics* 2018;**34**:i629–37. https://doi.org/10.1093/bioinformatics/bty575.
60. Lee S, Lim S, Lee T. *et al*. Cancer subtype classification and modeling by pathway attention and propagation. *Bioinformatics* 2020;**36**:3818–24. https://doi.org/10.1093/bioinformatics/btaa203.
61. Yuan L, Lai J, Zhao J. *et al*. Path-ATT-CNN: a novel deep neural network method for key pathway identification of lung cancer. *Front Genet* 2022;**13**:896884. https://doi.org/10.3389/fgene.2022.896884.
62. Ma T, Zhang A. Incorporating biological knowledge with factor graph neural network for interpretable deep learning. arXiv preprint. 2019. Available from: https://arxiv.org/abs/1906.00537.
63. Cho HJ, Shu M, Bekiranov S. *et al*. Interpretable meta-learning of multi-omics data for survival analysis and pathway enrichment. *Bioinformatics* 2023;**39**:btad113. https://doi.org/10.1093/bioinformatics/btad113.
64. Zhang TH, Hasib MM, Chiu YC. *et al*. Transformer for Gene Expression Modeling (T-GEM): An interpretable deep learning model for gene expression-based phenotype predictions. *Cancers (Basel)* 2022;**14**(19):4763. https://doi.org/10.3390/cancers14194763.
65. Ramirez R, Chiu YC, Zhang S. *et al*. Prediction and interpretation of cancer survival using graph convolution neural networks. *Methods* 2021;**192**:120–30. https://doi.org/10.1016/j.ymeth.2021.01.004.
66. Bourgeais V, Zehraoui F, Ben Hamdoune M. *et al*. Deep GONet: self-explainable deep neural network based on gene ontology for phenotype prediction from gene expression data. *BMC Bioinform* 2021;**22**:455. https://doi.org/10.1186/s12859-021-04370-7.
67. Choi Y, Li R, Quon G. Interpretable deep generative models for genomics. *bioRxiv*. 2022. https://doi.org/.10.1101/2021.09.15.460498.
68. Alessandri L, Cordero F, Beccuti M. *et al*. Sparsely-connected autoencoder (SCA) for single cell RNAseq data mining. *NPJ Syst Biol Appl* 2021;**7**:1. https://doi.org/10.1038/s41540-020-00162-6.
69. Lotfollahi M, Rybakov S, Hrovatin K. *et al*. Biologically informed deep learning to infer gene program activity in single cells. *bioRxiv* 2022. https://doi.org/10.1101/2022.02.05.479217.
70. Seninge L, Anastopoulos I, Ding H. *et al*. VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics. *Nat Commun* 2021;**12**:5684. https://doi.org/10.1038/s41467-021-26017-0.
71. Theodoris CV, Xiao L, Chopra A. *et al*. Transfer learning enables predictions in network biology. *Nature* 2023;**618**:616–24. https://doi.org/10.1038/s41586-023-06139-9.
72. Lemsara A, Ouadfel S, Fröhlich H. PathME: pathway based multi-modal sparse autoencoders for clustering of patient-level multi-omics data. *BMC Bioinformatics*. 2020;**21**:146. https://doi.org/10.1186/s12859-020-3465-2.
73. Pan X, Burgman B, Sahni N. *et al*. Deep learning based on multi-omics integration identifies potential therapeutic targets in breast cancer. *bioRxiv*. 2022. https://doi.org/10.1101/2022.01.18.476842.
74. Elmarakeby HA, Hwang J, Arafeh R. *et al*. Biologically informed deep neural network for prostate cancer discovery. *Nature*. 2021;**598**:348–52. https://doi.org/10.1038/s41586-021-03922-4.
75. Azher ZL, Vaickus LJ, Salas LA. *et al*. Development of biologically interpretable multimodal deep learning model for cancer prognosis prediction. In: *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing; 2022; Virtual Event*. New York, NY, USA: Association for Computing Machinery, 2022, p. 636–44. https://doi.org/10.1145/3477314.3507032.
76. Kaczmarek E, Jamzad A, Imtiaz T. *et al*. Multi-omic graph transformers for cancer classification and interpretation. *Pac Symp Biocomput* 2022;**27**:373–84.
77. Levy JJ, Chen Y, Azizgolshani N. *et al*. MethylSPWNet and MethylCapsNet: biologically motivated organization of DNAm neural networks, inspired by capsule networks. *NPJ Syst Biol Appl* 2021;**7**:33. https://doi.org/10.1038/s41540-021-00193-7.
78. Cai Z, Poulos RC, Aref A. *et al*. Transformer-based deep learning integrates multi-omic data with cancer pathways. *bioRxiv* 2022. https://doi.org/10.1101/2022.10.27.514141.
79. Zhou M, Zhang H, Baii Z. *et al*. Single-cell multi-omic topic embedding reveals cell-type-specific and COVID-19 severity-related immune signatures. *bioRxiv [Preprint]* 2023, 2023.01.31.526312. https://doi.org/10.1101/2023.01.31.526312.
80. Fiosina J, Fiosins M, Bonn S. Explainable deep learning for augmentation of small RNA expression profiles. *J Comput Biol* 2020;**27**(2):234–47. https://doi.org/10.1089/cmb.2019.0320.
81. Huang Z, Wang J, Yan Z. *et al*. Differentially expressed genes prediction by multiple self-attention on epigenetics data. *Brief Bioinform* 2022;**23**(3). https://doi.org/10.1093/bib/bbac117.
82. Liu G, Bichindaritz I. An explainable deep network framework with case-based reasoning strategies for survival analysis in cancer. *Research Square* 2022. https://doi.org/10.21203/rs.3.rs-2184342/v1.
83. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907. 2017. Available from: https://arxiv.org/abs/1609.02907.

84. Xing X, Yang F, Li H. *et al*. Multi-level attention graph neural network based on co-expression gene modules for disease diagnosis and prognosis. *Bioinformatics* 2022;**38**(8):2178–86. https://doi.org/10.1093/bioinformatics/btac088.
85. Yingtaweesittikul H, Suphavilai C. Network-guided supervised learning on gene expression using a graph convolutional neural network. *bioRxiv* 2021. https://doi.org/10.1101/2021.12.27.474240.
86. Zhao Y, Cai H, Zhang Z. *et al*. Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nat Commun* 2021;**12**:5261. https://doi.org/10.1038/s41467-021-25534-2.
87. Minoura K, Abe K, Nam H. *et al*. A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. *Cell Rep Methods* 2021;**1**(5):100071. https://doi.org/10.1016/j.crmeth.2021.100071.
88. Janizek JD, Spiro A, Celik S. *et al*. PAUSE: principled feature attribution for unsupervised gene expression analysis. *Genome Biol* 2023;**24**:81. https://doi.org/10.1186/s13059-023-02901-4.
89. Wang H, Wu Z, Xing EP. Removing confounding factors associated weights in deep neural networks improves the prediction accuracy for healthcare applications. *Pac Symp Biocomput* 2019;**24**:54–65.
90. Zhao Q, Adeli E, Pohl KM. Training confounder-free deep learning models for medical applications. *Nat Commun* 2020;**11**:6010. https://doi.org/10.1038/s41467-020-19784-9.
91. Holzscheck N, Falckenhayn C, Söhle J. *et al*. Modeling transcriptomic age using knowledge-primed artificial neural networks. *NPJ Aging Mech Dis* 2021;**7**:15. https://doi.org/10.1038/s41514-021-00068-5.
92. Albaradei S, Albaradei A, Alsaedi A. *et al*. MetastaSite: Predicting metastasis to different sites using deep learning with gene expression data. *Front Mol Biosci* 2022;**9**:913602. https://doi.org/10.3389/fmolb.2022.913602.
93. Kokhlikyan N, Miglani V, Martin M. *et al*. Captum: A unified and generic model interpretability library for PyTorch. *arXiv* 2020. Available from: https://arxiv.org/abs/2009.07896.
94. Ribeiro MT, 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; 2016 Aug 13–17; San Francisco, CA, USA*. New York, NY: Association for Computing Machinery, 2016, p. 1135–44.
95. Meudec R. *tf-explain [software]*. Version 0.3.1. Zenodo, 2021. Available from: https://github.com/sicara/tf-explain. https://doi.org/10.5281/zenodo.5711704.
96. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, *et al*., (eds). *Advances in Neural Information Processing Systems* 30. Curran Associates, Inc., 2017, p. 4765–74.
97. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: *International Conference on Machine Learning*. PMLR, 2017, p. 3145–53.
98. Neagu C-D, Avouris N, Kalapanidas E. *et al*. Neural and neuro-fuzzy integration in a knowledge-based system for air quality prediction. *Appl Intell* 2002;**17**:141–69. https://doi.org/10.1023/A:1016108730534.
99. Pal NR, Sharma A, Sanadhya SK. Deriving meaningful rules from gene expression data for classification. *J Intell Fuzzy Syst* 2008;**19**:171–80.
100. Chen C-F, Feng X, Szeto J. Identification of critical genes in microarray experiments by a neuro-fuzzy approach. *Comput Biol Chem* 2006;**30**:372–81. https://doi.org/10.1016/j.compbiolchem.2006.08.004.
101. Shams Z, Dimanov B, Kola S. *et al*. REM: an integrative rule extraction methodology for explainable data analysis in healthcare. *medRxiv* 2021. https://doi.org/10.1101/2021.01.25.21250459.
102. Jha A, Quesnel-Vallières M, Wang D. *et al*. Identifying common transcriptome signatures of cancer by interpreting deep learning models. *Genome Biol* 2022;**23**:117. https://doi.org/10.1186/s13059-022-02681-3.
103. Dwivedi K, Rajpal A, Rajpal S. *et al*. An explainable AI-driven biomarker discovery framework for Non-Small Cell Lung Cancer classification. *Comput Biol Med* 2023;**153**:106544. https://doi.org/10.1016/j.compbiomed.2023.106544.
104. Chatzianastasis M, Vazirgiannis M, Zhang Z. Explainable multilayer graph neural network for cancer gene prediction. *Bioinformatics* 2023;**39**(11). https://doi.org/10.1093/bioinformatics/btad643.
105. Real KSD, Rubio A. Discovering the mechanism of action of drugs with a sparse explainable network. *EBioMedicine* 2023;**95**:104767. https://doi.org/10.1016/j.ebiom.2023.104767.
106. Chereda H, Bleckmann A, Menck K. *et al*. Explaining decisions of graph convolutional neural networks: patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer. *Genome Med* 2021;**13**:42. https://doi.org/10.1186/s13073-021-00845-7.
107. Mieth B, Rozier A, Rodriguez JA. *et al*. DeepCOMBI: explainable artificial intelligence for the analysis and discovery in genome-wide association studies. *NAR Genomics Bioinformatics* 2021;**3**(3). https://doi.org/10.1093/nargab/lqab065.
108. de Lima Camillo LP, Lapierre LR, Singh R. A pan-tissue DNA-methylation epigenetic clock based on deep learning. *NPJ Aging* 2022;**8**(1):4. https://doi.org/10.1038/s41514-022-00085-y.
109. Yap M, Johnston RL, Foley H. *et al*. Verifying explainability of a deep learning tissue classifier trained on RNA-seq data. *Sci Rep* 2021;**11**:2641. https://doi.org/10.1038/s41598-021-81773-9.
110. Benkirane H, Pradat Y, Michiels S. *et al*. CustOmics: A versatile deep-learning based strategy for multi-omics integration. *PLoS Comput Biol* 2023;**19**(3). https://doi.org/10.1371/journal.pcbi.1010921.
111. Yu T. AIME: Autoencoder-based integrative multi-omics data embedding that allows for confounder adjustments. *PLoS Comput Biol* 2022;**18**(1). https://doi.org/10.1371/journal.pcbi.1009826.
112. Sun T, Wei Y, Chen W. *et al*. Genome-wide association study-based deep learning for survival prediction. *Stat Med* 2020;**39**(30):4605-4620. https://doi.org/10.1002/sim.8743.
113. Magnusson R, Tegnér JN, Gustafsson M. Deep neural network prediction of genome-wide transcriptome signatures – beyond the black-box. *NPJ Syst Biol Appl* 2022;**8**:9. https://doi.org/10.1038/s41540-022-00218-9.
114. van de Leur RR, Bos MN, Taha K. *et al*. Improving explainability of deep neural network-based electrocardiogram interpretation using variational auto-encoders. *Eur Heart J Digit Health* 2022;**3**(3):390–404. https://doi.org/10.1093/ehjdh/ztac038.
115. Liu L, Meng Q, Weng C. *et al*. Explainable deep transfer learning model for disease risk prediction using high-dimensional genomic data. *PLoS Comput Biol* 2022;**18**(7). https://doi.org/10.1371/journal.pcbi.1010328.
116. Watson M, Hasan BAS, Al Moubayed N. Using model explanations to guide deep learning models towards consistent explanations for EHR data. *Sci Rep* 2022;**12**(1):

19899. https://doi.org/10.1038/s41598-022-24356-6. Erratum in: *Sci Rep*. 2023;**13**(1):1349. https://doi.org/10.1038/s41598-023-28610-3.

117. Jin J, Yu Y, Wang R. *et al*. iDNA-ABF: multi-scale deep biological language learning model for the interpretable prediction of DNA methylations. *Genome Biol* 2022;**23**:219. https://doi.org/10.1186/s13059-022-02780-1.
118. Jin Y, Ren Z, Wang W. *et al*. Classification of Alzheimer's disease using robust TabNet neural networks on genetic data. *Math Biosci Eng* 2023;**20**(5):8358-8374. https://doi.org/10.3934/mbe.2023366.
119. Liu G, Zeng H, Gifford DK. Visualizing complex feature interactions and feature sharing in genomic deep neural networks. *BMC Bioinformatics* 2019;**20**:401. https://doi.org/10.1186/s12859-019-2957-4.
120. Withnell E, Zhang X, Sun K. *et al*. XOmiVAE: an interpretable deep learning model for cancer classification using high-dimensional omics data. *Brief Bioinform* 2021;**22**(6). https://doi.org/10.1093/bib/bbab315.
121. Van den Broeck, Lykov A, Schleich M. *et al*. On the tractability of shap explanations. *J Artif Intell Res* 2022;**74**:851–86. https://doi.org/10.1613/jair.1.13283.
122. Sun T, Wei Y, Chen W. *et al*. Genome-wide association study-based deep learning for survival prediction. *Stat Med* 2020;**39**: 4605–20. https://doi.org/10.1002/sim.8743.
123. Märtens K, Yau C. BasisVAE: Translation-invariant feature-level clustering with Variational Autoencoders. In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS)*. *Proceedings of Machine Learning Research*. 2020;**108**:2928–37. Available from: https://proceedings.mlr.press/v108/martens20b.html.
124. Liang B, Gong H, Lu L. *et al*. Risk stratification and pathway analysis based on graph neural network and interpretable algorithm. *BMC Bioinform* **23**:394.
125. Ghafouri-Fard S, Taheri M, Omrani MD. *et al*. Application of single-nucleotide polymorphisms in the diagnosis of autism spectrum disorders: a preliminary study with artificial neural networks. *J Mol Neurosci* **68**:515–21.
126. Qin R, Mahal LK, Bojar D. Deep learning explains the biology of branched glycans from single-cell sequencing data. *iScience*. 2022;**25**(10):105163. https://doi.org/10.1016/j.isci.2022.105163.
127. Keyl P, Bischoff P, Dernbach G. *et al*. Single-cell gene regulatory network prediction by explainable AI. *Nucleic Acids Res*. 2023;**51**(4). https://doi.org/10.1093/nar/gkac1212.
128. Jin T, Nguyen ND, Talos F. *et al*. ECMarker: interpretable machine learning model identifies gene expression biomarkers predicting clinical outcomes and reveals molecular mechanisms of human disease in early stages. *Bioinformatics* **37**: 1115–24.
129. Chen RJ, Lu MY, Wang J. *et al*. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans Med Imaging* **41**: 757–70.
130. Søgaard A. Shortcomings of interpretability taxonomies for deep neural networks. In: *Proceedings of the 2022 CIKM Workshops*. 2022, p. 1–6.
131. Hayakawa J, Seki T, Kawazoe Y. *et al*. Pathway importance by graph convolutional network and shapley additive explanations in gene expression phenotype of diffuse large b-cell lymphoma. *PloS One* 2022;**17**:e0269570. https://doi.org/10.1371/journal.pone.0269570.
132. Nguyen ND, Jin T, Wang D. Varmole: a biologically drop-connect deep neural network model for prioritizing disease risk variants and genes. *Bioinformatics* 2021;**37**:1772–5. https://doi.org/10.1093/bioinformatics/btaa866.
133. Heinze-Deml C, Maathuis MH, Meinshausen N. Causal structure learning. *Annu Rev Stat Appl* 2018;**5**:371–91. https://doi.org/10.1146/annurev-statistics-031017-100630.
134. van Hilten A, van Rooij J, BIOS consortium. *et al*. Phenotype prediction using biologically interpretable neural networks on multi-cohort multi-omics data. *bioRxiv* 2023;**2023**:537073. https://doi.org/10.1101/2023.04.16.537073.
135. Esser-Skala W, Fortelny N. Reliable interpretability of biology-inspired deep neural networks. *NPJ Syst Biol Appl* 2023;**9**:50. https://doi.org/10.1038/s41540-023-00310-8.
136. Urbanowicz RJ, Kiralis J, Sinnott-Armstrong NA. *et al*. GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. *BioData mining* 2012;**5**:1–14. https://doi.org/10.1186/1756-0381-5-16.
137. Blumenthal DB, Viola L, List M. *et al*. Epigen: an epistasis simulation pipeline. *Bioinformatics* 2020;**36**:4957–9. https://doi.org/10.1093/bioinformatics/btaa245.
138. Yang W, Gu CC. A whole-genome simulator capable of modeling high-order epistasis for complex disease. *Genet Epidemiol* 2013;**37**:686–94. https://doi.org/10.1002/gepi.21761.
139. Jain S, Wallace BC. Attention is not explanation. In: Burstein J, Doran C, Solorio T (eds). *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume **1** (Long and Short Papers); 2019 Jun; Minneapolis, Minnesota. Association for Computational Linguistics, 2019, p. 3543–56. https://doi.org/10.18653/v1/N19-1357.
140. Bastings J, Filippova K. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In: Alishahi A, Belinkov Y, Chrupała G, Hupkes D, Pinter Y, Sajjad H (eds). *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*; 2020 Nov; Online. Association for Computational Linguistics, 2020, p. 149–55. https://doi.org/10.18653/v1/2020.blackboxnlp-1.14.
141. Li Y, Wu F-X, Ngom A. A review on machine learning principles for multi-view biological data integration. *Brief Bioinform* 2018;**19**:325–40.
142. Cui C, Yang H, Wang Y. *et al*. Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review. *Prog Biomed Eng* 2023;**5**(2):022001. https://doi.org/10.1088/2516-1091/acc2fe.
143. Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: a review. *Brief Bioinform* 2022;**23**:bbab569. https://doi.org/10.1093/bib/bbab569.