Received: 25 September 2022 | Revised: 18 March 2023 | Accepted: 6 April 2023

DOI: 10.1002/wics.1617

ADVANCED REVIEW

WIREs COMPUTATIONAL STATISTICS WILEY

A spectrum of explainable and interpretable machine learning approaches for genomic studies

Ashley Mae Conard[1,2,3] | Alan DenAdel[1] | Lorin Crawford[1,4,5]

[1]Center for Computational Molecular Biology, Brown University, Providence, Rhode Island, USA

[2]Department of Computer Science, Brown University, Providence, Rhode Island, USA

[3]Microsoft Research Redmond, Redmond, Washington, USA

[4]Microsoft Research New England, Cambridge, Massachusetts, USA

[5]Department of Biostatistics, Brown University, Providence, Rhode Island, USA

Correspondence
Lorin Crawford, Microsoft Research New England, Cambridge, MA, USA.
Email: lcrawford@microsoft.com

Funding information
David and Lucile Packard Foundation

Edited by: David Scott, Review Editor and Co-Editor-in-Chief

Abstract

The advancement of high-throughput genomic assays has led to enormous growth in the availability of large-scale biological datasets. Over the last two decades, these increasingly complex data have required statistical approaches that are more sophisticated than traditional linear models. Machine learning methodologies such as neural networks have yielded state-of-the-art performance for prediction-based tasks in many biomedical applications. However, a notable downside of these machine learning models is that they typically do not reveal how or why accurate predictions are made. In many areas of biomedicine, this "black box" property can be less than desirable—particularly when there is a need to perform *in silico* hypothesis testing about a biological system, in addition to justifying model findings for downstream decision-making, such as determining the best next experiment or treatment strategy. Explainable and interpretable machine learning approaches have emerged to overcome this issue. While explainable methods attempt to derive *post hoc* understanding of what a model has learned, interpretable models are designed to inherently provide an intelligible definition of their parameters and architecture. Here, we review the model transparency spectrum moving from black box and explainable, to interpretable machine learning methodology. Motivated by applications in genomics, we provide background on the advances across this spectrum, detailing specific approaches in both supervised and unsupervised learning. Importantly, we focus on the promise of incorporating existing biological knowledge when constructing interpretable machine learning methods for biomedical applications. We then close with considerations and opportunities for new development in this space.

This article is categorized under:

Statistical Models > Nonlinear Models
Applications of Computational Statistics > Genomics/Proteomics/Genetics
Applications of Computational Statistics > Computational and Molecular Biology

Ashley Mae Conard and Alan DenAdel contributed equally to this work.

KEYWORDS

explainability, genomics, interpretability, machine learning

1 | INTRODUCTION

A major focus in precision medicine has been to use computational tools to accurately predict disease outcomes and identify associated biomarkers for effective follow-up evaluation. Over the last two decades, linear models have been widely implemented to identify differentially expressed genes and enriched signaling pathways in functional genomics (Love et al., 2014; Nueda et al., 2014; Ritchie et al., 2015; Robinson et al., 2009), characterize complex trait architecture in genome-wide association studies (, 2010; Hayeck et al., 2015; Heckerman et al., 2016; Jiang et al., 2019; Kang et al., 2008; Korte et al., 2012; Lippert et al., 2011; Loh et al., 2018; Price et al., 2010; Runcie & Crawford, 2019; Zeng & Zhou, 2017; Zhou & Stephens, 2012), estimate the underlying generative model of gene networks (Karlebach & Shamir, 2008; Ma et al., 2018; Manno et al., 2018), and perform effective normalization and dimensionality reduction between studies performed across different time points, data collection sites, and tissue types (Hasin et al., 2017; Lähnemann et al., 2020). Part of the utility of these approaches is their ability to provide statistical significance measures such as *p*-values, posterior inclusion probabilities (PIPs), or Bayes factors which then can be used to facilitate downstream tasks (e.g., selecting which molecular mechanism to target with drugs or choosing which clinical interventions would be effective for a particular patient). Unfortunately, strict additive assumptions often hinder the performance of linear models; and the most powerful of these approaches rely on training algorithms that are computationally inefficient and unreliable for large-scale sets of data.

The continued advancement of imaging and sequencing technologies (Stephens et al., 2015) has provided the opportunity to integrate multimodal, nonparametric approaches as state-of-the-art tools within biological and clinical applications. Indeed, machine learning methods are well-known to have the ability to learn complex nonlinear patterns in data and they are often most powered in settings when large sets of training examples are available (LeCun et al., 2015). However, it has been heavily cited in the literature that many machine learning techniques suffer from a "black box" limitation in that they do not naturally carry out classic statistical hypothesis testing like linear models, which is critical for decision-making in precision medicine. One of the key characteristics that lead to better predictive performance for nonlinear algorithms is the automatic inclusion of higher-order interactions between features being put into the model (Crawford et al., 2018; Jiang & Reif, 2015). For example, neural networks leverage activation functions between layers that implicitly enumerate all possible (polynomial) interaction effects between input features (Demetci et al., 2021; Murdoch et al., 2019; Tsang, Cheng, & Liu, 2018; Tsang, Liu, et al., 2018; Wahba, 1990). This has been shown to make a difference in accurately predicting traits of model organisms where phenomena like epistasis (i.e., the interaction between multiple loci and/or genes) can play a large role in variation across species (Bellot et al., 2018; Runcie et al., 2021; Swain et al., 2016; Weissbrod et al., 2016). While this is a partial mathematical explanation for model improvement, in many biomedical applications, we often wish to know precisely which subsets of genomic features (e.g., variants, genes, and pathways) are most important in defining the architecture of a phenotype or disease outcome.

The main purpose of this manuscript is to review the considerable amount of methodological research that has been put toward developing more "explainable" and "interpretable" machine learning in computational biology. Throughout this paper, we will use the classic viewpoint that "explainability" has to do with the *post hoc* ability to use model parameters to justify predictions (also sometimes referred to as performing "variable importance" in certain areas of the literature) (Crawford et al., 2019; Lundberg & Lee, 2016, 2017; Ribeiro et al., 2016; Shrikumar et al., 2017); while "interpretability" is where a model inherently provides an intelligible definition to its parameters and architecture (Hira et al., 2019; Marcinkevics & Vogt, 2020; Shmueli, 2010). Both concepts can be divided into classes of methods that seek to achieve explainability or interpretability on either (i) a global scale, where the goal is to rank/select input features on their contributions to overall variation in an observed population, or (ii) on the local level, which aims to detail on how important a variant is to any particular individual in the dataset. Here, we will focus on describing global-scale approaches in neural networks with a particular motivation stemming from association mapping-based applications in genomics. Our main contribution in this review is to provide a comprehensive landscape of what we describe as the "transparency spectrum" for supervised and unsupervised learning algorithms as we move from black box, to explainable, and finally to interpretable methods (Figure 1).

We want to highlight that this review frames interpretability as taking a statistical model with a large number of parameters and mapping onto a space of sparse solutions. While this scope is commonly used in many scientific disciplines, we acknowledge that there are other ways to build interpretable methods (e.g., rule-based machine learning) (Wang et al., 2017). We also leave human interaction with models, also commonly known as human-in-the-loop machine learning, for another review (Lage et al., 2018; Mosqueira-Rey et al., 2022; Settles, 2009; Sverchkov & Craven, 2017). The precise manner in which models are considered interpretable is domain specific. Our focus is motivated by the fact that many genomic applications have more measured features than observed samples, thus rendering simple linear models to be underdetermined. Here, sparse solutions are useful in taking steps toward the ultimate goal of real-world decision-making. Specifically, in both wet-lab and clinical research, the objective is often to choose a small

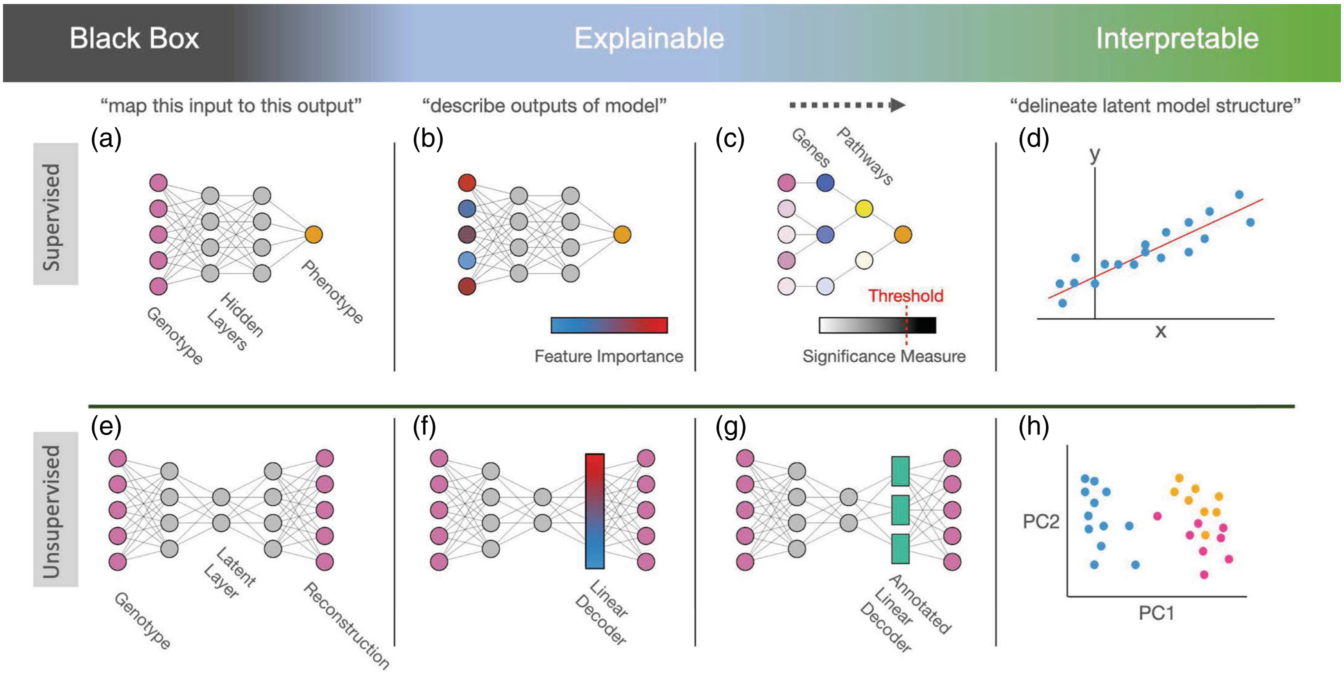

FIGURE 1 An overview of the model transparency spectrum: moving from black box to more interpretable methods in supervised and unsupervised learning. Supervised learning utilizes labeled datasets to learn a generalizable function to predict unobserved outcomes, while unsupervised machine learning discovers relationships from unlabeled data based on a given task. Here, we define an "explainable" model as being one that has the *post hoc* ability to use model parameters to justify predictions; however, "interpretable" methods inherently provide an intelligible definition of parameters. (a) Depiction of a black box neural network architecture with genotypes as inputs and phenotypes as output (Battle et al., 2015; Bell et al., 2011; Corsello et al., 2020; Rahman et al., 2019; Rakitsch & Stegle, 2016; Xiong et al., 2015). This model is said to be "fully connected" since all neurons from one layer are connected to every activation unit of the next layer. The downside of this formulation is that the classic notion of an effect size for each feature of the model is lost which prevents the ability to naturally carry out classic statistical hypothesis testing. (b) By integrating a *post hoc* method after training, the inputs of the model can be ranked by their influence on the phenotype, rendering the model explainable (Bourgeais et al., 2022; Crawford et al., 2018; Crawford et al., 2019; Elmarakeby et al., 2021; Gray-Davies et al., 2016; Huang et al., 2021; Lundberg & Lee, 2017; Samek et al., 2019; Shrikumar et al., 2017; Simonyan et al., 2013; Simonyan et al., 2014). (c) Interpretable supervised methods have three key components: (i) a motivating probabilistic model, (ii) the notion of an effect size for each input feature, and (iii) a significance metric for variable selection. One intuitive way of accomplishing this in machine learning is to use partially connected network architectures that are based on biological annotations or scientific knowledge (Cheng et al., 2022; Demetci et al., 2021; Kim et al., 2021; Ma & Wang, 1999; van Bergen et al., 2020; Videla Rodriguez et al., 2022). (d) Linear models are interpretable supervised methods because they naturally possess all components needed to carry out well-controlled hypothesis tests (Du et al., 2019; Molnar, 2020). (e) Autoencoders attempt to simultaneously encode input data in a lower dimensional representation using a neural network and decode that reduced representation into an approximation of the input (Kingma & Welling, 2014). (f) By constraining the decoder to be linear, *post hoc* methods can be used to explain the decoder as if it were a linear factor model (Svensson et al., 2020). (g) By incorporating biological annotations, sparse unannotated factors, and dense unannotated factors into a linear decoder, an autoencoder becomes interpretable by examining the subcomponents of the model (Rybakov et al., 2020). (h) Principal component analysis (PCA) is a classic linear dimensionality reduction technique. It still requires manual interpretation to understand the learned latent factors (Tipping & Bishop, 1999).

number of specific biomarkers or drug target candidates for follow-up. With this in mind, we position this review under the framework that a major goal of achieving interpretability in biology is to directly map from a model to an actionable set of sparse solutions.

Over the next three sections, we highlight the recent trend toward the development of more "biologically inspired" machine learning where interpretability is achieved by strategies such as imposing partially connected model architectures based on real-world annotations between features (Demetci et al., 2021; Elmarakeby et al., 2021; Lotfollahi et al., 2023; Rybakov et al., 2020) or by implementing physics-based loss functions to mirror processes that one would observe in nature (Karniadakis et al., 2021; Raissi et al., 2019). Finally, we provide a discussion on some important considerations about the need for model transparency in certain biological and clinical applications in practice and conclude with a review summary.

2 | SUPERVISED LEARNING

In this section, we give an overview of the "transparency spectrum" for supervised learning models in the literature (Figure 1a–d). Here, a popular position taken by previous studies assumes that an interpretable supervised method is made up of three key components: (i) a motivating probabilistic model, (ii) a notion of an effect size (or regression coefficient) for each genetic variant, and (iii) a statistical metric that determines marker significance according to a well-defined null hypothesis. In contrast, while an explainable supervised learning method may also provide these three components, the distinct difference is that the important measure that explainable methods provide does not naturally test against a calibrated null model.

3 | REVIEW OF GENERALIZED LINEAR MODELS

Supervised learning utilizes labeled datasets to learn a generalizable function to predict unobserved outcomes. Consider a biological study with N training observations (e.g., the number of individuals, cells, tissues) that have been phenotyped for some response $\mathbf{y}=(y_1,...,y_N)$. Assume that the i-th sample has been genotyped, sequenced, or profiled for J features $(x_{i1},...,x_{iJ})$ (e.g., gene expression, single nucleotide polymorphisms, pixels in a clinical image). Collectively, this results in a $N\times J$ matrix $\mathbf{X}$. Without loss of generality to the approaches that are described in this section, we will broadly refer to the input data in $\mathbf{X}$ as "genotypes" and the outcome response that we wish to model in $\mathbf{y}$ as the "phenotype". To build intuition about explainable and interpretable models in the supervised learning space, first, consider a standard generalized linear model (GLM)

$$g(\boldsymbol{\mu})=\boldsymbol{f}, \quad \boldsymbol{f}=\mathbf{X}\boldsymbol{\beta} \tag{1}$$

where $\mathbb{E}[\mathbf{y}|\mathbf{X}]=\boldsymbol{\mu}$ is the expected value of the phenotype, $g(\bullet)$ is a general link function, and $\boldsymbol{f}$ is an N-dimensional vector that is assumed to be a linear combination of each feature in $\mathbf{X}$ and their respective effects denoted by the J-dimensional vector $\boldsymbol{\beta}=(\beta_1,...,\beta_J)$ additive coefficients. Classically, the distribution of phenotypes being studied will determine the appropriate choice for the link function. For example, when the phenotype of interest is a continuous measurement [e.g., height (Loh et al., 2015; Orliac et al., 2022), treatment efficacy (Ritchie et al., 2015)], $g(\bullet)$ is set to the identity. In a classification problem, where the phenotype of interest is a binary label (e.g., case–control studies (Jiang et al., 2021; Wu et al., 2010)), $g(\bullet)$ is often chosen to be a logit or probit function. Lastly, in the case where the phenotype is a collection of discrete counts [e.g., RNA sequencing data (Love et al., 2014; Robinson et al., 2009)], then $g(\bullet)$ can be set to a log-link function.

A central goal of many biomedical applications is to jointly infer the true global effect and statistical significance of each genotype on the phenotype. One classic strategy for estimating the regression coefficients in Equation (1) is to use least squares where the phenotypic vector is projected onto the column space of the genotypic data $\widehat{\boldsymbol{\beta}}=\text{Proj}(\mathbf{X},\mathbf{y})=\mathbf{X}^{\dagger}\mathbf{y}$, with $\mathbf{X}^{\dagger}$ denoting the Moore-Penrose generalized inverse. We commonly refer to the vector $\widehat{\boldsymbol{\beta}}=\left(\widehat{\beta}_1,...,\widehat{\beta}_J\right)$ as the effect size estimates for each genotypic feature in the data set. Then, together with an estimated set of standard errors for each feature, it is common to assess the null hypothesis that the true effect of each genotype is equal to zero

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_A : \beta_j \neq 0. \tag{2}$$

The above results in p-values detailing the statistically significant relationship between each j-th input genotypic vector and the phenotype. Importantly, under this traditional framework, a well-calibrated model will produce uniformly distributed p-values and preserve type I error rates in the setting when the phenotypes are generated under the null hypothesis (i.e., zero effects from the genotype). Moreover, under the alternative hypothesis, various corrections have been established to adjust p-values derived from multiple statistical tests to correct for the increased probability of identifying false positives (Barber & Candès, 2015; Datta & Datta, 2005; Efron, 2008; Efron et al., 2001; Efron & Tibshirani, 2002; Greenland & Robins, 1991; Joo et al., 2016; Muralidharan, 2010; Sesia et al., 2020; Stephens, 2016).

With the improvement of sequencing and imaging technologies, datasets in many current biomedical applications have the ability to measure more (correlated) features than sampled observations (i.e., $J > N$). Due to the multi-collinear and overdetermined nature of these data, it is not possible to use the ordinary least squares solution reliably (Tibshirani, 1996). As a result, regularization techniques have been developed where a set of solutions for estimating the effect sizes of Equation (1) can be written as

$$\widetilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \ell(\mathbf{y}, \widehat{\boldsymbol{\mu}}) + \lambda\left[(1-\alpha)\,\|\boldsymbol{\beta}\|_1 + \alpha\,\|\boldsymbol{\beta}\|_2^2\right], \tag{3}$$

with $\ell(\mathbf{y}, \widehat{\boldsymbol{\mu}})$ denoting a loss function between the phenotypic vector $\mathbf{y}$ and the fitted values $\widehat{\boldsymbol{\mu}}$ based on the predictions made by the model. For regression problems, the loss function is typically the squared error loss; while, for applications with discrete classes or counts, one could use the logistic or Poisson loss, respectively. Here, $\|\bullet\|_1$ and $\|\bullet\|_2^2$ denote L_1 and L_2, respectively, λ represents a free regularization parameter, and the term α distinguishes the type of regularization to be used. Specifically, $\alpha = 0$ corresponds to the "Least Absolute Shrinkage and Selection Operator" or lasso solution, where the effect size of unimportant genotypic features are shrunken to be exactly zero (Tibshirani, 1996), $\alpha = 1$ equates to ridge regression (Hoerl & Kennard, 1970) where the effect size of unimportant genotypic features are shrunken toward zero, while $0 < \alpha < 1$ results in the grouped elastic net penalty (Zou & Hastie, 2005).

Classical hypothesis testing assumes the model is selected before observing the data. When regularized regression is used, hypothesis testing is typically conducted only on the selected variables. Maintaining the notion of the null hypothesis in Equation (2), it is also common to choose significant features based on the magnitude of the regularized coefficients stemming from Equation (3). In practice, this can be done a few ways, including permutation (Arbet et al., 2017) and covariance test statistics (Lockhart et al., 2014) that have been previously developed. This "double dipping" of the data, however, results in inflated p-values. Recently, methods have been developed for performing hypothesis testing while conditioning on the selection procedure (Lee et al., 2016; Taylor & Tibshirani, 2015; Tibshirani et al., 2016). For example, for linear models that have performed variable selection via forward selection or the lasso, the distribution of post-selective estimators can be determined exactly, allowing for computation of well-calibrated p-values for post-selective hypothesis testing (Lee et al., 2016; Tibshirani et al., 2016).

As an alternative to regularization, one can also perform variable selection within a Bayesian framework. In this paradigm, the coefficients of the generalized linear model in Equation (1) are treated as random variables with prior distributions that reflect how one believes genotypic effects are associated with the phenotype of interest. A common approach in genomics is to assume *a priori* that the distribution of null and non-null genotypes simply reflect the hypotheses in Equation (2)—that is, in the statistical sense, by assuming that input variables are either in or out of the final model. This can be formulated probabilistically by placing a spike-and-slab prior distribution on each of the J coefficients (Chipman et al., 2001; George & McCulloch, 1993; Ishwaran & Rao, 2005)

$$\beta_j \sim \pi \mathcal{N}\left(0, \sigma^2\right) + (1-\pi)\,\delta_0 \tag{4}$$

where π denotes the total proportion of genotypic measurements that have a nonzero effect on the phenotype of interest, σ^2 is a scale variance parameter of the "slab" proportion of the mixture, and δ_0 is a point mass at zero. To facilitate posterior computation and inference, many approaches also introduce a J-dimensional binary indicator variable $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J) \in \{0,1\}^J$ where they implicitly assume *a priori* that $\Pr\left[\gamma_j = 1\right] = \pi$. Intuitively, $\gamma_j = 1$ when $\beta_j \neq 0$ and the j-th genotypic variable is included in the model. We can derive marginal PIPs as statistical evidence that the effect of the j-th input does not follow the null hypothesis

$$\mathrm{PIP}(j) \equiv \Pr\left[\beta_j \neq 0 \mid \mathbf{y}, \mathbf{X}\right] \tag{5}$$

where we can define a candidate set or signature of important genotypic measures as those that satisfy $\{j : \mathrm{PIP}(j) > T\}$, for some threshold T. In practice, this threshold T may be chosen subjectively (Hoti & Sillanpää, 2006), selected through cross-validation to control for a desired family-wise error or false discovery rate (Stephens & Balding, 2009), derived with the goal to construct posterior credible intervals in the setting of highly collinear features (Du et al., 2021; Hormozdiari et al., 2014; Hutchinson et al., 2020; Wang et al., 2020; Zou et al., 2022), or simply taken to be $T = 0.5$ to obtain an equivalence of the Bayesian "median probability model" (Barbieri et al., 2021; Barbieri & Berger, 2004). For any set of significant features, further analyses may be carried out involving the relative costs of false positives and false negatives to make an explicitly reasoned decision about which measurements to pursue in downstream tasks.

Lastly, knockoffs were recently developed for performing variable selection while controlling the false discovery rate by constructing negative control variables (Barber & Candès, 2015; Candès et al., 2018; Sesia et al., 2020). For a set of genotypes $\mathbf{X}$ and a phenotype $\mathbf{y}$, knockoff variables $\widetilde{\mathbf{X}}$ are generated such that (i) $\mathrm{corr}(\mathbf{x}_j, \mathbf{x}_k) \approx \mathrm{corr}(\widetilde{\mathbf{x}}_j, \widetilde{\mathbf{x}}_k)$ (ii) $\mathrm{corr}(\mathbf{x}_j, \mathbf{x}_k) \approx \mathrm{corr}(\mathbf{x}_j, \widetilde{\mathbf{x}}_k)$, and (iii) $\mathbf{y}$ is not used to create $\widetilde{\mathbf{X}}$. A model is then constructed for predicting $\mathbf{y}$ from the combined dataset of $\mathcal{D} = \left[\mathbf{X}; \widetilde{\mathbf{X}}\right]$. Since the knockoff variables have no relationship to $\mathbf{y}$, they can be used to calibrate the false discovery rate adaptively for a feature importance measure because they represent the null distribution of the feature importance values.

Summary. Generalized linear regression is the standard supervised learning framework that allows biological researchers to estimate the effects of genotypic features on phenotypic variation. As sequencing and imaging technologies improve, so does the measured feature space, making it no longer possible to fit GLMs using least squares. Instead, regularization techniques and Bayesian shrinkage strategies have been developed to overcome issues that arise in high-dimensional data settings and to provide sparse solutions for downstream tasks. The advantage of GLMs is that they are computationally efficient, provide uncertainty estimates through standard errors and confidence intervals, are often robust to violations of their assumptions (such as non-normality and heteroscedasticity), and their coefficients are easy to interpret. Perhaps most importantly, GLMs provide a natural setup to perform rigorous hypothesis testing. For example, p-values characterize the statistical significance of the relationship between each genomic feature and the phenotype. Importantly, p-values are defined with respect to a null hypothesis (e.g., typically assuming that a feature has no effect or that there is no difference between two populations). Nevertheless, GLMs are restrictive since they assume some sort of linear relationship between features in the data; which is often underpowered in biological contexts where nonlinear effects can be drivers of variation between outcomes (e.g., Brown et al., 2014). Further, GLMs require full and manual enumeration of all biological interactions, which is impossible. In many cases, neural networks provide a robust framework to model the nonlinearities seen in biology and can easily enumerate many biological interactions. What is more, they are often more robust to violations of their assumptions. Further, probabilistic neural networks can provide more insight into the underlying relationships between the predictor variables and the response variable and the uncertainty in those relationships.

3.1 | Feedforward neural networks

In this section, we introduce relevant notation for general classes of feedforward neural networks for supervised learning. While there exist many nonlinear models, neural networks are well-known to have the ability to approximate complex biological systems. For simplicity, we will focus on multi-layer perceptrons throughout this paper; however, we also want to emphasize that the theoretical concepts we describe can also be applied broadly to other architectures (e.g., convolutional and graph neural networks, etc.). Formally, we can specify a K-layer neural network to mirror a generalized nonlinear model via the following

$$g(\boldsymbol{\mu}) = \boldsymbol{f}, \;\; \boldsymbol{f} = \mathbf{Z}_K \boldsymbol{\Theta}_K + \mathbf{b}_K, \;\; \cdots, \;\; \mathbf{Z}_k = h(\mathbf{Z}_{k-1}\boldsymbol{\Theta}_{k-1} + \mathbf{b}_{k-1}), \;\; \cdots, \;\; \mathbf{Z}_1 = h(\mathbf{X}\mathbf{W} + \mathbf{u}) \tag{6}$$

where, in addition to previous notation, $\mathbf{Z}_k$ denotes the matrix of nonlinear neurons from the k-th hidden layer with corresponding weight matrix $\boldsymbol{\Theta}_k$, $\mathbf{u}$ and $\mathbf{b}_k$ are deterministic biases that are produced during the network training phase for the input and k-th hidden layer, $h(\bullet)$ is a nonlinear activation function, and $\mathbf{W}$ is a matrix of weights for the input layer. Throughout this section, we will use H_k to represent the number of neurons in the k-th hidden layer such that the dimensions of $\mathbf{W}$ is $J \times H_1$ (i.e., the number of input genotypic features in $\mathbf{X}$ by the number of neurons in the first hidden layer). Here, we also assume the penultimate layer is of dimension $H_K = 1$ since neurons in the last layer feed into a single node which is used to estimate the latent function and make predictions (see Figure 1a).

There are a few key takeaways from the modeling framework in Equation (6). First, the ability of the neural network to model non-additive effects is primarily done through the choice of the nonlinear activation function $h(\bullet)$. A common practice in many biomedical applications is to use these nonlinear transformations to model the effects of phenomena like gene-by-gene and gene-by-environmental interactions on complex traits and diseases (Cheng et al., 2019; de los Campos et al., 2009; de los Campos et al., 2010; Morota & Gianola, 2014; Swain et al., 2016; Weissbrod et al., 2016; Weissbrod et al., 2019). For example, in statistical genetics, it has been shown that the Taylor series expansion of the Gaussian kernel function enumerates higher-order interaction terms between genetic variants (Crawford et al., 2018; Jiang & Reif, 2015), thus alleviating the computational burden of having to explicitly enumerate all possible combinations of interacting pairs in a given model (Crawford et al., 2017). One common choice for an activation function in neural networks is the rectified linear unit (ReLU) (Xu et al., 2015) family of operators which shares a similar property for implicitly encoding interactive effects. To see this, consider a general example of the ReLU function applied to the first hidden layer of Equation (6) where $h(\mathbf{XW}+\mathbf{u}) = \max\{\mathbf{0}, \mathbf{XW}+\mathbf{u}\}$. Under this formulation, the effect that any one genotypic feature has in determining the output of the ReLU activation jointly depends on the effects of all other genotypic measurements in the data. Hence, $h(\bullet)$ captures the dependence or interaction between the inputs in the function. The bias term $\mathbf{u}$ allows each node across hidden layers to change slope for different combinations of inputs. Theoretically, as more nodes and hidden layers are added to the network architecture (i.e., mirroring more "deep learning" methodologies), the model will have an even greater ability to account for non-additive variation (acting similarly to classic Gaussian process regression methods) (Lee et al., 2018).

The second key takeaway from the formulation in Equation (6) is that, when all weights in the matrices $\mathbf{W}, \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_K$ are nonzero, the neural network is said to be "fully-connected" since all neurons from one layer are connected to every activation unit of the next layer (Figure 1a). The downside of this formulation is that the classic notion of an effect size for each feature of the model is lost. In other words, if we let $\mathbf{w}_{j\bullet}$ be the set of weights in the j-th row of $\mathbf{W}$ corresponding to the j-th feature, then the effect of the j-th feature is only considered null during the model training when each element in the vector $\mathbf{w}_{j\bullet} = \mathbf{0}$. This has led to the development of many different approaches to overcome this limitation.

Summary. There are many benefits to using feedfoward neural networks including having the ability to model highly complex nonlinear relationships between input features, learning and extracting latent patterns from data without explicit feature engineering, and having the flexibility multiple data modalities for various tasks. Unfortunately, the concept of variable selection is lost within these models because, with a large number of fully connected hidden layers, there is no classic notion of an effect size for each feature. Recall that an advantage of GLMs is being able to produce significance metrics such as p-values that are both well calibrated and preserve type I error rates with respect to a null hypothesis. Despite the powerful ability of fully-connected neural networks to model nonlinear relationships and interactions between features, rigorous hypothesis is not as straightforward. Modern neural networks can also be highly over-parameterized which further contributes to difficult interpretations. Together, these issues motivate a growing body of work in the sub-field of explainable machine learning. Explainability methods often seek to develop importance metrics for both input features and model parameters as a proxy for regression effect sizes and significance testing.

3.2 | Explainable approaches for feedforward neural networks

Explainable machine learning methods employ auxiliary techniques where features are assessed for importance after model training is completed (Figure 1b). In this section, we examine several types of such explainable approaches which commonly provide a variable importance score that attempts to replicate the statistical notion of an effect size for nonlinear models. There are a wide variety of different variable importance scores in the literature. While many of these techniques share theoretical connections between them (Lundberg & Lee, 2016), it is also common to characterize these measures as broadly fitting into one of two categories. The first are "salience methods" (also commonly known as "saliency maps" (Simonyan et al., 2013, Simonyan et al., 2014)) which seek to describe the marginal effect of features on an output by comparing the fitted model to a baseline. The second class of explainable methods are known as "sensitivity scores" which quantify variable importance by measuring the amount of predictive accuracy that is lost when a particular feature is perturbed. Common examples in this second class of methods include relative centrality measures (Crawford et al., 2019; Paananen et al., 2019; Paananen et al., 2021; Woo et al., 2015), DeepLIFT (Shrikumar et al., 2017), and Shapley Additive Explanations (SHAP) (Lundberg & Lee, 2017), among others. Note that many of the approaches we discuss here can be used with any model (e.g., random forests, kernel regression machines) and not just neural networks.

Saliency maps are attribution methods that are usually used in imaging-based applications and are typically presented via heatmaps to help users delineate relevant pixels for image classifications made by neural networks. Scores of greater magnitude correspond to the regions that have the largest impact on the model predictions. One of the initial implementations of this method from Simonyan et al. (2013) calculates the gradient of the model loss function with respect to the input pixels for a class of interest. This approach has parallels to classic backpropagation given an image of interest and a classification task. The first step is to perform a forward pass with the data, followed by a backward pass to compute the gradient, and ends with visualizing the gradients as either positive, negative, or absolute values. While backpropagation typically does not change the inputs, the method from Simonyan et al. backpropagates back to the input layer to determine which pixels influence the output classification most, denoted by a classification score. Specifically, this predicted class score f_i for the class of interest $y_i = 1$ and with respect to an input image $\mathbf{x}_i$, is calculated as

$$f_i \approx \mathbf{x}_i^\top \mathbf{w} + u \qquad w_j = \frac{\partial f_i}{\partial x_{ij}}, \tag{7}$$

where the j-th element of the weight vector w_j is the partial derivative of the function f_i with respect to the j-th pixel in the image x_{ij}, and u is the model bias term (Simonyan et al., 2013). Equation (7) provides an effect size-like element which prioritizes the J pixels in order of importance for a given image class. Note that this class score can be thought of as taking the network's first-order Taylor expansion on the input. The backward pass to compute the gradients in Equation (7) can pose a challenge due to nonlinear units removing signs (i.e., similar to the ReLU activation function with positive or negative weights). Simonyan et al. devised a backpropagation technique to address this ambiguity. Consider a two-layer neural network where $\mathbf{x}_i$ is the input image and $\mathbf{z}_1$ denotes the single hidden layer. Then the backpropagation issue is mitigated by using the following update

$$\frac{\partial f_i}{\partial \mathbf{x}_i} = \frac{\partial f_i}{\partial \mathbf{z}_1} \times \mathbb{I}\{\mathbf{x}_i > 0\} \tag{8}$$

where $\mathbf{z}_1$ is the hidden neuron after running data only on a single sample, and $\mathbb{I}\{\bullet\}$ is the element-wise indicator function, resulting in one when the activation at the lower layer is nonnegative, and zero otherwise. Assigning global importance with saliency maps can be done by taking the mean absolute values of the gradients $\sum_{i=1}^{N} |\partial f_i / \partial x_{ij}| / N$. The associated absolute value of the gradients (i.e., coefficients) of each feature in the model's linear representation represents the feature ranking or importance across all samples in the dataset (Simonyan et al., 2013).

Another widely used technique for providing explainability in supervised learning models is DeepLIFT where the core idea is to compare the activation of a hidden neuron for a specific observation in the data to a "reference activation" (Shrikumar et al., 2017). In practice, DeepLIFT requires the user to specify the reference. Using the probabilistic setup in Equation (6), we say that a hidden neuron with inputs can be formulated by $\mathbf{z}_1 = h\left(\mathbf{x}_i^\top \mathbf{w} + u\right)$ with output variable f_i. Given some reference inputs, which we will denote by $\mathbf{x}_i^0$, the reference activation can be determined by $\mathbf{z}_1^0 = h\left(\left(\mathbf{x}_i^0\right)^\top \mathbf{w}^0 + u^0\right)$ which leads to a reference output f_i^0. For example, in biomedical applications, reference data could be determined by sequencing gene expression for healthy individuals rather than disease cases. If we define the "difference-from-reference" between outputs in the general form $\Delta t = t - t^0$, then DeepLIFT assigns scores as

$$\sum_{j=1}^{J} C_{\Delta x_{ij} \Delta f_i} = \Delta f_i. \tag{9}$$

Shrikumar et al. (2017) refer to the above as the "summation-to-delta property" where $C_{\Delta x_{ij} \Delta f_i}$ measures the amount of deviation from the true prediction f_i that is attributed by altering a given feature x_{ij}. Unlike the saliency map approach described in Equations (7) and (8), it is possible for a contribution score $C_{\Delta x_{ij}\Delta f_i} \neq 0$ even when the gradient $\partial f_i / \partial x_{ij} = 0$. Rather than working with gradients directly, DeepLIFT uses finite differences with respect to input perturbations via "multipliers" defined as

$$m_{\Delta x_{ij}\Delta f_i} = \frac{C_{\Delta x_{ij}\Delta f_i}}{\Delta x_{ij}}. \tag{10}$$

The authors develop a calculus for multipliers and define (i) rules for assigning contribution scores for linear layers, (ii) a rescaling rule for nonlinear activation functions, and (iii) a rule for dealing with saturation and thresholding. These rules allow them to calculate each $C_{\Delta x_{ij}\Delta f_i}$ and use it as a feature importance score.

While DeepLIFT chooses linearization root points based on reference value propagation through the network, there are other ways to determine the magnitude of a given feature's effect on a model's prediction. For example, Shapley Additive Explanations (SHAP) (Lundberg & Lee, 2017) assign feature importance weights based on game theoretic principles (Roth, 1988; Shapley, 1951). Shapley values essentially determine a payoff for all players when each player might have contributed more or less than the others when attempting to achieve the desired outcome. In genomics, this is done by considering all possible "coalitions" (or subsets) of genotypes $\mathcal{S} \subseteq \mathcal{J}$ and then simply comparing the performance of a model trained with and without the j-th genotype added to each coalition. Mathematically, this can be represented as the following

$$\phi_j = \sum_{\mathcal{S} \subseteq \mathcal{J}\setminus\{j\}} \left[\frac{|\mathcal{S}|!(J-|\mathcal{S}|-1)!}{|J|!}\right]\left(f_{\mathcal{S}\cup\{j\}} - f_{\mathcal{S}}\right) \tag{11}$$

where $|\mathcal{S}|$ is the cardinality of the given coalition, $|\mathcal{J}| = J$ is the dimensionality of all genotypes in the data, and $f_{\mathcal{S}\cup\{j\}}$ and $f_{\mathcal{S}}$ is the model fit with and without the j-th genotype added to the coalition set $\mathcal{S}$, respectively. Unfortunately, considering all possible combinations of measured features means that one must fit 2^J different models which are often computationally infeasible for many modern biomedical datasets. Given this challenge, approximate methods have been developed, including those such as DeepSHAP which utilize efficient Shapley value estimators and other rescaling rules to provide explanations for complex series of models (Chen et al., 2022; Lundberg & Lee, 2017). For instance, DeepLift can be used to compute Shapley values when the reference is taken to be the sample average (Lundberg & Lee, 2017).

Applying these auxiliary approaches to trained supervised black box models has been shown to contribute to the *post hoc* comprehension of what signal machine learning detects in many biomedical applications. For example, several studies have used saliency maps to detect genotypic aberrations and prognoses from histopathology slides (Adebayo et al., 2018; Fu et al., 2020; Kather et al., 2020). Maslova et al. (2020) used DeepLIFT to identify regulatory DNA sequences that are important when trying to infer cell type-specific chromatin accessibility. Lastly, Tasaki et al. (2020) used DeepSHAP to find biomarkers from the differential expression based on genome-wide binding sites on both genome promoters and RNA.

Summary. Taken together, explainable machine learning methods have gained prominence for understanding variable importance in nonlinear models (e.g., Lipton, 2018). Each of the explainable ML methods described in this section provide scores that are meant to serve as a proxy to the statistical notion of an effect size which is naturally given in traditional linear regression. Explainability is useful to glean insights into an otherwise black box method. Providing ranked lists that focus on prioritizing the most relevant features can help improve the usability and effectiveness of machine learning models (Stites et al., 2021). Nevertheless, without a framework for rigorous hypothesis testing with these *post hoc* methods, there is an inability to control for type I error rates. For example, Shapley values have been shown to be misleading in the presence of correlated features, although there have been extensions to mitigate this issue (Aas et al., 2021; Li et al., 2017; Tolosi & Lengauer, 2011). Moreover, since *post hoc* explainable approaches are performed after a model is trained, they can be unaware of the latent structure of black box algorithms and can make erroneous elucidations (Aas et al., 2021). More importantly, these types of misleading explanations can hinder decision-making and can contribute to distrust in machine learning (Herman, 2017). This is particularly relevant in the biomedical applications where the transparency of an underlying method is often crucial for making the most informed decisions for downstream tasks.

3.3 | Interpretable feedforward neural networks

Recently, there have been efforts to move beyond explainability by leveraging more classic group regularization and variable selection shrinkage priors to control the sparsity-inducing behavior of neural networks (Chen et al., 2020; Kassani et al., 2022; Lu et al., 2018; Song & Li, 2021). Recall that, in a fully connected feedforward model such as the one specified in Equation (6), the effect of the j-th genotype $\mathbf{x}_j$ is considered null when all weights in the j-th row of $\mathbf{W}$ are set to zero or $\mathbf{w}_{j\cdot} = \mathbf{0}$. Feng and Simon (2017) developed the "sparse input neural networks" (SPINN) which uses a "sparse group lasso" penalty to solve the following optimization problem

$$\widetilde{\mathbf{w}}_{j\cdot} = \arg\min_{\boldsymbol{\eta}} \ell\left(\mathbf{y}, \boldsymbol{f}_{\boldsymbol{\eta}}\right) + \lambda_0 \sum_{k=1}^{K} \| \boldsymbol{\Theta}_k \|_2^2 + \lambda\left[(1-\alpha) \| \mathbf{w}_{j\cdot} \|_1 + \alpha \| \mathbf{w}_{j\cdot} \|_2\right] \tag{12}$$

where, similar to the notation used in Equation (3), $\boldsymbol{\eta} = \{\mathbf{W}, \mathbf{u}, \boldsymbol{\Theta}_{1:K}, \mathbf{b}_{1:K}\}$ represents all of the parameters in the neural network, λ_0 and λ are free regularization parameters, and $\ell\left(\mathbf{y}, \boldsymbol{f}_{\boldsymbol{\eta}}\right)$ now denotes a loss function between the observed output $\mathbf{y}$ and the model prediction $\boldsymbol{f}_{\boldsymbol{\eta}}$. Feng and Simon (2017) describe three key components in this criterion. First, the ridge penalty $\| \boldsymbol{\Theta}_k \|_2^2$ controls the magnitude of the weights in the K hidden layers. The $0 < \alpha < 1$ is a mixture parameter that balances between the lasso and group lasso regularization schemes. When $\alpha = 0$, the group lasso encourages that all weights corresponding to the j-th genotypic variable are zero. Alternatively, when $\alpha = 1$, the lasso encourages sparsity across all weights in the input layer.

There are several ways to induce sparsity across a group of weights corresponding to a single input node using Bayesian shrinkage priors (Fortuin, 2022). For example, Ghosh, Yao, and Doshi-Velez (2019) perform general variable selection on features using horseshoe priors where they assume a conditionally independent Gaussian scale mixture distribution over each of the weights in a neural network

$$\mathbf{w}_{j\cdot} \sim \mathcal{N}\left(\mathbf{0}, \tau_j^2 \nu_0^2 \mathbf{I}\right), \quad \tau_j^2 \sim \mathcal{C}^+(0, a_\tau), \quad \nu_0^2 \sim \mathcal{C}^+(0, a_\nu) \tag{13}$$

with τ_j^2 being a genotype-specific (i.e., local) scale parameter, ν_0^2 is a scale parameter that is shared across all weights in the input layer (i.e., global), $\mathcal{C}^+(0, a)$ is used to denote the half-Cauchy distribution with $a > 0$, and $\mathbf{I}$ represents the identity matrix. The horseshoe prior has flat and heavy tails with an infinite spike at zero (Carvalho et al., 2009) which allows for genotypes with significantly large effects on the phenotype to not be penalized by any shrinkage. This is in contrast, for example, to the lasso and other sparse priors, which perform uniform shrinkage across all weights in the model. Ghosh, Yao, and Doshi-Velez (2019) show that by forcing all the weights corresponding to a single node to share scale parameters, Equation (13) induces sparsity at the genotype level and turns off the effect of genotypes that do not contribute to the variation of the phenotype. Carvalho et al. (2010) showed that the horseshoe prior also has the additional property of producing inclusion probability-like measures. Here, the authors describe that if we define a "shrinkage weight" $\gamma_j = 1/\left(1 + \tau_j^2\right)$, which lives on the unit interval $[0,1]$, then the quantity $1 - \gamma_j$ can play a similar role in probabilistic neural networks and can be used to select sets of genotypes that are most significantly associated with the phenotype.

As an alternative to the horseshoe prior, Cheng et al. (2022) proposed using a "grouped single-effect" shrinkage prior on the input weights of neural networks. Framed as a nonlinear extension to the "single effects regression" (SER) model by Wang et al. (2020), this method aims to overcome the challenge in high-dimensional biomedical settings where there are many correlated genotypic measurements but only a few directly affect the phenotype of interest. Here, the strategy is to build "credible sets" which determine a candidate subset of genotypic features that appear to have some sort of statistical association with the phenotype even though we are unclear as to which specific ones are truly causal. Keeping consistent with our notation, the grouped "single-effect" shrinkage prior for neural networks can be specified as

$$\mathbf{W} = \mathbf{A} \circ \boldsymbol{\Gamma}, \quad \mathbf{a}_{j\cdot} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}\right), \quad \boldsymbol{\gamma} \sim \mathrm{Mult}(1, \boldsymbol{\pi}). \tag{14}$$

where $\mathbf{a}_{j\cdot}$ is a H_1-dimensional row-vector of continuous weights in $\mathbf{A} = [\mathbf{a}_{1\cdot}, \ldots, \mathbf{a}_{J\cdot}]$, $\circ$ denotes the Hadamard product (i.e., element-wise multiplication) between two parameters, $\boldsymbol{\gamma}$ is a binary indicator that determines which genotype in $\mathbf{X}$

is to have a nonzero effect, $\boldsymbol{\Gamma}$ is a matrix that is H_1 copies of the binary vector $\boldsymbol{\gamma}$, and $\text{Mult}(m,\boldsymbol{\pi})$ denotes a multinomial distribution with m samples drawn with class probability $\boldsymbol{\pi}$. Note that by setting $m=1$, only one row in $\mathbf{W}$ (i.e., the weights corresponding to only one genotypic input) will have a nonzero entry for modeling the single effect. Hence, under this formulation, when $\gamma_j = 1$, the rest of the input genotypes are excluded from the model. Together, Cheng et al. (2022) refer to the model above as a "single-effect neural network" (SNN). The SNN resembles the SER model in that it assumes that only one input has an effect on the response and, thus, posterior inclusion probabilities can be computed via

$$\text{PIP}(j) \equiv \Pr\left[\mathbf{w}_{j\boldsymbol{\cdot}} \neq \mathbf{0} \mid \mathbf{y}, \mathbf{X}\right]. \tag{15}$$

which is then used to determine variable significance. This approach can also be extended to flexibly allow for many genotypes to have an effect on a phenotypic outcome by taking an ensemble of single-effect neural networks. A ρ-level credible set can be constructed from this method by simply sorting genotypes in descending order according to their PIPs and then including each feature into the set until the cumulative probability exceeds ρ (Du et al., 2021; Hormozdiari et al., 2014; Hutchinson et al., 2020; Wang et al., 2020; Zou et al., 2022).

3.3.1 | Biologically inspired feedforward neural networks

Each of the previous advances in sparse model development for machine learning has motivated researchers to develop customized neural network architectures that are inspired by biological systems. Here, rather than implementing fully connected and (often times) overparameterized models, these novel frameworks have partially connected architectures that are based on biological annotations in the literature or derived from other functional relationships that have been identified through experimental validation (Figure 1c). The biological system is governed by coordinated gene and pathway regulation, where interactions happen at various molecular levels. These relationships can be seamlessly integrated into neural networks such that their components and parameters have a true genomic interpretation. Partially connected neural networks are considered to be more interpretable because each node encodes some biological unit (e.g., single nucleotide polymorphisms, genes, or pathways) and each weight represents a known relationship between the corresponding entities. For example, Demetci et al. (2021) developed "Biologically Annotated Neural Networks" or BANNs which are a class of feedforward Bayesian models to study non-additive variation in complex traits from genome-wide association studies. Here, the input layer encodes SNP-level effects and the hidden layer models the effects among SNP-sets. The Deep GONet model by Bourgeais et al. (2021) used the Gene Ontology (GO) database to preserve the hierarchical relationships between cellular mechanisms and used the interpretable architecture for an explanation of cancer detection on three different levels: disease, subdisease, and on the individual patient-level. Similarly, Elmarakeby et al. (2021) developed P-NET to stratify prostate cancer patients by treatment status and evaluated molecular drivers of therapeutic resistance. In P-NET, the input layer represents multimodal measurements of genes, and each subsequent hidden layer represents different signaling cascades throughout the cell.

Biologically inspired neural networks can be probabilistically formulated by taking a slight reformulation of Equation (6). Assume that we have a set of predefined annotations $\{\mathcal{A}_1, \ldots, \mathcal{A}_{G_k}\}$ which are used to functionally group the different genotypes in $\mathbf{X}$ on the k-th genomic scale (with $k=0$ representing the input layer). We will use these annotations to construct a K partially connected network that represents K different scales of genomic units. Formally, we can specify the following

$$g(\boldsymbol{\mu}) = \boldsymbol{f}, \quad \cdots, \quad \boldsymbol{f} = \sum_{l=1}^{G_K} \mathbf{Z}_{Kl}\boldsymbol{\Theta}_{Kl} + \mathbf{b}_{Kl}, \quad \cdots, \quad \mathbf{Z}_1 = \sum_{g=1}^{G_0} h\left(\mathbf{X}_g \mathbf{W}_g + \mathbf{u}_g\right) \tag{16}$$

where, in addition to previous notation, $\mathbf{X}_g$ is the subset of genotypes that have been annotated for the g-th set on the input layer; $\mathbf{Z}_{kl}$ is the subset of hidden neurons that have been annotated for the l-th set on the k-th genomic scale; and $\mathbf{W}_g$ and $\boldsymbol{\Theta}_{kl}$ are the corresponding weight matrices, respectively. In practice, the weight matrices $\mathbf{W}_g = \mathbf{W} \circ \mathbf{M}_g$ and $\boldsymbol{\Theta}_{kl} = \boldsymbol{\Theta}_k \circ \mathbf{M}_{kl}$ are constructed by taking the Hadamard product with a binary mask matrix $\mathbf{M}$ to zero-out all the connections that do not exist in the set of annotations. It is worth noting that the hierarchical structure of the joint likelihood in Equation (16) can be seen as a nonlinear take on classical integrative and structural regression models frequently used in many areas of computational biology (Califano et al., 2012; Carbonetto & Stephens, 2013; Kichaev et al., 2019; van der Wijst et al., 2018; Wang et al., 2008; Yang et al., 2017; Zhu & Stephens, 2018).

The key innovation of biologically inspired neural networks is that they allow practitioners to perform multi-scale inference on multiple genomic levels (e.g., association mapping of SNPs and enrichment analyses of functional pathways), simultaneously. Similar to other classes of feedforward models, there are many ways to conduct inference within this framework. In the P-NET model, Elmarakeby et al. (2021) utilize DeepLIFT to prioritize a novel candidate gene signature of therapeutic resistance which they then validate with wet-lab experimentation (Equations (9) and (10)). The Deep GONet model uses regularization to identify important genes and ontologies (Equation (12)). In the BANNs method, Demetci et al. (2021) take a Bayesian approach by treating the weights of the input and hidden layers as random variables. Here, the authors specify scale-specific priors to reflect how effect size distributions of non-null genomic markers can take different forms depending on the phenotype of interest. Generally, these take on forms similar to the point mass mixture distributions commonly used in variable selection (e.g., Equation (4))

$$w_{j\bullet}^{(g)} \sim \pi_w \mathcal{N}\left(0,\sigma_w^2\right) + (1-\pi_w)\delta_0, \quad \theta_{k\bullet}^{(l)} \sim \pi_{\theta_k} \mathcal{N}\left(0,\sigma_{\theta_k}^2\right) + (1-\pi_{\theta_k})\delta_0 \tag{17}$$

with $w_{j\bullet}^{(g)}$ and $\theta_{k\bullet}^{(l)}$ being individual elements in the sparse weight matrices in $\mathbf{W}_g$ and $\mathbf{\Theta}_{kl}$, respectively. Additionally, π_w and π_{θ_k} denote the total proportion of genotypes and annotated sets that have a nonzero effect on the phenotype of interest. By modifying a variational inference algorithm assuming point-normal priors in multiple linear regression, Demetci et al. (2021) derive posterior probabilities that any weight of a given connection in the neural network is nonzero

$$\mathrm{PIP}(jg) \equiv \Pr\left[w_{j\bullet}^{(g)} \neq 0 \mid \mathbf{y},\mathbf{X}\right], \quad \mathrm{PIP}(kl) \equiv \Pr\left[\theta_{k\bullet}^{(l)} \neq 0 \mid \mathbf{y},\mathbf{X}\right]. \tag{18}$$

With biologically annotated units and the ability to perform statistical inference on explicitly defined parameters, biologically inspired methods like BANNs, Deep GONet, and P-NET represent a step toward fully interpretable extensions of neural networks to biomedical applications.

Summary. In recent years, there has been a movement toward controlling the sparsity-inducing behavior of neural networks through classic group regularization and variable selection shrinkage priors, rather than relying solely on *post hoc* explainability approaches. By inducing sparsity on the weights between nodes in the network, these methods allow for identification of genotypes that are most significantly associated with the variation of a phenotype, all while improving the interpretability, flexibility, and generalizability of the model. Biologically inspired feedforward neural networks make further improvements by ensuring that all components of the model have a contextual interpretation and offer a unified frameowrk for conducting multi-scale inference and biomarker discovery. One key disadvantage of this approach is that it depends on reliable domain knowledge to generate these interpretable architectures. When this level of real world evidence is not available, as is the case for many practical scientific problems, implementing this strategy can be extremely challenging. Inferring data-driven relationships between molecular features without any *a priori* knowledge can also be difficult (e.g., Demetci et al., 2021). As a result, many of the current biologically inspired methods treat neural network architectures as fixed based on a predefined set of annotations. As we continue to build more robust domain knowledge, the strength and quantity of labeled data will improve, leading to an opportunity for these biologically inspired methods to have greater impact.

4 | UNSUPERVISED LEARNING

In this section, we now give an overview of the "transparency spectrum" for a subset of unsupervised learning models. Unsupervised machine learning discovers relationships from unlabeled data based on a given task. In biomedical applications, these tasks can include dimensionality reduction over large genetic data sets to identify latent relationships between human populations (Novembre et al., 2008; Patterson et al., 2006), clustering to identify cell types with similar molecular profiles (Ringnér, 2008), visualization of diverse data modalities (Becht et al., 2019; Ringnér, 2008) and others. While supervised learning assumes that an interpretable model has three clearly defined components with the single task of learning a function to predict outcomes, it is more difficult to delineate these components within unsupervised models as these methods may be used to address one of many research questions. Here, we will focus on reviewing popular unsupervised learning approaches that perform dimensionality reduction and aim to interpret the learned latent space (i.e., the representation of the data with reduced dimensionality). Specifically, we will focus on autoencoders which have become widely used in many biomedical applications (Figure 1e–h). We will begin by reviewing principal component analysis, which can be thought of as a linear autoencoder, and then we will examine nonlinear autoencoders along the model transparency spectrum.

4.1 | Review of probabilistic PCA

The general goal of dimensionality reduction techniques is to find a low-dimensional representation of a high-dimensional dataset. The prototypical example of this approach is principal components analysis (PCA), commonly used as a preprocessing technique in genome-wide association studies (Patterson et al., 2006), the analysis of gene expression and functional genomics (Townes et al., 2019), and many other areas of biomedicine including clinical imaging (Mwangi et al., 2013). As in the supervised case, we will consider a biological study with N training observations profiled for J genotypic features. Here, we will use $\mathbf{x}_i = (x_{i1}, ..., x_{ij})$ to now represent a row-vector of these measurements for the i-th observation, resulting in an $N \times J$ matrix $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]$. Probabilistic PCA (Tipping & Bishop, 1999) assumes that, for each $\mathbf{x}_i$, there exists a normally distributed K-dimensional latent variable $\mathbf{z}_i$ such that

$$\mathbf{x}_i \mid \mathbf{z}_i \sim \mathcal{N}(\boldsymbol{\mu} + \mathbf{W}\mathbf{z}_i, \sigma^2\mathbf{I}) \quad \mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{19}$$

where the dimensionality $K < J$, $\boldsymbol{\mu}$ is the mean of the data, σ^2 is the error variance, and $\mathbf{W}$ is a $J \times K$ dimensional matrix that contains "principal axis directions". To draw an analogy with the supervised machine learning models we previously discussed, note that the principal directions in $\mathbf{W}$ define a linear function that relates the input data to its latent representation. If the input genotype data $\mathbf{X}$ is assumed to be mean-centered column-wise at zero, then the above simplifies to

$$\mathbf{x}_i \mid \mathbf{z}_i \sim \mathcal{N}(\mathbf{W}\mathbf{z}_i, \sigma^2\mathbf{I}), \tag{20}$$

and by marginalizing out the latent variable, we obtain

$$\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}). \tag{21}$$

Note that as $\sigma^2 \to 0$, we obtain classical principal component analysis where the sample covariance of the input genotype data can be represented by the product of the principal axes (Tipping & Bishop, 1999).

Probabilistic PCA models high dimensional data via a linear transformation to a lower dimensional latent space where data are assumed to be normally distributed. In the next section, we will generalize this framework to the variational autoencoder, which nonlinearly models data via lower dimensional latent spaces. PCA is probably the optimal method for performing linear dimensionality reduction with a given number of latent variables according to the Eckart-Young-Mirsky theorem (Eckart & Young, 1936). Since PCA is a linear method, it might naively be considered interpretable, yet it still requires practitioners to examine principal component (PC) directions manually. There is no built-in notion of statistical significance for selecting the number of PCs to use nor to test which genotypes are significant contributors to a particular PC. Although the Marchenko–Pastur distribution asymptotically describes the distribution of singular values of random matrices, this is not often used in practice as a null distribution (Johnstone, 2001; Marčenko & Pastur, 1967). Rather, practitioners will qualitatively choose the number of PCs using an "elbow plot" or by using the "jackstraw" resampling procedure for an empirical null distribution (Chung & Storey, 2015). Though a principal component will often be related to confounders like batch effects or population structure, it can represent meaningful scientific results. For example, a well-known result in population genetics is that a scatterplot of the first two principal components of SNP data in individuals of self-identified European ancestry is visually similar to a map of Europe (Novembre et al., 2008).

4.2 | Autoencoders

Autoencoders have become widely used tools in biomedicine and played key roles in tasks such as (Eraslan et al., 2019; Gayoso et al., 2022; Lopez et al., 2018) like imputation, batch correction, visualization, and clustering. Although these models are capable of expressing highly sophisticated data-generating processes, they are algorithms that can often be highly opaque and "black box" due to the often "entangled" nature of their learned latent spaces (Radford et al., 2016). This is dissimilar to PCA, which generates a new system of orthogonal axes that may have a meaning that is interpretable to the modeler. In fact, PCA can be viewed as a linear implementation of an autoencoder, while autoencoders are more general nonlinear methods for dimensionality reduction. We now survey examples of autoencoders that span the transparency spectrum between black box and interpretable. In the particular case of interpretable autoencoders, the interpretability comes from models that incorporate prior knowledge in the form of biological annotations.

An autoencoder is defined by two functions: an encoder $e:\mathcal{X}\rightarrow\mathcal{Z}$, and a decoder $d:\mathcal{Z}\rightarrow\mathcal{X}$ —both of which map between an input space $\mathcal{X}$ and a learned latent space $\mathcal{Z}$ (Hinton & Salakhutdinov, 2006). Keeping our notation consistent, we assume that the dimensionality of these spaces are $\dim(\mathcal{X})=J$ and $\dim(\mathcal{Z})=K$, respectively, with $K<J$. Taking the genotypes for a given observation $\mathbf{x}_i\in\mathcal{X}$, a lower-dimensional representation can be obtained as a function of the encoder $\mathbf{z}_i=e(\mathbf{x}_i)$ and the decoder is used to approximately reconstruct the data $\mathbf{x}_i\approx d(\mathbf{z}_i)$ (Goodfellow et al., 2016). The functions $e(\bullet)$ and $d(\bullet)$ are usually modeled using fully connected neural network architectures, and their corresponding weights are typically estimated using an algorithm based on stochastic gradient descent (Figure 1e). Intuitively, a successfully trained model aims to minimize the distance between the original and reconstructed data via some loss function

$$\ell(\mathbf{x}_i,d(\mathbf{z}_i))=\ell(\mathbf{x}_i,d(e(\mathbf{x}_i))) \tag{22}$$

where, similar to previous notation, $\ell\left(\mathbf{x}_i,\mathbf{x}_i'\right)$ denotes a loss function of choice. A common choice in the literature is to simply take this function to be the squared loss $\ell\left(\mathbf{x}_i,\mathbf{x}_i'\right)=\sum_i\|\mathbf{x}_i-\mathbf{x}_i'\|^2$ across all N observations in the data.

4.3 | Variational autoencoders

Autoencoders are effective at performing dimensionality reduction but the standard formulation has no distributional assumptions on the latent space. Consequently, this makes generating new data a challenge for autoencoders where, in theory, this would be done by randomly drawing a new $\mathbf{z}_i^*\in\mathcal{Z}$ and generating new points $\mathbf{x}_i^*=d\left(\mathbf{z}_i^*\right)\in\mathcal{X}$ by passing them through the decoder. Variational autoencoders (VAEs) were developed in order to make autoencoders more effective as generative models that can produce new data (Kingma & Welling, 2014). They solve the data generation problem by augmenting the modeling framework to assume that that variable in the latent space follows standard multivariate Gaussian distributions. This allows for a straightforward sampling scheme to generate new data. Under these assumptions, we can write the VAE similarly to probabilistic PCA where

$$\mathbf{x}_i\mid\mathbf{z}_i\sim\mathcal{N}\left(d(\mathbf{z}_i),\sigma^2\mathbf{I}\right),\quad \mathbf{z}_i\sim\mathcal{N}(\mathbf{0},\mathbf{I}). \tag{23}$$

Under this formulation, one can represent the encoder $\mathbf{z}_i=e(\mathbf{x}_i)$ as the conditional distribution $p_\eta(\mathbf{z}_i|\mathbf{x}_i)$ and the decoder $\mathbf{x}_i\approx d(\mathbf{z}_i)$ as $p_\eta(\mathbf{x}_i|\mathbf{z}_i)$, respectively, where once again we use $\boldsymbol{\eta}$ to represent the free parameters in the neural network architecture. Intuitively, a successfully trained model achieves a minimized loss function $\ell(\mathbf{x}_i,d(\mathbf{z}_i))$ between the original and reconstructed data by maximizing the marginal log-likelihood which takes on the form

$$\log p_\eta(\mathbf{x}_1,\ldots,\mathbf{x}_N)=\sum_{i=1}^{N}\log p_\eta(\mathbf{x}_i),\quad p_\eta(\mathbf{x})=\int p_\eta(\mathbf{x}_i|\mathbf{z}_i)p_\eta(\mathbf{z}_i)\mathbf{z}_i. \tag{24}$$

Notably, when autoencoders utilize (nonlinear) neural network architectures, optimization becomes difficult due to the intractability of the marginal likelihood. This in turn makes estimating the posterior distribution (i.e., the encoder) $p_\eta(\mathbf{z}_i|\mathbf{x}_i)=p_\eta(\mathbf{x}_i|\mathbf{z}_i)p_\eta(\mathbf{z}_i)/p_\eta(\mathbf{x}_i)$ also intractable. As an alternative approach, one can use variational inference to formulate a lower bound to the marginal log-likelihood. Here, the main idea is to replace the true posterior distribution $p_{\boldsymbol{\eta}}(\mathbf{z}_i|\mathbf{x}_i)$ with an approximating family of distributions $q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathbf{x}_i)$. It can be shown that finding a suitable approximation amounts to selecting the variational parameters $\boldsymbol{\phi}$ that minimize Kullback–Leibler (KL) divergence between $p_{\boldsymbol{\eta}}(\mathbf{z}_i|\mathbf{x}_i)$ and $q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathbf{x}_i)$. The intuition behind this objective becomes clear when we rewrite Equation (24) as

$$\log p_{\boldsymbol{\eta}}(\mathbf{x}_i) = \mathscr{L}(\boldsymbol{\eta},\boldsymbol{\phi};\mathbf{x}_i) + \mathrm{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathbf{x}_i) \,\|\, p_{\boldsymbol{\eta}}(\mathbf{z}_i|\mathbf{x}_i)\right). \tag{25}$$

Since the KL-divergence is a non-negative quantity and equates to zero if and only if $q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathbf{x}_i) = p_{\boldsymbol{\eta}}(\mathbf{z}_i|\mathbf{x}_i)$, the first term $\mathscr{L}(\boldsymbol{\eta},\boldsymbol{\phi};\mathbf{x}_i)$ denotes the (variational) lower bound to the marginal log-likelihood. Thus,

$$\begin{aligned} \log p_{\boldsymbol{\eta}}(\mathbf{x}_i) \geq \mathscr{L}(\boldsymbol{\eta},\boldsymbol{\phi};\mathbf{x}_i) &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathbf{x}_i)}\left[\log p_{\boldsymbol{\eta}}(\mathbf{x}_i,\mathbf{z}_i) - \log q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathbf{x}_i)\right] \\ &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathbf{x}_i)}\left[\log p_{\boldsymbol{\eta}}(\mathbf{x}_i|\mathbf{z}_i)\right] - \mathrm{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathbf{x}_i) \,\|\, p_{\boldsymbol{\eta}}(\mathbf{z}_i)\right). \end{aligned} \tag{26}$$

To conduct (posterior) inference, VAEs are usually implemented by taking the gradient of Equation (26) and optimizing the model with respect to both the variational parameters $\boldsymbol{\phi}$ and the generative network parameters $\boldsymbol{\eta}$. In practice, it is common to consider a stochastic gradient variational Bayes (SGVB) view on VAEs which yields the following general expression for the lower bound of the log-likelihood

$$\mathscr{L}(\boldsymbol{\eta},\boldsymbol{\phi};\mathbf{x}_i) = \frac{1}{L}\sum_{l=1}^{L} \log p_{\boldsymbol{\eta}}\left(\mathbf{x}_i|\mathbf{z}_{\boldsymbol{\phi}}^{(l)}\right) - \lambda \mathrm{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathbf{x}_i) \,\|\, p_{\boldsymbol{\eta}}(\mathbf{z}_i)\right). \tag{27}$$

The first term on the right-hand side is commonly referred to as the "reconstruction loss" (Ghosh, Sajjadi, et al., 2019; Tolstikhin et al., 2018); while the second term can then be viewed as a "regularized loss" function where the variational posterior distribution $q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathbf{x}_i)$ is being adjusted to approximate a prior distribution specified over the latent space $p_{\boldsymbol{\eta}}(\mathbf{z}_i)$ (Ghosh, Sajjadi, et al., 2020; Tolstikhin et al., 2018). In practice, the term $\lambda \geq 0$ represents a weight parameter for the KL term to balance the trade-off with reconstruction (Ghosh, Sajjadi, et al., 2020; Higgins et al., 2017). Lastly, $\mathbf{z}_{\boldsymbol{\phi}}^{(l)}$ is used to denote an empirical Monte Carlo sample of $\mathbf{z}_i$ which is computed using a re-parameterization trick of Gaussian distributions. By taking L Monte Carlo samples of $\mathbf{z}_{\boldsymbol{\phi}}^{(1)},\dots,\mathbf{z}_{\boldsymbol{\phi}}^{(L)}$, the reconstruction loss and the regularized loss in Equation (27) can be analytically computed.

One common final step for VAEs that are used in genomics is to have the final layer of the autoencoder output parameters of probability distributions rather than reconstructed or newly generated data. For example, the scVI model (Gayoso et al., 2022; Lopez et al., 2018) takes in single-cell gene expression measurements as input and outputs hyperparameters for a Gamma distribution that are used as parameters for a Poisson model to generate expression profiles of new single-cell data. In this setting

$$x'_{ij} \sim \mathrm{Poisson}(w_{ij}), \quad w_{ij} \sim \mathrm{Gamma}\left(\psi_{ij},\vartheta_j\right), \quad \boldsymbol{\psi}_i = d(\mathbf{z}_i), \quad \mathbf{z}_i \sim \mathcal{N}(\mathbf{0},\mathbf{I}) \tag{28}$$

where ϑ_j is a gene-specific over-dispersion parameter. Note that the scVI model includes additional variables for confounding such as library size which are not included here for simplicity (Lopez et al., 2018).

Summary. Without additional modeling assumptions, autoencoders can be considered complex models powered by black box algorithms. They can be useful for (nonlinear) dimensionality reduction-based tasks but do not yield a straightforward parametric way of generating new data (Carraro et al., 2020). Due to this limitation, so-called vanilla autoencoders are not often used in practice (particularly in biomedical applications). The variational autoencoder improves upon this framework by using scalable inference algorithms to approximate posterior estimates of the model parameters. By regularizing the distribution over the latent space toward a standard Gaussian, learned generative models by VAEs can simulate new data by simply generating Gaussian noise and passing it through the decoder function. A major limitation of VAEs is that they usually rely on fully connected architectures—as a result, the learned latent space often lacks a direct interpretation. In addition, latent variables can be "entangled" after training and may contain correlations with the training data rather than being independent (Radford et al., 2016).

4.4 | Explainable variational autoencoders

VAEs can be considered more interpretable than basic autoencoders due to the standard multivariate Gaussian structure imposed on their latent space, but these VAE latent spaces are not always directly interpretable and the nonlinear nature of the decoder can render explainability more challenging (Mathieu et al., 2019). As previously described, explainable machine learning methods employ auxiliary methods in order to understand the learned model parameters *post hoc*. Although the modeling goals of unsupervised learning are different than those of supervised learning, the objectives and methods of explainable unsupervised learning overlap considerably with explainable supervised learning. In this section, we will give two examples of explainable autoencoders used in genomics. The first applies feature importance scores on a standard (non-variational) autoencoder, while the second imposes additional structure on the VAE framework to increase its explainability (Figure 1f).

Kong et al. (2021) recently applied the DeepLIFT method for computing feature importance scores to a standard autoencoder and used the resulting gene scores to find a minimal set of genes that can be used to infer the expression of the genes not included in the minimal set. The goal of this model was to improve upon the L1000 assay, which is a tool for measuring the expression of 978 "landmark genes," and using those 978 measurements to infer the expression of 11,350 additional genes.

Svensson et al. (2020) recently developed the linearly decoded variational autoencoder (LDVAE) model, a class of autoencoder composed with a neural network encoder and a linear decoder. With the LDVAE model, one can represent the encoder $\mathbf{z}_i = e(\mathbf{x}_i)$ as the conditional distribution $p_{\boldsymbol{\eta}}(\mathbf{z}_i|\mathbf{x}_i)$ and the decoder $\mathbf{x}_i \approx \mathbf{W}\mathbf{z}_i$ as $p_{\boldsymbol{\eta}}(\mathbf{x}_i|\mathbf{z}_i)$, respectively, where $\mathbf{W}$ is a matrix rather than a nonlinear function. Once again, we use $\boldsymbol{\eta}$ to represent the free parameters in the neural network architecture. Just as in the standard VAE, a successfully trained model achieves a minimized loss function $\ell(\mathbf{x}_i, \mathbf{W}\mathbf{z}_i)$ between the original and reconstructed data with the full model taking on the following form

$$\mathbf{x}_i \mid \mathbf{z}_i \sim \mathcal{N}\left(\mathbf{W}\mathbf{z}_i, \sigma^2\mathbf{I}\right), \quad \mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{29}$$

Training again then seeks to maximize the lower bound specified in Equation (27). Note that this is quite similar to probabilistic PCA (see again Equations (20) and (29)), yet the structure of the latent space is learned jointly and the encoder can be a flexible nonlinear neural network. This "factor model" is claimed by Svensson et al. to be interpretable due to its linear nature.

Summary. Explainable autoencoders attempt to overcome the limitations of interpreting the latent space of VAEs by using model-agnostic feature importance metrics (e.g., DeepLIFT) and by making additional linear or sparse modeling assumptions induced by prior knowledge (Shrikumar et al., 2017). Nevertheless, similar to understanding principal component analysis, interpreting latent factors is a non-trivial task due to the requirement of investigating the meaning of the factors using domain specific knowledge and *post hoc* methods to choose important contributing variables. LDVAE does not have a built-in methodology for explainability, and it must rely on feature importance scores and other auxiliary methods. As with explainable methods for feedforward neural networks in supervised learning, attempts can be made to interrogate the internal workings of VAEs using sensitivity scores like relative centrality measures, DeepLIFT, Shapley Additive Explanations (SHAP), as well as with saliency methods. These explainability methods can be used to draw conclusions about otherwise black box models but, as with supervised learning, these must be used carefully as they may lead to users of the model to misinterpreting results (Stites et al., 2021).

4.5 | Interpretable variational autoencoders

To further increase the interpretability of VAEs, biological annotations have been used in order to impose structure in two models that incorporate a linear decoder like in LDVAE framework (Figure 1g). Seninge et al. (2021) developed the VAE Enhanced by Gene Annotations (VEGA), which uses a linear decoder that is sparsified using a mask matrix generated from annotated gene sets. Rybakov et al. (2020) developed an interpretable variational autoencoder that incorporates ideas from LDVAE and from f-scLVM, an interpretable latent variable model. In this section, we will focus on the Rybakov et al. modeling approach. As described in the previous section, we will let the encoder $\mathbf{z}_i = e(\mathbf{x}_i)$ be represented by the conditional distribution $p_\eta(\mathbf{z}_i|\mathbf{x}_i)$ and the decoder $\mathbf{x}_i \approx \mathbf{W}\mathbf{z}_i$ be represented as $p_\eta(\mathbf{x}_i|\mathbf{z}_i)$. Here, however, the matrix $\mathbf{W}$ is not just a dense matrix that results in a dense layer within a network. Instead, $\mathbf{W}$ has three contributing components:

- Annotated factors that come from gene-set databases. Due to the small number of genes in most gene sets relative to the total number of genes in an organism, these are typically sparse factors.
- Sparse factors those are unannotated and estimated through regularized inference.
- Dense unannotated and unregularized factors consisting of a large number of genes and estimated through standard inference.

With this in mind, the full biologically inspired model can be displayed by the following

$$\mathbf{x}_i \mid \mathbf{z}_i \sim \mathcal{N}\left(\mathbf{W}\mathbf{z}_i, \sigma^2\mathbf{I}\right), \quad \mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad \mathbf{W} = [\mathbf{A}; \mathbf{S}; \mathbf{U}] \tag{30}$$

where $\mathbf{A}$ represents the annotated factor matrix, $\mathbf{S}$ represents the sparse unannotated factor matrix, and $\mathbf{U}$ represents the dense unannotated factor matrix. Model training seeks to minimize the same loss in Equation (27) and parameters are learned using stochastic gradient descent.

A similar model called the "explainable programmable mapper" (expiMap) by Lotfollahi et al. (2023) also uses a linear decoder while incorporating biological annotations. Here, the authors additionally developed gene importance scores, latent score directions, and implement differential testing for gene programs in order to interpret the expiMap model. The gene importance scores are simply computed as the absolute value of the weight of the gene in a given program and the latent score directions are a heuristic used to determine whether a latent variable generally upregulates or downregulates genes. There are two heuristic measures for assessing how aligned a latent variable is toward up or down-regulation: one that is count-based and one that is sum-based

$$\gamma_j^{\text{sum}} = \text{sign}\left(\sum_k w_{kj}\right) \quad \gamma_j^{\text{count}} = \text{sign}\left(\sum_k \text{sign}(w_{kj})\right) \tag{31}$$

where w_{kj} is the weight of the k-th gene in the j-th gene program. Finally, to test the statistical hypotheses $H_0 : z_i^{(a)} > z_i^{(b)}$ versus $H_1 : z_i^{(a)} \le z_i^{(b)}$, where $z_i^{(a)}$ and $z_i^{(b)}$ are the i-th latent variables for observations in two different groups a and b (e.g., cases versus controls, different experimental conditions), expiMap computes Bayes factors of the form

$$\text{BF}_{10} = \frac{p(\mathcal{D}|H_1)}{p(\mathcal{D}|H_0)} = \frac{p(H_1|\mathcal{D})}{p(H_0|\mathcal{D})} \times \frac{p(H_0)}{p(H_1)} \tag{32}$$

where $\mathcal{D} = \left[\mathbf{X}^{(a)}; \mathbf{X}^{(b)}\right]$ is the collective input data from both groups. In practice, these quantities are computed analytically by assuming a Gaussian distribution over the latent space or approximated via Monte Carlo sampling.

Summary. In contrast to explainable autoencoders, interpretable autoencoders include more rigorous statistical tests for interpreting their parameters. This includes the strategy of using predefined sparse (or partially connected) architectures that allow users to readily understand the inner workings of the model. In particular, modelers can use annotations that they believe are likely *a priori* to be related to the problem of interest in order to design a model that can be probed and interrogated. One weakness that remains is that the current VAE literature solely focuses on the interpretation of the latent space and typically leaves the encoder and decoder unexamined. To remedy this limitation, users may use explainability methods as previously described. However, to reiterate, when explainability methods are used they may lead to conclusions that are divergent from the model itself.

Additional comments. Compared to developments in the field of interpretable supervised methods, the field of interpretable unsupervised methods is still relatively nascent. There are many opportunities for latent space interpretation by making models probabilistic, developing effect size measures, and performing hypothesis testing. Although we focused on latent space interpretation here, there are other unsupervised learning tasks to be addressed with more explainable and interpretable methods. Recently, there has been criticism of the over-interpretation of UMAP, t-SNE, and other methods for visualization of high dimensional data (Chari et al., 2021). Finally, a common task is post-selective inference. This refers to the task of conducting statistical inference after some form of a selective algorithm. In supervised learning, this refers to conducting statistical inference on estimated parameters determined via regularized methods like the lasso which, as previously mentioned in the review of supervised learning methods, results in inflated *p*-values using classical methods (Taylor & Tibshirani, 2015). In unsupervised learning, it refers to determining differences between the groups inferred by a clustering algorithm. Due to the fact that the clustering algorithm grouped dissimilar data into different clusters, it is expected *a priori* that there are differences between groups and so any *p*-values computed by comparing the groups are highly inflated due to "data double dipping" (Gao et al., 2022) resulting in increased type I error. Methods have been developed for conditioning on the particular selective algorithm used before hypothesis testing and this is still a highly active area of research (Chen & Witten, 2022; Gao et al., 2022; Neufeld et al., 2022).

5 | CONCLUSIONS AND FUTURE PERSPECTIVES

In this article, we reviewed the model transparency spectrum from black box machine learning methods to those that are explainable and interpretable, focusing on supervised and unsupervised approaches in the domain of computational biology. Where one decides to sit on the model transparency spectrum should depend on the application of interest. Black box models are effective when the goal is solely to make predictions without the need for understanding how those predictions were generated. Explainable models have the additional benefit of providing ranked feature lists which are useful for *in silico* hypothesis generation and for prioritizing downstream tasks such as determining which genes to target first in the best next set of experiments. However, in settings with more high-stakes decision-making (e.g., clinical diagnostics), more interpretable and transparent models are needed (Figure 2) (Hira et al., 2019; Rahman et al., 2019).

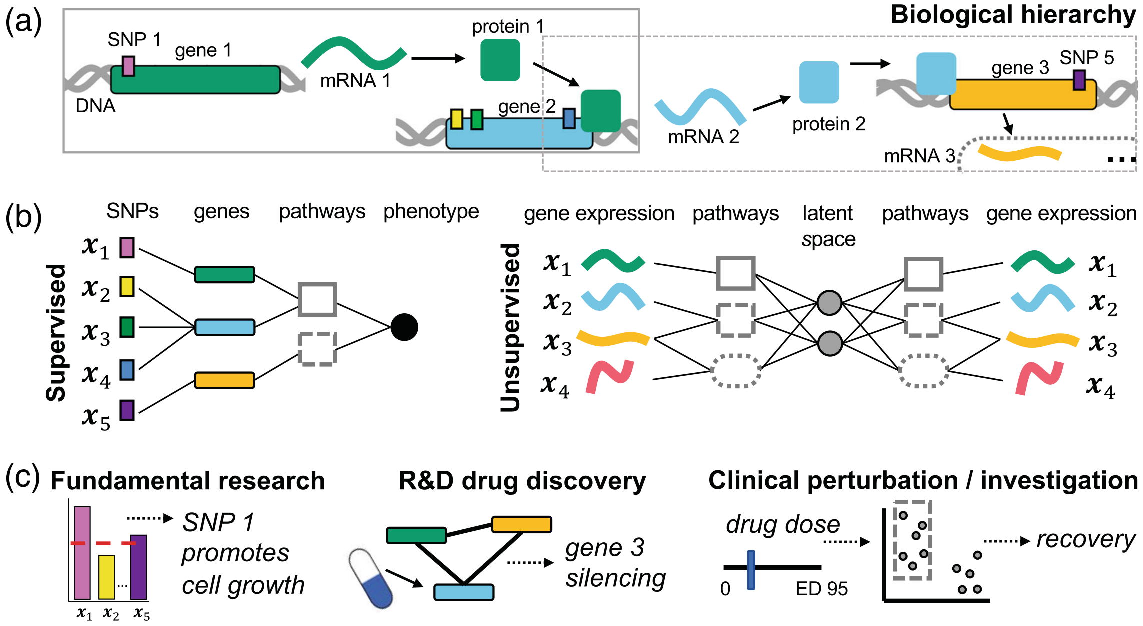

FIGURE 2 Interpretable models are essential for downstream decision-making in biomedical and clinical applications. (a) Biological processes are a series of molecular cascades of interactions between genes and gene products. (b) Such biological hierarchies can be modeled for example through biologically informed (i.e., partial) connections in supervised learning feed-forward neural networks or through unsupervised learning autoencoder frameworks. This modeling strategy has been used recently to identify genetic variants significantly contributing to phenotypic variation as well as to uncover latent structures that govern evolutionary processes among cell types.
(c) Interpretable models move beyond prediction tasks, performing variable selection to generate hypotheses and inform decision-making for downstream evaluation. Example areas of such applications include ranking genetic variants in fundamental research, deciding which molecules to test in drug development, and performing clinical intervention. ED, effective dose; R&D, research and development; SNP, single nucleotide polymorphism.

In biomedical machine learning, we are observing a shift in the field toward more interpretable frameworks that aim to mirror the underlying biological structure for the problem of interest. Such methods are valuable when needing to test against a null hypothesis, calibrate uncertainties, and control for type I error.

Here, we highlight the utility of techniques for developing explainable and interpretable models, albeit with some caveats (Nie et al., 2018; Sixt et al., 2020). While black box models and their explainable counterparts have proven to be valuable tools for generating hypotheses and performing prediction-based tasks, it should be noted that explanations may instill a false sense of confidence in users concerning a model's capabilities, thereby potentially increasing the acceptance of false positives and negatives (Stites et al., 2021). Indeed, complete assurance of accurate model interpretation is difficult to achieve; however, moving forward, it is suggested that researchers should seek to expand their focus beyond the interpretability spectrum and devise strategies to facilitate the identification of the best next action or recommendation based on model findings (i.e., providing summaries that are useful for the end user). For example, when aiming to discover novel biomarkers in genomics by employing *post hoc* methods to understand model behavior or to perform *in silico* hypothesis testing, one could also take into account the plausibility wet lab validation experiments or the viability of being targeted by a drug.

There are important caveats to consider when developing interpretable frameworks in biomedical applications. Firstly, we have recognized the potential power of incorporating known scientific knowledge during model development. Nevertheless, just as training data can contain biases, *a priori* annotations and domain knowledge can also bias what new associations scientists can discover. Secondly, what makes a model "explainable" or "interpretable" is domain specific, which introduces nuance and complexity based on contextual evidence. In biomedicine, this can make model evaluation a challenge in most contexts without performing wet-lab validation in a reproducible manner. One effective way to obtain, integrate, and verify domain knowledge is through user studies, where the expert is part of the model design and validation processes, sometimes through interviews or co-design sessions (Lage et al., 2018). Thirdly, human evaluation is crucial for understanding how machine learning models work and how they can be improved, as it can also introduce biases and inconsistencies if not done carefully (Herman, 2017; Molnar, 2020; Molnar et al., 2021; Stites et al., 2021). Lastly, the performance of any model is limited by the quality of the data that it is trained on. In other words, if the input-sequenced genotypes are flawed, any attempt to explain variation among the output phenotypes may also be compromised. Keeping these points in mind will enable the careful and robust development of biomedical machine learning across the model transparency spectrum.

AUTHOR CONTRIBUTIONS

Ashley Mae Conard: Conceptualization (supporting); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Alan DenAdel:** Conceptualization (supporting); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Lorin Crawford:** Conceptualization (lead); funding acquisition (equal); project administration (equal); supervision (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal).

FUNDING INFORMATION

We would like to the Editor, Associate Editor, and anonymous referees for their constructive comments. This research was supported by a David & Lucile Packard Fellowship for Science and Engineering awarded to Lorin Crawford. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funders.

CONFLICT OF INTEREST STATEMENT

The authors declare no competing interests.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Lorin Crawford https://orcid.org/0000-0003-0178-8242

RELATED WIREs ARTICLES

Genome-wide prediction of chromatin accessibility based on gene expression
Integrative clustering methods for multi-omics data
SAREV: A review on statistical analytics of single-cell RNA sequencing data

FURTHER READING

Newton, M. A. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, *5*(2), 155–176. https://doi.org/10.1093/biostatistics/5.2.155

REFERENCES

Aas, K., Jullum, M., & Løland, A. (2021. ISSN 0004-3702. https://www.sciencedirect.com/science/article/pii/S0004370221000539). Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, *298*, 103502. https://doi.org/10.1016/j.artint.2021.103502

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, *31*, 9525–9536.

Arbet, J., McGue, M., Chatterjee, S., & Basu, S. (2017). Resampling-based tests for lasso in genome-wide association studies. *BMC Genetics*, *18*(1), 70. https://doi.org/10.1186/s12863-017-0533-3

Barber, R. F., & Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, *43*(5), 2055–2085. https://doi.org/10.1214/15-aos1337

Barbieri, M. M., & Berger, J. O. (2004. ISSN 00905364). Optimal predictive model selection. *The Annals of Statistics*, *32*(3), 870–897 http://www.jstor.org/stable/3448578

Barbieri, M. M., Berger, J. O., George, E. I., & Ročková, V. (2021). The median probability model and correlated variables. *Bayesian Analysis*, *16*(4), 1085–1112.

Battle, A., Khan, Z., Wang, S. H., Mitrano, A., Ford, M. J., Pritchard, J. K., & Gilad, Y. (2015). Impact of regulatory variation from RNA to protein. *Science*, *347*(6222), 664–667.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., & Newell, E. W. (2019. ISSN 1546-1696). Dimensionality reduction for visualizing single-cell data using umap. *Nature Biotechnology*, *37*(1), 38–44. https://doi.org/10.1038/nbt.4314

Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., Gilad, Y., & Pritchard, J. K. (2011). DNA methylation patterns associate with genetic and gene expression variation in hapmap cell lines. *Genome Biology*, *12*(1), 1–13.

Bellot, P., de los Campos, G., & Pérez-Enciso, M. (2018). Can deep learning improve genomic prediction of complex human traits? *Genetics*, *21*(3), 809–819. https://doi.org/10.1534/genetics.118.301298

Bourgeais, V., Zehraoui, F., Hamdoune, M. B., & Hanczar, B. (2021). Deep GONet: Self-explainable deep neural network based on gene ontology for phenotype prediction from gene expression data. *BMC Bioinformatics*, *22*(S10), 455. https://doi.org/10.1186/s12859-021-04370-7

Bourgeais, V., Zehraoui, F., & Hanczar, B. (2022). Graphgonet: A self-explaining neural network encapsulating the gene ontology graph for phenotype prediction on gene expression. *Bioinformatics*, *38*(9), 2504–2511.

Brown, A. A., Buil, A., Viñuela, A., Lappalainen, T., Zheng, H.-F., Richards, J. B., Small, K. S., Spector, T. D., Dermitzakis, E. T., & Durbin, R. (2014). Genetic interactions affecting human gene expression identified by variance association mapping. *eLife*, *3*, e01381.

Califano, A., Butte, A. J., Friend, S., Ideker, T., & Schadt, E. (2012. ISSN 1546-1718). Leveraging models of cell regulation and gwas data in integrative network-based association studies. *Nature Genetics*, *44*(8), 841–847. https://doi.org/10.1038/ng.2355

Candès, E., Fan, Y., Janson, L., & Lv, J. (2018). Panning for gold: 'Model-x' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *8*(3), 551–577. https://doi.org/10.1111/rssb.12265

Carbonetto, P., & Stephens, M. (2013). Integrated enrichment analysis of variants and pathways in genome-wide association studies indicates central role for il-2 signaling genes in type 1 diabetes, and cytokine signaling genes in Crohn's disease. *PLoS Genetics*, *9*(10), 1–19. https://doi.org/10.1371/journal.pgen.1003770

Carraro, T., Polato, M., & Aiolli, F. (2020). A look inside the black-box: Towards the interpretability of conditioned variational autoencoder for collaborative filtering. In *Adjunct publication of the 28th ACM conference on user modeling, adaptation and personalization* (pp. 233–236). Association for Computing Machinery. ISBN 9781450379502. UMAP '20 Adjunct. https://doi.org/10.1145/3386392.3399305

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial intelligence and statistics* (pp. 73–80). PMLR.

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465–480.

Chari, T., Banerjee, J., & Pachter, L. (2021). The specious art of single-cell genomics. *bioRxiv*. https://doi.org/10.1101/2021.08.25.457696

Chen, H., Lundberg, S. M., & Lee, S.-I. (2022). Explaining a series of models by propagating shapley values. *Nature Communications*, *13*(1), 1–15.

Chen, Y., Gao, Q., Liang, F., & Wang, X. (2020). Nonlinear variable selection via deep neural networks. *Journal of Computational and Graphical Statistics*, *3*(2), 484–492. https://doi.org/10.1080/10618600.2020.1814305

Chen, Y. T., & Witten, D. M. (2022). Selective inference for k-means clustering. https://arxiv.org/abs/2203.15267

Cheng, L., Ramchandran, S., Vatanen, T., Lietzén, N., Lahesmaa, R., Vehtari, A., & Lähdesmäki, H. (2019). An additive gaussian process regression model for interpretable non-parametric analysis of longitudinal data. *Nature Communications*, *10*, 1798. https://doi.org/10.1038/s41467-019-09785-8

Cheng, W., Ramachandran, S., & Crawford, L. (2022). Uncertainty quantification in variable selection for genetic fine-mapping using Bayesian neural networks. *iScience*, *25*(7), 104553.

Chipman, H., George, E. I., & McCulloch, R. E. (2001). The practical implementation of bayesian model selection. In *Institute of Mathematical Statistics Lecture notes-monograph series* (pp. 65–116). Institute of Mathematical Statistics. https://doi.org/10.1214/lnms/1215540964

Chung, N. C., & Storey, J. D. (2015. ISSN 1367-4803). Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics*, *31*(4), 545–554. https://doi.org/10.1093/bioinformatics/btu674

Corsello, S. M., Nagari, R. T., Spangler, R. D., Rossen, J., Kocak, M., Bryan, J. G., Humeidi, R., Peck, D., Wu, X., Tang, A. A., Wang, V. M., Bender, S. A., Lemire, E., Narayan, R., Montgomery, P., Ben-David, U., Garvie, C. W., Chen, Y., Rees, M. G., … Golub, T. R. (2020). Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nature Cancer*, *1*(2), 235–248.

Crawford, L., Flaxman, S. R., Runcie, D. E., & West, M. (2019). Variable prioritization in nonlinear black box methods: A genetic association case study. *The Annals of Applied Statistics*, *13*(2), 958–989. https://doi.org/10.1214/18-AOAS1222

Crawford, L., Wood, K. C., Zhou, X., & Mukherjee, S. (2018). Bayesian approximate kernel regression with variable selection. *Journal of the American Statistical Association*, *113*(524), 1710–1721.

Crawford, L., Zeng, P., Mukherjee, S., & Zhou, X. (2017). Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *PLoS Genetics*, *13*(7), e1006869. doi:10.1371/journal.pgen.1006869

Datta, S., & Datta, S. (2005). Empirical Bayes screening of many p-values with applications to microarray studies. *Bioinformatics*, *21*(9), 1987–1994. https://doi.org/10.1093/bioinformatics/bti301

de los Campos, G., Gianola, D., Rosa, G. J. M., Weigel, K. A., & Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genetics Research*, *92*(4), 295–308. https://doi.org/10.1017/S0016672310000285

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., & Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, *182*(1), 375–385. https://doi.org/10.1534/genetics.109.101501

Demetci, P., Cheng, W., Darnell, G., Zhou, X., Ramachandran, S., & Crawford, L. (2021). Multi-scale inference of genetic trait architecture using biologically annotated neural networks. *PLoS Genetics*, *17*(8), e1009754.

Du, M., Andersen, S. L., Perls, T. T., & Sebastiani, P. (2021). Bayesian variable selection utilizing posterior probability credible intervals. *medRxiv*. https://doi.org/10.1101/2021.01.13.21249759

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, *63*(1), 68–77.

Eckart, C., & Young, G. (1936. ISSN 1860-0980). The approximation of one matrix by another of lower rank. *Psychometrika*, *1*(3), 211–218. https://doi.org/10.1007/BF02288367

Efron, B. (2008). Microarrays, empirical bayes and the two-groups model. *Statistical Science*, *23*(1), 1–22. https://doi.org/10.1214/07-STS236

Efron, B., & Tibshirani, R. (2002). Empirical bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, *23*(1), 70–86. https://doi.org/10.1002/gepi.1124

Efron, B., Tibshirani, R., Storey, J. D., & Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, *96*(456), 1151–1160. https://doi.org/10.1198/016214501753382129

Elmarakeby, H. A., Hwang, J., Arafeh, R., Crowdis, J., Gang, S., Liu, D., AlDubayan, S. H., Salari, K., Kregel, S., Richter, C., Arnoff, T. E., Park, J., Hahn, W. C., & van Allen, E. M. (2021). Biologically informed deep neural network for prostate cancer discovery. *Nature*, *598*(7880), 348–352.

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., & Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications*, *10*, 390. https://doi.org/10.1038/s41467-018-07931-2

J. Feng and N. Simon. Sparse-input neural networks for high-dimensional nonparametric regression and classification, 2017. https://arxiv.org/abs/1711.07592

Fortuin, V. (2022). Priors in bayesian deep learning: A review. *International Statistical Review*, *90*, 563–591.

Fu, Y., Jung, A. W., Torne, R. V., Gonzalez, S., Vöhringer, H., Shmatko, A., Yates, L. R., Jimenez-Linan, M., Moore, L., & Gerstung, M. (2020). Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer*, *1*(8), 800–810.

Gao, L. L., Bien, J., & Witten, D. (2022). Selective inference for hierarchical clustering. *Journal of the American Statistical Association*, 1–27. https://doi.org/10.1080/01621459.2022.2116331

Gayoso, A., Lopez, R., Xing, G., Boyeau, P., Amiri, V. V. P., Hong, J., Wu, K., Jayasuriya, M., Mehlman, E., Langevin, M., Liu, Y., Samaran, J., Misrachi, G., Nazaret, A., Clivio, O., Xu, C., Ashuach, T., Gabitto, M., Lotfollahi, M., … Yosef, N. (2022. ISSN 1546-1696). A python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology*, *4*(2), 163–166. https://doi.org/10.1038/s41587-021-01206-w

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881–889. https://doi.org/10.1080/01621459.1993.10476353

Ghosh, P., Sajjadi, M. S. M., Vergari, A., Black, M., & Schölkopf, B. (2020). From variational to deterministic autoencoders. In *International Conference on Learning Representations*.

Ghosh, S., Yao, J., & Doshi-Velez, F. (2019). Model selection in Bayesian neural networks via horseshoe priors. *Journal of Machine Learning Research*, *2*(182), 1–46 http://jmlr.org/papers/v20/19-236.html

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press http://www.deeplearningbook.org

Gray-Davies, T., Holmes, C. C., & Caron, F. (2016). Scalable Bayesian nonparametric regression via a Plackett-Luce model for conditional ranks. *Electronic Journal of Statistics*, *1*(2), 1807.

Greenland, S., & Robins, J. M. (1991). Empirical-Bayes adjustments for multiple comparisons are sometimes useful. *Epidemiology*, *2*(4), 244–251. https://doi.org/10.1097/00001648-199107000-00002

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology*, *18*(1), 83. https://doi.org/10.1186/s13059-017-1215-1

Hayeck, T. J., Zaitlen, N. A., Loh, P.-R., Vilhjalmsson, B., Pollack, S., Gusev, A., Yang, J., Chen, G.-B., Goddard, M. E., Visscher, P. M., Patterson, N., & Price, A. L. (2015). Mixed model with correction for case–control ascertainment increases association power. *The American Journal of Human Genetics*, *96*(5), 720–730. https://doi.org/10.1016/j.ajhg.2015.03.004

Heckerman, D., Gurdasani, D., Kadie, C., Pomilla, C., Carstensen, T., Martin, H., Ekoru, K., Nsubuga, R. N., Ssenyomo, G., Kamali, A., Kaleebu, P., Widmer, C., & Sandhu, M. S. (2016). Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proceedings of the National Academy of Sciences*, *113*(27), 7377–7382. https://doi.org/10.1073/pnas.1510497113

B. Herman. The promise and peril of human evaluation for model interpretability. *arXiv preprint arXiv:1711.07414*, 2017.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., & Lerchner, A. (2017). Beta-vae: Learning basic visual concepts with a constrained variational framework. *Iclr*, *2*(5), 6.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507. https://doi.org/10.1126/science.1127647

Hira, M. T., Razzaque, M., Angione, C., Scrivens, J., Sawan, S., & Sarker, M. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(1), 206–2014.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B., & Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. *Genetics*, *198*(2), 497–508. https://doi.org/10.1534/genetics.114.167908

Hoti, F., & Sillanpää, M. J. (2006). Bayesian mapping of genotype expression interactions in quantitative and qualitative traits. *Heredity*, *97*(1), 4–18. https://doi.org/10.1038/sj.hdy.6800817

Huang, X., Huang, K., Johnson, T., Radovich, M., Zhang, J., Ma, J., & Wang, Y. (2021). Parsvnn: Parsimony visible neural networks for uncovering cancer-specific and drug-sensitive genes and pathways. *NAR Genomics and Bioinformatics*, *3*(4), lqab097.

Hutchinson, A., Watson, H., & Wallace, C. (2020). Improving the coverage of credible sets in bayesian genetic fine-mapping. *PLoS Computational Biology*, *16*(4), e1007829.

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, *33*(2), 730–773. https://doi.org/10.1214/009053604000001147

Jiang, L., Zheng, Z., Fang, H., & Yang, J. (2021). A generalized linear mixed model association tool for biobank-scale data. *Nature Genetics*, *53*(11), 1616–1621. https://doi.org/10.1038/s41588-021-00954-4

Jiang, L., Zheng, Z., Qi, T., Kemper, K. E., Wray, N. R., Visscher, P. M., & Yang, J. (2019). A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics*, *51*(12), 1749–1755. https://doi.org/10.1038/s41588-019-0530-8

Jiang, Y., & Reif, J. C. (2015). Modeling epistasis in genomic selection. *Genetics*, *201*(2), 759–768. https://doi.org/10.1534/genetics.115.177907

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, *29*(2), 295–327. https://doi.org/10.1214/aos/1009210544

Joo, J. W. J., Hormozdiari, F., Han, B., & Eskin, E. (2016). Multiple testing correction in linear mixed models. *Genome Biology*, *17*(1), 62. https://doi.org/10.1186/s13059-016-0903-6

Kang, H. M., Sul, J. H., S. K. Service, Zaitlen, N. A., Kong, S. Y., Freimer, N. B., Sabatti, C., & Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, *42*(4), 348–354. https://doi.org/10.1038/ng.548

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, *178*(3), 1709–1723. https://doi.org/10.1534/genetics.107.080101

Karlebach, G., & Shamir, R. (2008). Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, *9*(10), 770–780.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, *3*(6), 422–440. https://doi.org/10.1038/s42254-021-00314-5

Kassani, P. H., Lu, F., Guen, Y. L., Belloy, M. E., & He, Z. (2022). Deep neural networks with controlled variable selection for the identification of putative causal genetic variants. *Nature Machine Intelligence*, *4*, 761–771. https://doi.org/10.1038/s42256-022-00525-0

Kather, J. N., Heij, L. R., Grabsch, H. I., Loeffler, C., Echle, A., Muti, H. S., Krause, J., Niehues, J. M., Sommer, K. A., Bankhead, P., Kooreman, L. F. S., Schulte, J. J., Cipriani, N. A., Buelow, R. D., Boor, P., Ortiz-Brüchle, N., Hanby, A. M., Speirs, V., Kochanny, S., … Luedde, T. (2020). Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer*, *1*(8), 789–799.

Kichaev, G., Bhatia, G., Loh, P.-R., Gazal, S., Burch, K., Freund, M. K., Schoech, A., Pasaniuc, B., & Price, A. L. (2019). Leveraging polygenic functional enrichment to improve GWAS power. *The American Journal of Human Genetics*, *104*(1), 65–75. https://doi.org/10.1016/j.ajhg.2018.11.008

Kim, Q., Ko, J.-H., Kim, S., Park, N., & Jhe, W. (2021). Bayesian neural network with pretrained protein embedding enhances prediction accuracy of drug-protein interaction. *Bioinformatics*, *37*(20), 3428–3435.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. In Y. Bengio & Y. LeCun (Eds.), *2nd International Conference on learning representations*, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings http://arxiv.org/abs/1312.6114

Kong, L., Chen, Y., Xu, F., Xu, M., Li, Z., Fang, J., Zhang, L., & Pian, C. (2021. ISSN 1471-2105). Mining influential genes based on deep learning. *BMC Bioinformatics*, *22*(1), 27. https://doi.org/10.1186/s12859-021-03972-5

Korte, A., Vilhjálmsson, B. J., Segura, V., Platt, A., Long, Q., & Nordborg, M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics*, *44*(9), 1066–1071. https://doi.org/10.1038/ng.2376

Lage, I., Ross, A. S., Kim, B., Gershman, S. J., & Doshi-Velez, F. (2018). Human-in-the-loop interpretability prior. In *Proceedings of the 32nd International Conference on neural information processing systems*, NIPS'18 (pp. 10180–10189). Curran Associates Inc.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S.-O., Aparicio, S., Baaijens, J., Balvert, M., de Barbanson, B., Cappuccio, A., ... Schönhuth, A. (2020). Eleven grand challenges in single-cell data science. *Genome Biology*, *21*(1), 31. https://doi.org/10.1186/s13059-020-1926-6

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, J., Sohl-dickstein, J., Pennington, J., Novak, R., Schoenholz, S., & Bahri, Y. (2018). Deep neural networks as Gaussian processes. In *International conference on learning representations* https://openreview.net/forum?id=B1EA-M-0Z

Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, *44*(3), 907–927. https://doi.org/10.1214/15-aos1371

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, *5*(6), 1–45.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., & Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods*, *8*(10), 833–835. https://doi.org/10.1038/nmeth.1681

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57.

Lockhart, R., Taylor, J., Tibshirani, R. J., & Tibshirani, R. (2014). A significance test for the lasso. *The Annals of Statistics*, *42*(2), 413–468. https://doi.org/10.1214/13-aos1175

Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P., & Price, A. L. (2018). Mixed-model association for biobank-scale datasets. *Nature Genetics*, *5*(7), 906–908. https://doi.org/10.1038/s41588-018-0144-6

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., Patterson, N., & Price, A. L. (2015. ISSN 1546-1718). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, *47*(3), 284–290. https://doi.org/10.1038/ng.3190

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, *15*, 1053–1058. https://doi.org/10.1038/s41592-018-0229-2

Lotfollahi, M., Rybakov, S., Hrovatin, K., Hediyeh-zadeh, S., Talavera-López, C., Misharin, A. V., & Theis, F. J. (2023. ISSN 1476-4679). Biologically informed deep learning to query gene programs in single-cell atlases. *Nature Cell Biology*, *25*(2), 337–350. https://doi.org/10.1038/s41556-022-01072-x

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with deseq2. *Genome Biology*, *15*(12), 1–21.

Lu, Y., Fan, Y., Lv, J., & Stafford Noble, W. (2018). Deeppink: Reproducible feature selection in deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 31). Curran Associates, Inc https://proceedings.neurips.cc/paper/2018/file/29daf9442f3c0b60642b14c081b4a556-Paper.pdf

Lundberg, S. M., & Lee, S. (2016). An unexpected unity among methods for interpreting model predictions. *CoRR*, 7478 http://arxiv.org/abs/1611.07478

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 4765–4774). Curran Associates, Inc http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

Ma, J., Yu, M. K., Fong, S., Ono, K., Sage, E., Demchak, B., Sharan, R., & Ideker, T. (2018). Using deep learning to model the hierarchical structure and function of a cell. *Nature Methods*, *15*(4), 290–298. https://doi.org/10.1038/nmeth.4627

Ma, Q., & Wang, J. T. (1999). Biological data mining using Bayesian neural networks: A case study. *International Journal on Artificial Intelligence Tools*, *8*(4), 433–451.

Manno, G. L., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., ... Kharchenko, P. V. (2018). RNA velocity of single cells. *Nature*, *56*(7719), 494–498. https://doi.org/10.1038/s41586-018-0414-6

Marčenko, V. A., & Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, *1*(4), 457–483. https://doi.org/10.1070/sm1967v001n04abeh001994

Marcinkevics, R., & Vogt, J. E. (2020). Interpretability and explainability: A machine learning zoo mini-tour. *CoRR*, 1805 https://arxiv.org/abs/2012.01805

Maslova, A., Ramirez, R. N., Ma, K., Schmutz, H., Wang, C., Fox, C., Ng, B., Benoist, C., Mostafavi, S., & Immunological Genome Project. (2020). Deep learning of immune cell differentiation. *Proceedings of the National Academy of Sciences*, *117*(41), 25655–25666. https://doi.org/10.1073/pnas.2011795117

Mathieu, E., Rainforth, T., Siddharth, N., & Teh, Y. W. (2019). Disentangling disentanglement in variational autoencoders. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on machine learning, volume 97 of proceedings of machine learning research* (pp. 4402–4412. PMLR, 09–15) https://proceedings.mlr.press/v97/mathieu19a.html

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Molnar, C., Casalicchio, G., & Bischl, B. (2021). Interpretable machine learning–a brief history, state-of-the-art and challenges. In *ECML PKDD 2020 workshops: Workshops of the European Conference on machine learning and knowledge discovery in databases (ECML PKDD 2020): SoGood 2020, PDFL 2020, MLCS 2020, NFMCP 2020, DINA 2020, EDML 2020, XKDD 2020 and INRA 2020, Ghent, Belgium, September 14–18, 2020, proceedings* (pp. 417–431). Springer.

Morota, G., & Gianola, D. (2014). Kernel-based whole-genome prediction of complex traits: A review. *Frontiers in Genetics*, *5*, 363. https://doi.org/10.3389/fgene.2014.00363

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ros, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2022). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, *56*, 3005–3054.

Muralidharan, O. (2010). An empirical Bayes mixture method for effect size and false discovery rate estimation. *The Annals of Applied Statistics*, *4*(1), 422–438. https://doi.org/10.1214/09-AOAS276

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116

Mwangi, B., Tian, T. S., & Soares, J. C. (2013). A review of feature reduction techniques in neuroimaging. *Neuroinformatics*, *12*(2), 229–244. https://doi.org/10.1007/s12021-013-9204-3

Neufeld, A. C., Gao, L. L., & Witten, D. M. (2022). Tree-values: Selective inference for regression trees. *Journal of Machine Learning Research*, *23*(305), 1–43 http://jmlr.org/papers/v23/21-0722.html

Nie, W., Zhang, Y., & Patel, A. (2018). A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In *International conference on machine learning* (pp. 3809–3818). PMLR.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., & Bustamante, C. D. (2008. ISSN 1476-4687). Genes mirror geography within Europe. *Nature*, *456*(7218), 98–101. https://doi.org/10.1038/nature07331

Nueda, M. J., Tarazona, S., & Conesa, A. (2014). Next masigpro: Updating masigpro bioconductor package for rna-seq time series. *Bioinformatics*, *3*(18), 2598–2602.

Orliac, E. J., Banos, D. T., Ojavee, S. E., Läll, K., Mägi, R., Visscher, P. M., & Robinson, M. R. (2022). Improving gwas discovery and genomic prediction accuracy in biobank data. *Proceedings of the National Academy of Sciences*, *119*(31), e2121279119. https://doi.org/10.1073/pnas.2121279119

Paananen, T., Andersen, M. R., & Vehtari, A. (2021). Uncertainty-aware sensitivity analysis using rényi divergences. In *Uncertainty in artificial intelligence* (pp. 1185–1194). PMLR.

Paananen, T., Piironen, J., Andersen, M. R., & Vehtari, A. (2019). Variable selection for gaussian processes via sensitivity analysis of the posterior predictive distribution. In *The 22nd international conference on artificial intelligence and statistics* (pp. 1743–1752). PMLR.

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), 1–20. https://doi.org/10.1371/journal.pgen.0020190

Price, A. L., Zaitlen, N. A., Reich, D., & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 459–463. https://doi.org/10.1038/nrg2813

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In Y. Bengio & Y. LeCun (Eds.), *4th international conference on learning representations, ICLR 2016, San Juan, Puerto Rico, may 2–4, 2016, conference track proceedings* http://arxiv.org/abs/1511.06434

Rahman, R., Dhruba, S. R., Ghosh, S., & Pal, R. (2019). Functional random forest with applications in dose–response predictions. *Scientific Reports*, *9*(1), 1–14.

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, *378*, 686–707.

Rakitsch, B., & Stegle, O. (2016). Modelling local gene networks increases power to detect trans-acting genetic effects on gene expression. *Genome Biology*, *17*(1), 1–13.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you? In *Proceedings of the 22nd ACM SIGKDD International Conference on knowledge discovery and data mining*. ACM. https://doi.org/10.1145/2939672.2939778

Ringnér, M. (2008). What is principal component analysis? *Nature Biotechnology*, *26*, 303–304.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015. ISSN 0305-1048). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009. ISSN 1367–4803). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Roth, A. E. (1988). *The Shapley value: Essays in honor of Lloyd S. Shapley*. Cambridge University Press.

Runcie, D. E., & Crawford, L. (2019). Fast and flexible linear mixed models for genome-wide genetics. *PLoS Genetics*, *15*(2), e1007978. https://doi.org/10.1371/journal.pgen.1007978

Runcie, D. E., Qu, J., Cheng, H., & Crawford, L. (2021). MegaLMM: Mega-scale linear mixed models for genomic predictions with thousands of traits. *Genome Biology*, *22*(1), 213. https://doi.org/10.1186/s13059-021-02416-w

Rybakov, S., Lotfollahi, M., Theis, F. J., & Wolf, F. A. (2020). Learning interpretable latent autoencoder representations with annotations of feature sets. In *Machine learning in computational biology*. Cold Spring Harbor Laboratory. https://doi.org/10.1101/2020.12.02.401182

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning* (Vol. 11, p. 700). Springer Nature.

Seninge, L., Anastopoulos, I., Ding, H., & Stuart, J. (2021. ISSN 2041-1723). Vega is an interpretable generative model for inferring biological network activity in single-cell transcriptomics. *Nature Communications*, *12*(1), 5684. https://doi.org/10.1038/s41467-021-26017-0

Sesia, M., Katsevich, E., Bates, S., Candès, E., & Sabatti, C. (2020). Multi-resolution localization of causal variants across the genome. *Nature Communications*, *11*(1), 1093. https://doi.org/10.1038/s41467-020-14791-2

Settles, B. (2009). *Active learning literature survey*. Computer Sciences Technical Report 1648. University of Wisconsin–Madison.

Shapley, L. S. (1951). *Notes on the N-person game–I: Characteristic-point solutions of the four-person game*. Rand Corporation.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on machine learning- Volume 70*, ICML'17 (pp. 3145–3153). JMLR.org.

K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In Y. Bengio & Y. LeCun (Eds.), *2nd international conference on learning representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, workshop track proceedings* http://arxiv.org/abs/1312.6034

Sixt, L., Granz, M., & Landgraf, T. (2020). When explanations lie: Why many modified bp attributions fail. In *International conference on machine learning* (pp. 9046–9057). PMLR.

Song, Z., & Li, J. (2021). Variable selection with false discovery rate control in deep neural networks. *Nature Machine Intelligence*, *3*(5), 426–433. https://doi.org/10.1038/s42256-021-00308-z

Stephens, M. (2016). False discovery rates: A new deal. *Biostatistics*, *18*(2), 275–294. https://doi.org/10.1093/biostatistics/kxw041

Stephens, M., & Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, *1*(10), 681–690. https://doi.org/10.1038/nrg2615

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., & Robinson, G. E. (2015). Big data: Astronomical or genomical? *PLoS Biology*, *13*(7), 1–11. https://doi.org/10.1371/journal.pbio.1002195

Stites, M. C., Nyre-Yu, M., Moss, B., Smutz, C., & Smith, M. R. (2021). Sage advice? The impacts of explanations for machine learning models on human decision-making in spam detection. In H. Degen & S. Ntoa (Eds.), *Artificial Intelligence in HCI* (pp. 269–284). Springer International Publishing ISBN 978-3-030-77772-2.

Svensson, V., Gayoso, A., Yosef, N., & Pachter, L. (2020. ISSN 1367-4803). Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics*, *36*(11), 3418–3421. https://doi.org/10.1093/bioinformatics/btaa169

Sverchkov, Y., & Craven, M. (2017). A review of active learning approaches to experimental design for uncovering biological networks. *PLoS Computational Biology*, *13*(6), e1005466.

Swain, P. S., Stevenson, K., Leary, A., Montano-Gutierrez, L. F., Clark, I. B., Vogel, J., & Pilizota, T. (2016). Inferring time derivatives including cell growth rates using gaussian processes. *Nature Communications*, *7*(1), 13766. https://doi.org/10.1038/ncomms13766

Tasaki, S., Gaiteri, C., Mostafavi, S., & Wang, Y. (2020). Deep learning decodes the principles of differential gene expression. *Nature Machine Intelligence*, *2*(7), 376–386.

Taylor, J., & Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, *112*(25), 7629–7634. https://doi.org/10.1073/pnas.1507583112

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288 ISSN 00359246. http://www.jstor.org/stable/2346178

Tibshirani, R. J., Taylor, J., Lockhart, R., & Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, *111*(514), 600–620. https://doi.org/10.1080/01621459.2015.1108848

Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *61*(3), 611–622 ISSN 13697412, 14679868. http://www.jstor.org/stable/2680726

Tolosi, L., & Lengauer, T. (2011). Classification with correlated features: Unreliability of feature ranking and solutions. *Bioinformatics*, *27*(14), 1986–1994.

Tolstikhin, I., Bousquet, O., Gelly, S., & Schoelkopf, B. (2018). Wasserstein auto-encoders. In *International Conference on Learning Representations*.

Townes, F. W., Hicks, S. C., Aryee, M. J., & Irizarry, R. A. (2019. ISSN 1474-760X). Feature selection and dimension reduction for single-cell RNA-seq based on a multinomial model. *Genome Biology*, *2*(1), 295. https://doi.org/10.1186/s13059-019-1861-6

Tsang, M., Cheng, D., & Liu, Y. (2018). Detecting statistical interactions from neural network weights. In *International conference on learning representations* https://openreview.net/forum?id=ByOfBggRZ

Tsang, M., Liu, H., Purushotham, S., Murali, P., & Liu, Y. (2018). Neural interaction transparency (nit): Disentangling learned interactions for improved interpretability. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 31). Curran Associates, Inc https://proceedings.neurips.cc/paper/2018/file/74378afe5e8b20910cf1f939e57f0480-Paper.pdf

van Bergen, G. H., Duenk, P., Albers, C. A., Bijma, P., Calus, M. P., Wientjes, Y. C., & Kappen, H. J. (2020). Bayesian neural networks with variable selection for prediction of genotypic values. *Genetics Selection Evolution*, *52*(1), 1–14.

van der Wijst, M. G. P., de Vries, D. H., Brugge, H., Westra, H.-J., & Franke, L. (2018. ISSN 1756-994X). An integrative approach for building personalized gene regulatory networks for precision medicine. *Genome Medicine*, *1*(1), 96. https://doi.org/10.1186/s13073-018-0608-4

Videla Rodriguez, E. A., Pértille, F., Guerrero-Bosagna, C., Mitchell, J. B., Jensen, P., & Smith, V. A. (2022). Practical application of a Bayesian network approach to poultry epigenetics and stress. *BMC Bioinformatics*, *23*(1), 1–16.

Wahba, G. (1990). *Spline models for observational data*. United States: Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611970128

Wang, G., Sarkar, A., Carbonetto, P., & Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(5), 1273–1300. https://doi.org/10.1111/rssb.12388

Wang, L., Zhang, B., Wolfinger, R. D., & Chen, X. (2008). An integrated approach for the analysis of biological pathways using mixed models. *PLoS Genetics*, *4*(7), 1–9. https://doi.org/10.1371/journal.pgen.1000115

Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., & MacNeille, P. (2017). A Bayesian framework for learning rule sets for interpretable classification. *Journal of Machine Learning Research*, *18*(70), 1–37 http://jmlr.org/papers/v18/16-003.html

Weissbrod, O., Geiger, D., & Rosset, S. (2016). Multikernel linear mixed models for complex phenotype prediction. *Genome Research*, *26*(7), 969–979. https://doi.org/10.1101/gr.201996.115

Weissbrod, O., Kaufman, S., Golan, D., & Rosset, S. (2019). Maximum likelihood for Gaussian process classification and generalized linear mixed models under case–control sampling. *Journal of Machine Learning Research*, *2*(108), 1–30 http://jmlr.org/papers/v20/18-298.html

Woo, J. H., Shimoni, Y., Yang, W. S., Subramaniam, P., Iyer, A., Nicoletti, P., Marítnez, M. R., López, G., Mattioli, M., Realubit, R., Karan, C., Stockwell, B. R., Bansal, M., & Califano, A. (2015). Elucidating compound mechanism of action by network perturbation analysis. *Cell*, *162*(2), 441–451.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., & Lin, X. (2010). Powerful SNP-set analysis for case–control genome-wide association studies. *The American Journal of Human Genetics*, *86*(6), 929–942. https://doi.org/10.1016/j.ajhg.2010.05.002

Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K., Hua, Y., Gueroussov, S., Najafabadi, H. S., Hughes, T. R., Morris, Q., Barash, Y., Krainer, A. R., Jojic, N., Scherer, S. W., Blencowe, B. J., & Frey, B. J. (2015). The human splicing code reveals new insights into the genetic determinants of disease. *Science*, *347*(6218), 1254806.

Xu, B., Wang, N., Chen, T., & Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. In *ICML deep learning workshop*.

Yang, J., Fritsche, L. G., Zhou, X., & Abecasis, G. (2017). A scalable Bayesian method for integrating functional information in genome-wide association studies. *The American Journal of Human Genetics*, *101*(3), 404–416. https://doi.org/10.1016/j.ajhg.2017.08.002

Zeng, P., & Zhou, X. (2017). Non-parametric genetic prediction of complex traits with latent dirichlet process regression models. *Nature Communications*, *8*(1), 456. https://doi.org/10.1038/s41467-017-00470-2

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, *44*(7), 821–824. https://doi.org/10.1038/ng.2310

Zhu, X., & Stephens, M. (2018). Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nature Communications*, *9*(1), 4361. https://doi.org/10.1038/s41467-018-06805-x

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.

Zou, Y., Carbonetto, P., Wang, G., & Stephens, M. (2022). Fine-mapping from summary data with the "sum of single effects" model. *PLoS Genetics*, *18*(7), e1010299. https://doi.org/10.1371/journal.pgen.1010299

How to cite this article: Conard, A. M., DenAdel, A., & Crawford, L. (2023). A spectrum of explainable and interpretable machine learning approaches for genomic studies. *WIREs Computational Statistics*, *15*(5), e1617. https://doi.org/10.1002/wics.1617