Human Genetics
https://doi.org/10.1007/s00439-021-02387-9

ORIGINAL INVESTIGATION

Interpretable machine learning for genomics

David S. Watson[1]

Received: 20 April 2021 / Accepted: 8 October 2021

Abstract

High-throughput technologies such as next-generation sequencing allow biologists to observe cell function with unprecedented resolution, but the resulting datasets are too large and complicated for humans to understand without the aid of advanced statistical methods. Machine learning (ML) algorithms, which are designed to automatically find patterns in data, are well suited to this task. Yet these models are often so complex as to be opaque, leaving researchers with few clues about underlying mechanisms. Interpretable machine learning (iML) is a burgeoning subdiscipline of computational statistics devoted to making the predictions of ML models more intelligible to end users. This article is a gentle and critical introduction to iML, with an emphasis on genomic applications. I define relevant concepts, motivate leading methodologies, and provide a simple typology of existing approaches. I survey recent examples of iML in genomics, demonstrating how such techniques are increasingly integrated into research workflows. I argue that iML solutions are required to realize the promise of precision medicine. However, several open challenges remain. I examine the limitations of current state-of-the-art tools and propose a number of directions for future research. While the horizon for iML in genomics is wide and bright, continued progress requires close collaboration across disciplines.

Introduction

Technological innovations have made it relatively cheap and easy to observe biological organisms at molecular resolutions. High-throughput methods, such as next-generation sequencing and the full suite of "omic" platforms—e.g., genomic, proteomic, metabolomic, and related technologies—have inaugurated a new era of systems biology, providing data so abundant and detailed that researchers are not always sure just what to do with the newfound embarrassment of riches. One of the most salient traits of these datasets is their sheer size. Sequencing technologies can record anywhere from a few thousand to a few billion features per sample. Another important factor, related but distinct, is that genomic data are not immediately intelligible to humans. Whereas a small child can accurately classify pictures of animals, experts cannot generally survey a genetic sequence and predict health outcomes—at least not without the aid of advanced statistical models.

Machine learning (ML) algorithms are designed to automatically mine data for insights using few assumptions and lots of computational power. With their ability to detect and exploit complex relationships in massive datasets, ML techniques are uniquely suited to the challenges of modern genomics. In this article, I will focus specifically on supervised learning methods, which attempt to estimate a function from inputs (e.g., gene expression) to outputs (e.g., disease diagnosis).[1] ML algorithms have become enormously popular in medical research (Topol 2019), especially in imaging tasks, such as radiological screening (Mazurowski et al. 2019) and tumor identification (McKinney et al. 2020). They have also been successfully applied to complex molecular problems such as antibiotic discovery (Stokes et al. 2020) and predicting regulatory behavior from genetic variation (Eraslan et al. 2019). ML promises to advance our understanding of fundamental biology and revolutionize the practice of medicine, enabling personalized treatment regimens tailored to a patient's unique biomolecular profile.

✉ David S. Watson
david.watson@ucl.ac.uk

[1] Department of Statistical Science, University College London, London, UK

[1] This is distinct from *unsupervised* learning, which searches for structure without predicting outcomes, and *reinforcement* learning, where agents select a policy to maximize rewards in a particular environment. Though both methods have been applied in genomics, supervised learning is more prevalent and remains the focus of almost all contemporary research in interpretability.

Published online: 20 October 2021

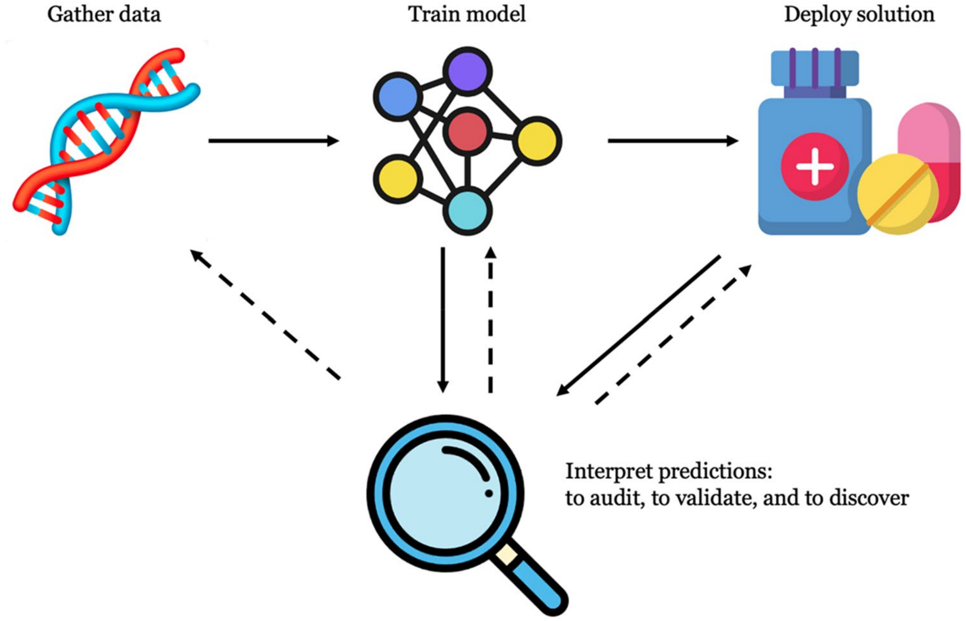


Fig. 1 The classic bioinformatics workflow spans data collection, model training, and deployment. iML augments this pipeline with an extra interpretation step, which can be used during training and throughout deployment (incoming solid edges). Algorithmic explanations (outgoing dashed edges) can be used to guide new data collection, refine training, and monitor models during deployment

Despite all their strengths and achievements, supervised learning techniques pose a number of challenges, some of which are especially troubling in biomedical contexts. Foremost among these is the issue of interpretability. Successful ML models are often so complex that no human could possibly follow the reasoning that leads to individual predictions. Inputs may pass through a long sequence of recursive nonlinearities, spanning thousands or millions of parameters, before a prediction emerges out the other side. How can such a black box, no matter how accurate, advance our knowledge of biological mechanisms? How can we trust what we do not understand?

Interpretable machine learning (iML)—also known as explainable artificial intelligence (xAI) or, more simply, explainability—is a fast-growing subfield of computational statistics devoted to helping users make sense of the predictions of ML models. Cataloging the state-of-the-art in iML has become a whole meta-literature unto itself. Recent examples include (but are almost certainly not limited to): Adadi and Berrada (2018), Gilpin et al. (2018), Guidotti et al. (2018b), Mueller et al. (2019), Murdoch et al. (2019), Barredo Arrieta et al. (2020), Das and Rad (2020), Mohseni et al. (2020), Vilone and Longo (2020), Xu et al. (2020), Marcinkevičs and Vogt (2020), Linardatos et al. (2021) and Rudin et al. (2021). Holzinger et al. (2019) frame their survey with a focus on medical applications, while Azodi et al. (2020) specialize more narrowly on iML for genetics. For reviews of interpretable deep learning in genomics, see Talukder et al. (2021) and Treppner et al. (2021).

My goal in this article is not to add yet another survey to this overpopulated field. Instead, I aim to provide a gentle and critical introduction to iML for genomic researchers. I define relevant concepts, motivate prominent approaches, and taxonomize popular methodologies. I examine important opportunities and challenges for iML in genomics, arguing that more intelligible algorithms will ultimately advance our understanding of systems biology and play a crucial role in realizing the promise of precision medicine. I show how iML algorithms augment the traditional bioinformatics workflow with explanations that can be used to guide data collection, refine training procedures, and monitor models throughout deployment (see Fig. 1). To do so, however, the field must overcome several conceptual and technical obstacles. I outline these issues and suggest a number of future research directions, all of which require close interdisciplinary collaboration.

The remainder of this article is structured as follows. First, I review relevant background material and provide a typology of iML. Next, I examine three motivations for iML, each of which has a role to play in computational biology. I proceed to introduce a number of popular iML methodologies that have been used in recent genomic research. Following a discussion, in which I consider three open challenges for iML that are especially urgent in bioinformatics, I conclude with some reflections on the future of the field.

Background

In supervised learning, we assume access to a finite training dataset of n input–output pairs, $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$. The vector $\boldsymbol{x}_i$ denotes a point in p-dimensional space, $\boldsymbol{x}_i = \left(x_{i1}, \dots, x_{ip}\right)$, with coordinate x_{ij} corresponding to the ith value of feature X_j. For instance, $\boldsymbol{x}_i$ may represent an individual's transcriptomic profile and x_{ij} their read count for gene X_j. Samples are presumed to be independently and identically distributed instances of some fixed but unknown joint distribution $P(\boldsymbol{X}, Y)$. The goal is to infer a function $f : \mathcal{X} \to \mathcal{Y}$ that maps $\boldsymbol{x}$-vectors to y-outcomes. For example, we may want to group cancer patients into subtypes with different prognostic trajectories based on gene expression. If outcomes are continuous, we say that f is a regressor; if they are categorical, we say that f is a classifier. In either case, performance is evaluated via some loss function that quantifies model error. While the expected loss—also known as the risk—cannot be directly calculated without knowledge of the underlying distribution, it can be estimated either on the training data or, preferably, on an independent test set sampled from the same distribution. Empirical risk minimization is the learning strategy whereby we select the top performing model from some predefined function class (Vapnik 1995). Popular algorithms of this type include (potentially regularized) linear models, neural networks, and tree-based ensembles, such as random forests and gradient boosting machines. See (Hastie et al. 2009) for a good introduction.

It is not always obvious what would constitute a successful explanation of a given model prediction. Indeed, explanation itself is an epistemologically contested concept, the subject of ancient and modern philosophical debates (Woodward 2019). It should perhaps come as no surprise then to learn that algorithmic explanations come in many flavors. The first point to acknowledge is that iML tools are used for different analytic purposes. For instance, they may help to estimate or understand a true functional relationship presumed to hold in nature. Alternatively, they may be used to analyze the behavior of a fitted model—to illuminate the black box, as it were. Finally, they may be involved in the design of so-called "glass box" algorithms, i.e., some novel function class specifically built for transparency. These goals may overlap at the edges, and methods originally intended for one may be repurposed for another. However, each represents a distinct task with its own challenges. Not all are equally prevalent in genomics, though this review will discuss examples of each. Keeping these aims separate is crucial to avoid the conceptual pitfalls addressed in "Open challenges".

The following typology is adapted from Molnar's (2019) textbook guide to iML, which provides a helpful overview of technical approaches and the current state-of-the-art. Roughly put, there are three key dichotomies that orient iML research: intrinsic vs. post-hoc, model-specific vs. model-agnostic, and global vs. local. A final consideration is what type of output the method generates. Each point is considered in turn.

Intrinsic vs. post-hoc

An intrinsically explainable algorithm is one that raises no intelligibility issues in the first place—i.e., a glass box algorithm. Canonical examples include (sparse) linear regression and (short) rule lists. Parametric models typically include interpretable parameters that correspond to meaningful quantities, e.g., a linear coefficient denoting a gene's log fold change in a microarray experiment. Some have argued that such models are the only ones that should be allowed in high-risk settings, such as those arising in healthcare (Rudin 2019). Unfortunately, many interesting real-world problems cannot be adequately modeled with intrinsically explainable algorithms. Watson et al. (2019) argue that in clinical medicine, doctors are obligated to use whatever available technology leads to the best health outcomes on average, even if that involves opaque ML algorithms. Of course, flexible ML models are also prone to overfit the training data, especially in high-dimensional settings. The choice of which approach to use invariably depends on contextual factors—the task at hand, the prior knowledge available, and what assumptions the analyst deems reasonable. Should researchers choose to use some black box method, interpreting predictions will require post-hoc tools, which take a target model f as input and attempt to explain its predictions, at least near some region of interest.

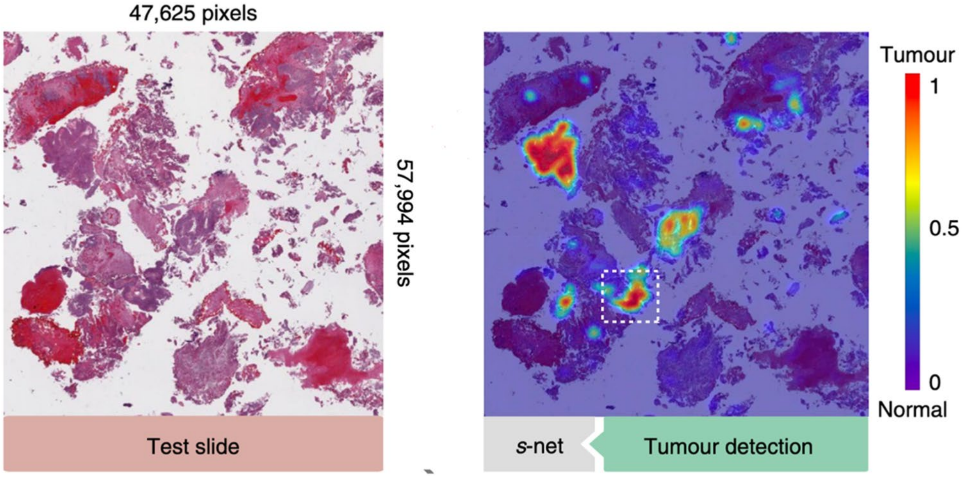
Model-specific vs. model-agnostic

Model-specific iML solutions take advantage of the assumptions and architectures upon which particular algorithms are built to generate fast and accurate explanations. For example, much work in iML has been specifically devoted to deep neural networks (Bach et al. 2015; Montavon et al. 2017; Shrikumar et al. 2017; Sundararajan et al. 2017), an especially rich class of functions with unique explanatory affordances and constraints. Model-agnostic tools, on the other hand, strive for more general applicability. Treating the fitted function f as a black box, they attempt to explain its predictions with few or no assumptions about the data generating process. Model-agnostic approaches are especially useful in cases where f is inaccessible (for example if an algorithm is protected by intellectual property laws), while model-specific methods are generally more efficient and reliable when f's structure is known.

Global vs. local

A global explanation helps the user understand the behavior of the target model f across all regions of the feature space. This is difficult to achieve when f is complex and/or high-dimensional. A local explanation, by contrast, is only meant to apply to the area near some particular point of interest.

Fig. 2 A saliency map visually explains a cancer diagnosis based on whole-slide pathology data. The highlighted regions on the right pick out the elements of the image that the algorithm deemed most strongly associated with malignancy. From (Zhang et al. 2019, p. 237)



For instance, a properly specified linear regression is globally explainable in the sense that the model formula holds with equal probability for any randomly selected data point. However, a local linear approximation to some nonlinear f will fit best near the target point, and does not in general tell us anything about how the model behaves in remote regions of the feature space. In biological contexts, we can think of global and local explanations applying at population and individual levels, respectively. These are poles of a spectrum that also admits of intermediate alternatives, e.g., subpopulation- or group-level explanations, which are possible as well.

A final axis of variation for iML tools is their output class. Typically, these methods explain predictions through some combination of images, statistics, and/or examples. Visual explanations are especially well suited to image classifiers (see Fig. 2). Other common visual approaches include plots that illustrate the partial dependence (Friedman 2001) or individual conditional expectation (Casalicchio et al. 2019) of variables, which can inform users about feature interactions and/or causal effects (Zhao and Hastie 2021). Statistical outputs, by contrast, may include rule lists, tables, or numbers quantifying explanatory value in some predefined way. Finally, exemplary methods report informative data points, either in the form of prototypes, which typify a given class (Chen et al. 2019), or counterfactuals, which represent the most similar sample on the opposite side of a decision boundary (Wachter et al. 2018). These latter methods are less common in genomics, although conceptually similar matching algorithms are used in clinical medicine (Bica et al. 2021).

Motivations

Why do we seek algorithmic explanations in the first place? Watson and Floridi (2021) offer three reasons: (1) to audit for potential bias; (2) to validate performance, guarding against unexpected errors; and (3) to discover underlying mechanisms of the data generating process. All three are relevant for genomics.

To audit

Healthcare often magnifies social inequalities. For recent evidence, look no further than the COVID-19 pandemic, which disproportionately affects minority populations in the US and the UK (Egede and Walker 2020). ML threatens to automate these injustices. Obermeyer et al. (2019) found evidence of significant racial bias in a healthcare screening algorithm used by millions of Americans. Simulations suggest that rectifying the disparity would nearly triple the number of Black patients receiving medical care. Similar problems are evident in genomic research. Individuals of white European ancestry make up some 16% of the global population but constitute nearly 80% of all genome-wide association study (GWAS) subjects, raising legitimate concerns that polygenic risk scores and other tools of precision medicine may increase health disparities (Martin et al. 2019). As genomic screening becomes more prevalent, there will be substantial and justified pressure to ensure that new technologies do not reinforce existing inequalities. Algorithmic fairness and explainability may even be legally required under the European Union's 2018 General Data Protection Regulation (GDPR), depending on one's interpretation of the relevant articles (Selbst and Powles 2017; Wachter et al. 2017). By making a model's reliance on potentially sensitive attributes more transparent, iML methods can help quantify and mitigate potential biases.

To validate

The second motivation concerns the generalizability of ML models. Supervised learning algorithms are prone to overfitting, which occurs when associations in the training data do not generalize to test environments. In a famous example, a neural network trained on data from a large New York hospital classified asthmatics with pneumonia as low risk, a result that came as a surprise to the doctors and data scientists working on the project (Caruana et al. 2015). The algorithm had not uncovered some subtle pulmonological secret. On the contrary, the apparent association was spurious. Asthmatics with pneumonia are at such great risk that emergency room doctors immediately send them to the intensive care unit, where their chances of survival are relatively high. It was the extra medical attention, not the underlying condition, which improved outcomes for these patients. Naively applying this algorithm in a new environment—e.g., some hospital where patient triage is performed by a neural network with little or no input from doctors—could have grave consequences.

Overfitting has been observed in GWAS models (Nicholls et al. 2020), where associations that appear informative in one population do not transfer to another. Such failures can be difficult to detect given the complexity of the underlying signals, which may depend on environmental factors or subtle multi-omic interactions. The problem of external validity or transportability is well known in the natural and social sciences, if not always well understood (Bareinboim and Pearl 2016; Pearl and Bareinboim 2014). Because causal dependencies are robust to perturbations of upstream variables, environmental heterogeneity—e.g., different patient subpopulations or data collection protocols—can help isolate and quantify causal effects (Heinze-Deml et al. 2018; Peters et al. 2016; Meinshausen et al. 2016). These methods have motivated novel GWAS methodologies that search for persistent associations across varying patterns of linkage disequilibrium (Li et al. 2021). iML algorithms, together with causal inference tools (Imbens and Rubin 2015; Pearl 2000; Peters et al. 2017), can help researchers identify and remove spurious signals, ensuring better generalizability to new environments.

To discover

A final goal of iML, less widely discussed than the previous two but arguably of greater interest in genomics, is to reveal unknown properties and mechanisms of the data generating process. In this case, the guiding assumption is not that the target model f is biased or overfit; on the contrary, we assume that it has found some true signal and investigate its internal logic to learn more. This may mean examining weights from a support vector machine, approximating a decision boundary with some local linear model, or extracting Boolean rules to describe the geometry of some complex regression surface. Examples of all these approaches and more will be examined below, demonstrating how iML can be—and to some extent already has been—integrated into genomic research workflows. By unpacking the reasoning that underlies high-performance statistical models, iML algorithms can mine for insights and suggest novel hypotheses in a flexible, data-driven manner. Rightly or wrongly, it is this capacity—not its potential utility for auditing or validation—that is likely to inspire more widespread adoption in bioinformatics.

Methodologies and applications

In this section, I introduce a number of prominent approaches to iML. As a running example, I will consider a hypothetical algorithm that classifies breast cancer patients into different subtypes on the basis of a diagnostic biomarker panel. Such tools have been in use for decades, although recent advances in ML have vastly increased their accuracy and sophistication (Cascianelli et al. 2020; Sarkar et al. 2021). I examine three particular iML methods at length—variable importance, local linear approximators, and rule lists—describing their basic forms and reviewing some recent applications. There is considerable variety within each subclass, and the choice of which to use for a given task is inevitably context dependent. While all three could be fruitfully applied for auditing, validation, or discovery, the latter has tended to dominate in genomic research to date.

Variable importance

Variable importance (VI) measures are hardly new. Laplace and Gauss, writing independently in the early nineteenth century, each described how standardized linear coefficients can be interpreted as the average change in response per unit increase of a feature, with all remaining covariates held fixed. Coefficients with larger absolute values therefore suggest greater VI. If our cancer subtyping algorithm were linear, then we might expect large weights on genes such as BRCA1, which is strongly associated with basal-like breast cancer (BLBC) (Turner and Reis-Filho 2006), and ESR1, a known marker of the luminal A subtype (Sørlie et al. 2003).

The ease of computing VI for linear models has been exploited to search for causal variants in single nucleotide polymorphism (SNP) arrays via penalized regression techniques like the lasso (Tibshirani 1996) and elastic net (Zou and Hastie 2005), both popular in GWAS (Waldmann et al. 2013). Random forests (Breiman 2001a), one of the most common supervised learning methods in genomics (Chen and Ishwaran 2012), can provide a range of marginal or conditional VI scores, typically based either on permutations or impurity metrics (Altmann et al. 2010; Nembrini et al. 2018; Strobl et al. 2008). Support vector machines (SVMs) may give intelligible feature weights depending on the underlying kernel (Schölkopf et al. 2004). For example, Sonnenburg et al. (2008) use a string kernel to predict splice sites in *C. elegans* and extract relevant biological motifs from the resulting positional oligomer importance matrices. The method has since been extended to longer sequence motifs and more general learning procedures (Vidovic et al. 2015, 2017). More recently, Kavvas et al. (2020) used an intrinsically interpretable SVM to identify genetic determinants of antimicrobial resistance (AMR) from whole-genome sequencing data. Whether these methods tell us about the importance of features in nature or just in some fitted model f depends on whether we assume that f accurately captures the functional form of the relationship between predictors and outcomes.

To go back to the typology above, VI measures are global parameters that may be intrinsic (as in linear models) or post-hoc (as in random forest permutation importance). Model-specific versions are popular, although several model-agnostic variants have emerged in recent years. These include targeted maximum likelihood measures (Hubbard et al. 2018; Williamson et al. 2021), nested model tests using conformal inference (Lei et al. 2018; Rinaldo et al. 2019), and permutation-based reliance statistics (Fisher et al. 2019). Such methods typically involve computationally intensive procedures, such as bootstrapping, permutations, and/or model refitting, which pose both computational and statistical challenges when applied in settings with large sample sizes and/or high-dimensional feature spaces, such as those commonly found in genomics.

A notable exception specifically designed for high-dimensional problems is the knockoff test for variable selection (Barber and Candès 2015; Candès et al. 2018). The basic idea behind this approach is to generate a set of "control" variables—the eponymous knockoffs—against which to test the importance of the original input features. For a given $n \times p$ design matrix $\boldsymbol{X}$, we say that $\tilde{\boldsymbol{X}}$ is the corresponding knockoff matrix if it meets the following two criteria:

(a) *Pairwise exchangeability*. For any proper subset $S \subset [p] = (1, \dots, p)$:

$$(\boldsymbol{X}, \tilde{\boldsymbol{X}})_{swap(S)} =^d (\boldsymbol{X}, \tilde{\boldsymbol{X}}),$$

where $=^d$ represents equality in distribution and the swapping operation is defined below.
(b) *Conditional independence*. $\tilde{\boldsymbol{X}} \perp Y \mid \boldsymbol{X}$.

A swap is obtained by switching the entries X_j and $\tilde{X}_j$ for each $j \in S$. For example, with $p = 3$ and $S = \{1, 3\}$:

$$(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)_{\mathrm{swap}(S)} =^d (\tilde{X}_1, X_2, \tilde{X}_3, X_1, \tilde{X}_2, X_3).$$

A supervised learning algorithm is then trained on the expanded $n \times 2p$ feature matrix, including original and knockoff variables. If X_j does not significantly outperform its knockoff $\tilde{X}_j$ by some model-specific importance measure—Candès et al. (2018) describe methods for lasso and random forests—then the original feature may be safely removed from the final model. The authors outline an adaptive thresholding procedure that provably provides finite sample false discovery rate (FDR) control for variable selection under minimal assumptions. They also describe a conditional randomization test (CRT) for asymptotic type I error rate control, in which observed values are compared to null statistics repeatedly sampled from the knockoff distribution. Experiments suggest the CRT is more powerful than the adaptive procedure, and therefore could be preferable when signals are sparse. However, Candès et al. caution that the CRT may be infeasible for large datasets.

The most challenging aspect of this pipeline is computing the knockoff variables themselves. However, when good generative models are available, the technique can be quite powerful. Watson and Wright (2021) use Candès et al.'s original semidefinite programming formulation to approximate knockoffs for a DNA microarray experiment. Sesia et al. (2019) extend the method to genotype data using a hidden Markov model to sample knockoffs. In a recent study, Bates et al. (2020) used knockoffs with GWAS data from parents and offspring to create a "digital twin test" based on the CRT, in which causal variants are identified by exploiting the natural randomness induced by meiosis. The method has greater power and localization than the popular transmission disequilibrium test.

It should be noted that VI measures can be applied either to raw or processed features. The latter can be especially valuable when the original data is difficult to interpret. In a recent paper, Chia et al. (2020) use direct wavelet transform to capture the location and frequency of bacterial Raman spectra as a preprocessing step toward building a more interpretable classifier. They use knockoffs for (processed) feature selection and train a multinomial logistic regression, reporting performance on par with the best black box models. Of course, not all genomic settings have comparably interpretable low-dimensional representations. I shall revisit the issue of variable abstraction and granularity as an open challenge below. .

Local linear approximators

All methods described above are global in scope. The goal in many iML settings, however, is to provide explanations for individual predictions. This could be useful if, for instance, a subject receives an unexpected diagnosis. Perhaps Alice shows few obvious signs of BLBC and deviates markedly from the classic patient profile, yet our algorithm assigns her to this class with high probability. Given the aggressive treatment regime likely to follow such a diagnosis,

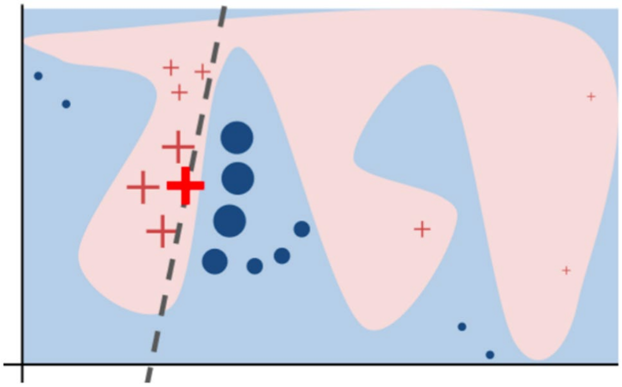


Fig. 3 A complex decision boundary (the pink blob/blue background) separates red crosses from blue circles. This function cannot be well-approximated by a linear model, but the boundary near the large red cross is roughly linear, as indicated by the dashed line. From (Ribeiro et al. 2016, p. 1138)

Alice wants to be certain that the classification is correct. A local explanation reveals that, though BRCA1 mutations account for many BLBC predictions, in her case, the feature is relatively unimportant. Instead, her local explanation turns largely on CXCR6—a gene associated with the Basal I subtype, which has a better prognosis on average than Basal II, and is therefore less likely to require high doses of chemotherapy (Milioli et al. 2017).

Local linear approximators have become the de facto standard method for computing local explanations of this sort (Bhatt et al. 2020). These algorithms assign weights to each input feature that sum to the model output. The idea derives from the insight that even though target functions may be highly complex and nonlinear, any point on a continuous curve will have a linear tangent (see Fig. 3). By estimating the formula for this line, we can approximate the regression surface or decision boundary at a particular point. Popular examples of local linear approximators include LIME (Ribeiro et al. 2016) and SHAP (Lundberg and Lee 2017), both of which are implemented in user-friendly Python libraries. The latter has additionally been incorporated into iML toolkits distributed by major tech firms, such as Microsoft,[2] Google,[3] and IBM.[4] I will briefly explicate the theory behind this method, which unifies a number of similar approaches, including LIME.

SHAP is founded on principles from cooperative game theory, where Shapley values were originally proposed as a way to fairly distribute surplus across a coalition of players (Shapley 1953). In iML settings, players are replaced by input features and Shapley values measure their contribution to a given prediction. Let $\boldsymbol{x}_i \in \mathbb{R}^p$ denote an input datapoint and $f(\boldsymbol{x}_i) \in \mathbb{R}$ the corresponding output of function f. Shapley values decompose this number into a sum of feature attributions:

$$f(\boldsymbol{x}_i) = \sum_{j=0}^{p} \phi_j,$$

where ϕ_0 denotes a baseline expectation (e.g., the mean response) and ϕ_j ($j \geq 1$) the weight assigned to feature X_j at point $\boldsymbol{x}_i$. Let $v : 2^p \to \mathbb{R}$ be a value function such that $v(S)$ is the payoff associated with feature subset $S \subseteq [p]$ and $v(\{\emptyset\}) = 0$. The Shapley value ϕ_j is given by j's average marginal contribution to all subsets that exclude it:

$$\phi_j = \frac{1}{p!} \sum_{S \subseteq [p] \backslash \{j\}} |S|!(p - |S| - 1)!\left[v(S \cup \{j\}) - v(S)\right].$$

It can be shown that this is the unique value satisfying a number of desirable properties, including efficiency, linearity, sensitivity, and symmetry.[5] Computing exact Shapley values is NP-hard, although several efficient approximations have been proposed (Sundararajan and Najmi 2019).

There is some ambiguity as to how one should calculate payoffs on a proper subset of features, since f requires p-dimensional input. Let S and R be a partition of $[p]$, such that we can rewrite any $\boldsymbol{x}_i$ as a pair of subvectors $(\boldsymbol{x}_i^S, \boldsymbol{x}_i^R)$. Then the payoff for feature subset S takes the form of an expectation, with $\boldsymbol{x}_i^S$ held fixed while $\boldsymbol{X}^R$ varies. Following Merrick and Taly (2020), we consider a general formulation of the value function, indexed by a distribution $\mathcal{D}_R$:

$$v_{\mathcal{D}_R}(S) = \mathbb{E}_{X^R \sim \mathcal{D}_R}\left[f\left(\boldsymbol{x}_i^S, \boldsymbol{X}^R\right)\right].$$

Popular options for $\mathcal{D}_R$ include the marginal distribution $P(\boldsymbol{X}^R)$, which is the default choice in SHAP; the conditional $P(\boldsymbol{X}^R | \boldsymbol{x}_i^S)$, implemented in the R package shapr (Aas et al. 2021); and the interventional $P(\boldsymbol{X}^R | do(\boldsymbol{x}_i^S))$, recently proposed by Heskes et al. (2020). Each reference distribution offers certain advantages and disadvantages, but the choice of which to use is ultimately dependent upon one's analytical goals (more on this below).

SHAP has been used to identify biomarkers in a number of genomic studies. A model-specific variant known as DeepSHAP—with close ties to related methods DeepLIFT (Shrikumar et al. 2017) and integrated gradients (Sundararajan et al. 2017), all techniques for explaining the predictions of deep neural networks—was recently used in conjunction with a model for predicting differential expression based on

[2] See https://github.com/interpretml/interpret.

[3] See https://pair-code.github.io/what-if-tool/.

[4] See http://aix360.mybluemix.net/.

[5] For formal statements of the Shapley axioms, see Lundberg & Lee (2017).

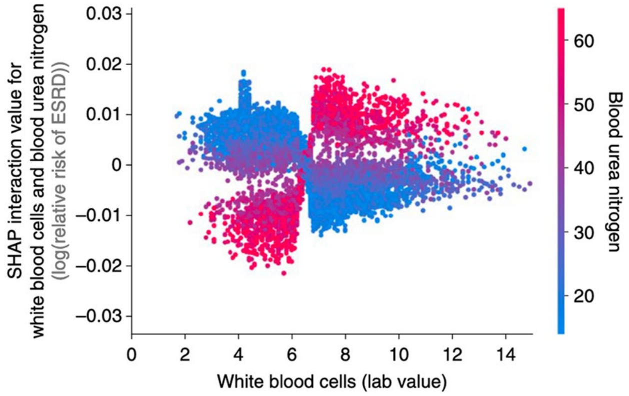


Fig. 4 Shapley values show that high white blood cell counts increase the negative risk conferred by high blood urea nitrogen for progression to end stage renal disease (ESRD). From (Lundberg et al. 2020, p. 61)

genome-wide binding sites on RNAs and promoters (Tasaki et al. 2020). The same tool was used to identify CpG loci in a DNA methylation experiment that best predicted a range of biological and clinical variables, including cell type, age, and smoking status (Levy et al. 2020). Yap et al. (2021) trained a convolutional neural network to classify tissue types using transcriptomic data, prioritizing top genes as selected by DeepSHAP. Another SHAP variant—TreeExplainer (Lundberg et al. 2020), which is optimized for tree-based ensembles such as random forests—was used to identify taxa in the skin microbiome most closely associated with various phenotypic traits (Carrieri et al. 2021). SHAP is also gaining popularity in mass spectrometry, where data heterogeneity can complicate more classical inference procedures (Tideman et al. 2021; Xie et al. 2020).

In each of these cases, further investigation was required to confirm the involvement of selected features in the target functions. The outputs of SHAP, or any other iML algorithm for that matter, are by no means decisive or infallible. However, they offer a principled and novel approach for feature ranking and selection, as well as exploring interactions that can reveal unexpected mechanisms and guide future experiments. For instance, Lundberg et al. (2020) use Shapley interaction values to demonstrate that white blood cells are positively associated with risk of end stage renal disease (ESRD) in patients with high blood urea nitrogen, but negatively associated with ESRD in patients with low blood urea nitrogen (see Fig. 4). The multivariate nature of these attributions makes them more informative than the probe-level analyses common in differential expression testing, even when shrinkage estimators are used to "pool information" across genes (Law et al. 2014; Love et al. 2014; Smyth 2004). They are potentially more meaningful to individuals than global estimates such as those provided by knockoffs, since Shapley values are localized to explain particular predictions rather than average behavior throughout an entire feature space. With their model-agnostic flexibility and axiomatic underpinnings, local linear approximators are an attractive tool for genomic research.

Rule lists

Rule lists are sequences of if-then statements, often visualized as a decision tree. Psychological studies have shown that humans can quickly comprehend explanations with such structures, at least when there are relatively few conditions, which is why rule lists are widely promoted as "intrinsically interpretable" (Lage et al. 2018). This accords with the privileged position of material implication in propositional logic, where → is typically regarded as a primitive relation, along with conjunction (∧), disjunction (∨), and negation (¬). These logical connectives form a functionally complete class, capable of expressing all possible Boolean operations. This flexibility, which allows for non-monotonic and discontinuous decision boundaries, affords greater expressive power than linear models. Thus, if the true reason for Alice's unexpected diagnosis lies neither in her BRCA1 allele nor her CXCR6 expression but rather in some non-linear interaction between the two, then she may be better off with a rule list that can concisely explain the (local or global) behavior of that function.

In statistical contexts, rule lists are generally learned through some process of recursive partitioning. For instance, the pioneering classification and regression tree (CART) algorithm (Breiman et al. 1984) predicts outcomes by dividing the feature space into hyperrectangles that minimize predictive error. Computing optimal decision trees is NP-complete (Hyafil and Rivest 1976), but CART uses greedy heuristics that generally work well in practice. Because individual decision trees can be unstable predictors, they are often combined through ensemble methods such as bagging (Breiman 2001a), in which predictions are averaged across trees trained on random bootstrap samples, and boosting (Friedman 2001), in which predictions are summed over a series of trees, each sequentially optimized to improve upon the last. While combining basis functions tends to improve predictions, it unfortunately makes it difficult if not impossible to extract individual rules for better model interpretation. However, some regularization schemes have been developed to post-process complex learning forests for precisely this purpose. For instance, Friedman and Popescu (2008) propose the RuleFit algorithm, which mines a collection of Boolean variables by extracting splits from a gradient boosted forest. These engineered features are then combined with the original predictors in a lasso regression, producing a sparse linear combination of splits and inputs. Nalenz and Villani (2018) develop a similar procedure using a Bayesian horseshoe prior instead of an L_1 penalty to induce

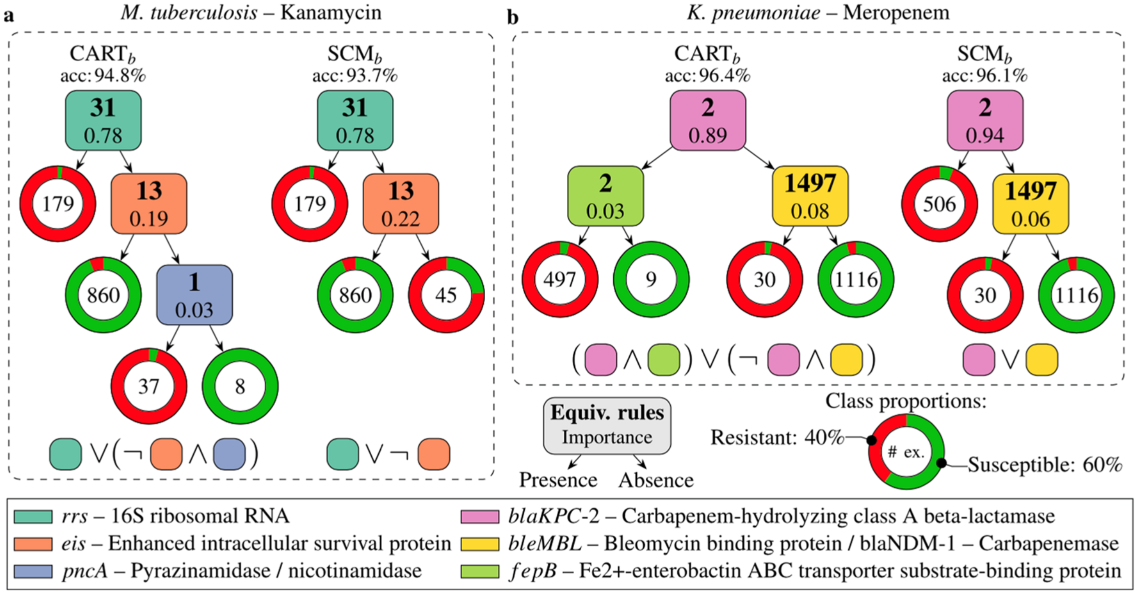

a
M. tuberculosis – Kanamycin
CART_b
acc: 94.8%
31
0.78
179
13
0.19
860
1
0.03
37
8
SCM_b
acc: 93.7%
31
0.78
179
13
0.22
860
45
b
K. pneumoniae – Meropenem
CART_b
acc: 96.4%
2
0.89
2
0.03
1497
0.08
497
9
30
1116
SCM_b
acc: 96.1%
2
0.94
506
1497
0.06
30
1116
Equiv. rules
Importance
Presence
Absence
Class proportions:
Resistant: 40%
# ex.
Susceptible: 60%
rrs – 16S ribosomal RNA
eis – Enhanced intracellular survival protein
pncA – Pyrazinamidase / nicotinamidase
blaKPC-2 – Carbapenem-hydrolyzing class A beta-lactamase
bleMBL – Bleomycin binding protein / blaNDM-1 – Carbapenemase
fepB – Fe2+-enterobactin ABC transporter substrate-binding protein

Fig. 5 Example rule lists for AMR prediction from genotype data. Each rule detects the presence/absence of a k-mer and is colored according to the genomic locus at which it was found. From (Drouin et al. 2019, p. 4)

shrinkage. They also add splits extracted from a random forest with those learned via gradient boosting to promote greater diversity.

Another strand of research in this area has focused on falling rule lists, which create monotonically ordered decision trees such that the probability of the binary outcome $Y = 1$ strictly decreases as one moves down the list. These models were originally designed for medical contexts, where doctors must evaluate patients quickly and accurately. For instance, Letham et al. (2015) design a Bayesian rule list to predict stroke risk, resulting in a model that outperforms leading clinical diagnostic methods while being small enough to fit on an index card. Falling rule lists can be challenging to compute—see the note above about NP-completeness—and subsequent work has largely focused on efficient optimization strategies. Specifically, researchers have developed fast branch-and-bound techniques to prune the search space and reduce training time (Chen and Rudin 2018; Yang et al. 2017), culminating in several tree-learning methods that are provably optimal under some restrictions on the input data (Angelino et al. 2018; Hu et al. 2019).

Less work has been done on localized rule lists, but there have been some recent advances in this direction. Ribeiro et al. (2018) followed up on their 2016 LIME paper with a new method, Anchors, which combines graph search with a multi-armed bandit procedure to find a minimal set of sufficient conditions for a given model prediction. Guidotti et al. (2018a) introduce LORE, which simulates a balanced dataset of cases using a genetic algorithm designed to sample heavily from points near the decision boundary. A decision tree is then fit to the synthetic dataset. Lakkaraju et al. (2019)'s MUSE algorithm allows users to specify particular features of interest. Explanations are computed as compact decision sets within this subspace.

To date, rule lists have not been as widely used in genomics as feature attributions or local linear approximations. This likely has more to do with computational obstacles than any preference for particular model assumptions, per se. Still, some recent counterexamples buck the trend. Drouin et al. (2019) combine sample compression theory with recursive partitioning to learn interpretable genotype-to-phenotype classifiers with performance guarantees. As depicted in Fig. 5, these lists—visualized as trees and formulated as logical propositions below—can predict AMR in *M. tuberculosis* and *K. pneumoniae* with high accuracy using just a small handful of indicator functions over the space of all k-mers. Though their experiments focus on AMR, the method can be applied more generally. Anguita-Ruiz et al. (2020) use a sequential rule mining procedure to uncover gene expression patterns in obese subjects from longitudinal DNA microarray data. Garvin et al. (2020) combined iterative random forests (Cliff et al. 2019), a method for gene regulatory network inference, with random intersection trees (Shah and Meinshausen 2014), which detect stable interactions in tree-based ensembles, to discover potentially adaptive SARS-CoV-2 mutations.

Open challenges

Despite all the recent progress in iML, the field is still struggling with several challenges that are especially important in genomics. I highlight three in particular: ambiguous targets, error rate control, and variable granularity.

Ambiguous targets

I have done my best above to be clear about the distinction between two tasks for which iML is often used: to better explain or understand (a) some fitted model f, or (b) some natural system that f models. It is not always obvious which goal researchers have in mind, yet model- and system-level analyses require entirely different tools and assumptions. Whereas a supervised learning algorithm does not generally distinguish between correlation and causation, the difference is crucial in nature. Clouds predict rain and rain predicts clouds, but the causal arrow runs in only one direction. Genomic researchers face a fundamental ambiguity when seeking to explain, say, why our algorithm diagnosed Alice with BLBC. Is the goal to explain why the classifier made the prediction it did, independent of the ground truth? Or, alternatively, is the goal to understand what biological conditions led to the diagnosis? The former, which I will call a model-level explanation, may be preferable in cases of auditing or validation, where the analyst seeks merely to understand what the algorithm has learned, without any further restrictions. In this case, we do not necessarily assume that the model is correct. The latter, which I will call a system-level explanation, is more useful in cases of discovery and/or planning, where real-world mechanisms cannot be ignored. In such instances, we (tentatively) presume that the prediction in question is accurate, at least to a first approximation.

Model-level explanations are generally easier to compute, since features can be independently perturbed one at a time. This is the default setting for popular iML tools, such as LIME and SHAP. System-level explanations, by contrast, require some structural assumptions about dependencies between variables. Such assumptions may be difficult or even impossible to test, raising legitimate questions about identifiability and under-determination. Yet, as Pearl (2000) has long argued, there is value in articulating one's assumptions clearly, opening them up for scrutiny and debate instead of burying them behind defaults. The last year has seen a burst of new papers on causally-aware iML tools (Heskes et al. 2020; Karimi et al. 2020a, b; Xu et al. 2020), indicating that researchers in computational statistics are increasingly sensitive to the distinction between model- and system-level analyses. Genomic practitioners should avail themselves of both explanatory modes, but always make sure the selected tool matches the stated aim. Addressing this challenge is difficult at both a conceptual level, because the distinction between model- and system-level analyses may not be immediately obvious to practitioners, and at a technical level, because causal approaches can require careful covariate adjustments and data reweighting. The sooner these issues are addressed head on, the more fruitful the results will be.

Error rate control

Another open challenge in iML concerns bounding error and quantifying uncertainty. Bioinformaticians are no strangers to p-values, which are typically fixed at low levels to control false positive rates in GWAS (Panagiotou and Ioannidis 2012), or else adjusted to control familywise error rates (Holm 1979) or FDR (Benjamini and Hochberg 1995) in other omic settings. Bayesians have their own set of inferential procedures for multiple testing scenarios (Gelman et al. 2012; Scott and Berger 2010), although there is some notable convergence with frequentism on the subject of q-values (Storey 2003). In any event, the error-statistical logic that guides testing in computational biology is largely absent from contemporary iML. This may be partially a result of cultural factors. As Breiman (2001b) observed some 20 years ago, there are two main cultures of statistical modeling—one focused on predicting outcomes, the other on inferring parameter values. Authors in contemporary iML, which grew almost exclusively out of the former camp, are generally less worried about error rates than their colleagues in the latter camp.

Several critics have pointed out that post-hoc methods do not generally provide standard errors for their estimates or goodness of fit measures for their approximations (Ribeiro et al. 2018; Wachter et al. 2018). Indeed, it would be difficult to do so without some nonparametric resampling procedure such as the bootstrap (Davison and Hinkley 1997), which would add considerable computational burden as the number of samples and/or features grows. It is not clear that such methods are even applicable in these settings, however, given the instability of bootstrap estimators in high dimensions (Karoui and Purdom 2018). This makes it impossible to reliably rank biomarkers or evaluate the reliability of experimental results. Knockoffs are a notable exception, given their focus on FDR control. However, the Candès et al. (2018) algorithm is not, strictly speaking, a post-hoc iML method, since it requires training an algorithm on an expanded $n \times 2p$ design matrix that includes both the original features and their knockoffs. This is just a preliminary step toward a final model, which contains only those variables that pass some predetermined FDR threshold. The modified method of Watson and Wright (2021) adapts knockoffs for post-hoc importance, but in so doing loses the finite sample error rate guarantees of the original adaptive thresholding procedure.

A handful of other iML methods make at least a nominal effort to quantify uncertainty (Gimenez and Zou 2019; Ribeiro et al. 2018; Schwab and Karlen 2019). Yet these examples are perhaps most notable for their scarcity. To gain more widespread acceptance in genomics—and the sciences more generally—iML algorithms will need to elevate rigorous testing procedures from an occasional novelty to a core requirement. Generic methods for doing so in high dimensions raise complex statistical challenges that remain unresolved at present.

Variable granularity

A final challenge I will highlight concerns variable granularity. This is not a major issue in the low- or moderate-dimensional settings for which most iML tools are designed. But it quickly becomes important as covariates increase, especially when natural feature groupings are either known a priori or directly estimable from the data. For instance, it is well-established that genes do not operate in isolation, but rather work together in co-regulated pathways. Thus, even when a classifier uses gene-level RNA-seq data as input features, researchers may want to investigate the prognostic value of pathways to test or develop new hypotheses. In multi-omic models, where features typically represent a range of biological processes, each measured using different platforms, analysts may want to know not just which variables are most predictive overall, but which biomarkers are strongest within a given class. Interactions across subsystems may also be of particular interest.

Few methods in use today allow users to query a target model at varying degrees of resolution like this, but such flexibility would be a major asset in systems biology. Once again, some exceptions are worth noting. Sesia et al. (2020) introduce a knockoff method for localizing causal variants at different resolutions using well-established models of linkage disequilibrium. This amounts to a global post-hoc feature attribution method for whole-genome sequencing data. The leave-out-covariates (LOCO) statistic (Lei et al. 2018; Rinaldo et al. 2019) can be used to quantify the global or local importance of arbitrary feature subsets, but only at the cost of extensive model refitting. Groupwise Shapley values have been formally described (Conitzer and Sandholm 2004) but are not widely used in practice. Resolving the granularity problem will help iML tools scale better in high-dimensional settings, with major implications for genomics. The problem is complicated, however, by the fact that hierarchical information regarding biomolecular function is not always available. Automated methods for discovering such hierarchies are prone to error, while data-driven dimensionality reduction techniques—e.g., the latent embeddings learned by a deep neural network—can be difficult or impossible to interpret. Promising directions of research in this area include causal coarsening techniques (Beckers et al. 2019; Chalupka et al. 2017) and disentangled representation learning (Locatello et al. 2019; Schölkopf et al. 2021).

Conclusion

The pace of advances in genomics and ML can make it easy to forget that both disciplines are relatively young. The subfield of iML is even younger, with the vast majority of work published in just the last three to five years. The achievements to date at the intersection of these research programs are numerous and varied. Feature attributions and rule lists have already revealed novel insights in several genomic studies. Exemplary methods have not yet seen similar uptake, but that will likely change with better generative models. As datasets grow larger, computers become faster, and theoretical refinements continue to accumulate in statistics and biology, iML will become an increasingly integral part of the genomics toolkit. I have argued that better algorithmic explanations can serve researchers in at least three distinct ways: by auditing models for potential bias, validating performance before and throughout deployment, and revealing novel mechanisms for further exploration. I provided a simple typology for iML and reviewed several popular methodologies with a focus on genomic applications.

Despite considerable progress and rapidly expanding research interest, iML still faces a number of important conceptual and technical challenges. I highlighted three with particular significance for genomics: ambiguous targets, limited error rate control, and inflexible feature resolution. These obstacles can be addressed by iML solutions in a post-hoc or intrinsic manner, with model-agnostic or model-specific approaches, via global or local explanations. All types of iML require further development, especially as research in supervised learning and genomics continues to evolve. Ideally, iML would become integrated into standard research practice, part of hypothesis generation and testing as well as model training and deployment. As the examples above illustrate, this vision is already on its way to becoming a reality.

The future of iML for genomics is bright. The last few years alone have seen a rapid proliferation of doctoral dissertations on the topic—e.g., Greenside (2018). Danaee (2019), Nikumbh (2019), Kavvas (2020), Ploenzke (2020) and Shrikumar (2020)—suggesting that early career academics in particular are being drawn to this highly interdisciplinary area of research. Existing work has been promising, though not without its challenges. As the field continues to gather more data, resources, and brainpower, there is every reason to believe the best is yet to come.

Funding This research was partially funded by ONR Grant N62909-19-1-2096.

Availability of data and material N/A.

Code availability N/A.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Aas K, Jullum M, Løland A (2021) Explaining individual predictions when features are dependent: more accurate approximations to Shapley values. Artif Intell 298:103502

Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6:52138–52160

Altmann A, Toloşi L, Sander O, Lengauer T (2010) Permutation importance: a corrected feature importance measure. Bioinformatics 26(10):1340–1347

Angelino E, Larus-Stone N, Alabi D, Seltzer M, Rudin C (2018) Learning certifiably optimal rule lists for categorical data. J Mach Learn Res 18(234):1–78

Anguita-Ruiz A, Segura-Delgado A, Alcalá R, Aguilera CM, Alcalá-Fdez J (2020) eXplainable artificial intelligence (XAI) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research. PLoS Comput Biol 16(4):e1007792

Azodi CB, Tang J, Shiu S-H (2020) Opening the black box: interpretable machine learning for geneticists. Trends Genet 36(6):442–455

Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE 10(7):1–46

Barber RF, Candès EJ (2015) Controlling the false discovery rate via knockoffs. Ann Stat 43(5):2055–2085

Bareinboim E, Pearl J (2016) Causal inference and the data-fusion problem. Proc Natl Acad Sci 113(27):7345–7352

Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Herrera F (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58:82–115

Bates S, Sesia M, Sabatti C, Candès E (2020) Causal inference in genetic trio studies. Proc Natl Acad Sci 117(39):24117 LP-24126 LP

Beckers S, Eberhardt F, Halpern JY (2019) Approximate causal abstraction. Uncertain Artif Intell 210

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Stat Methodol 57(1):289–300

Bhatt U, Xiang A, Sharma S, Weller A, Taly A, Jia Y, Eckersley P (2020) Explainable machine learning in deployment. Conf Fair Account Trans 648–657

Bica I, Alaa AM, Lambert C, van der Schaar M (2021) From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. Clin Pharmacol Ther 109(1):87–100

Breiman L (2001a) Random forests. Mach Learn 45(1):1–33

Breiman L (2001b) Statistical modeling: the two cultures. Stat Sci 16(3):199–231

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. Taylor & Francis, Boca Raton

Candès E, Fan Y, Janson L, Lv J (2018) Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. J R Stat Soc Ser B Stat Methodol 80(3):551–577

Carrieri AP, Haiminen N, Maudsley-Barton S, Gardiner L-J, Murphy B, Mayes AE, Pyzer-Knapp EO (2021) Explainable AI reveals changes in skin microbiome composition linked to phenotypic differences. Sci Rep 11(1):4565

Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N (2015) Intelligible models for healthcare. In: International conference on knowledge discovery and data mining, pp 1721–1730

Casalicchio G, Molnar C, Bischl B (2019) Visualizing the feature importance for black box models. In: Machine learning and knowledge discovery in databases. Springer International Publishing, Cham, pp 655–670

Cascianelli S, Molineris I, Isella C, Masseroli M, Medico E (2020) Machine learning for RNA sequencing-based intrinsic subtyping of breast cancer. Sci Rep 10(1):14071

Chalupka K, Eberhardt F, Perona P (2017) Causal feature learning: an overview. Behaviormetrika 44(1):137–164

Chen X, Ishwaran H (2012) Random forests for genomic data analysis. Genomics 99(6):323–329

Chen C, Li O, Tao D, Barnett A, Rudin C, Su JK (2019) This looks like that: deep learning for interpretable image recognition. Adv Neural Inf Process Syst 32:8930–8941

Chen C, Rudin C (2018) An optimization approach to learning falling rule lists. In: International conference on artificial intelligence and statistics, pp 604–612

Chia C, Sesia M, Ho C, Jeffrey S, Dionne J, Candès E, Howe R (2020) Interpretable classification of bacterial Raman spectra with knockoff wavelets. arXiv:2006.04937

Cliff A, Romero J, Kainer D, Walker A, Furches A, Jacobson D (2019) A high-performance computing implementation of iterative random forest for the creation of predictive expression networks. Genes 10(12):996

Conitzer V, Sandholm T (2004) Computing Shapley values, manipulating value division schemes, and checking core membership in multi-issue domains. In: Conference on artificial intelligence, pp 219–225

Danaee P (2019) Interpretable machine learning: applications in biology and genomics. Doctoral dissertation, Oregon State University

Das A, Rad P (2020). Opportunities and challenges in explainable artificial intelligence (XAI): a survey. arXiv:2006.11371

Davison AC, Hinkley DV (1997) Bootstrap methods and their application. Cambridge University Press, Cambridge

Drouin A, Letarte G, Raymond F, Marchand M, Corbeil J, Laviolette F (2019) Interpretable genotype-to-phenotype classifiers with performance guarantees. Sci Rep 9(1):4071

Egede LE, Walker RJ (2020) Structural racism, social risk factors, and Covid-19—a dangerous convergence for Black Americans. N Engl J Med 383(12):e77

Eraslan G, Avsec Ž, Gagneur J, Theis FJ (2019) Deep learning: new computational modelling techniques for genomics. Nat Rev Genet 20(7):389–403

Fisher A, Rudin C, Dominici F (2019) All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J Mach Learn Res 20(177):1–81

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232

Friedman JH, Popescu BE (2008) Predictive learning via rule ensembles. Ann Appl Stat 2(3):916–954

Garvin MRT, Prates E, Pavicic M, Jones P, Amos BK, Geiger A, Jacobson D (2020) Potentially adaptive SARS-CoV-2 mutations discovered with novel spatiotemporal and explainable AI models. Genome Biol 21(1):304

Gelman A, Hill J, Yajima M (2012) Why we (usually) don't have to worry about multiple comparisons. J Res Educ Eff 5(2):189–211

Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L (2018) Explaining explanations: an overview of interpretability of machine learning. In: 2018 international conference on data science and advanced analytics, pp 80–89

Gimenez JR, Zou J (2019) Discovering conditionally salient features with statistical guarantees. In: International conference on machine learning, pp 2290–2298

Greenside P (2018) Interpretable machine learning methods for regulatory and disease genomics. Doctoral dissertation, Stanford University

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2018b) A survey of methods for explaining black box models. ACM Comput Surv 51(5):1–42

Guidotti R, Monreale A, Ruggieri S, Pedreschi D, Turini F, Giannotti F (2018a) Local rule-based explanations of black box decision systems. arXiv:1805.10820

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, New York

Heinze-Deml C, Peters J, Meinshausen N (2018) Invariant causal prediction for nonlinear models. J Causal Inference 6(2):20170016

Heskes T, Sijben E, Bucur IG, Claassen T (2020) Causal Shapley values: exploiting causal knowledge to explain individual predictions of complex models. Adv Neural Inf Process Syst 33:4778–4789

Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6(2):65–70

Holzinger A, Langs G, Denk H, Zatloukal K, Müller H (2019) Causability and explainability of artificial intelligence in medicine. Data Min Knowl Discov 9(4):e1312

Hu X, Rudin C, Seltzer M (2019) Optimal sparse decision trees. Adv Neural Inf Process Syst 32:7267–7275

Hubbard AE, Kennedy CJ, van der Laan MJ (2018) Data-adaptive target parameters. In: van der Laan MJ, Rose S (eds) Targeted Learning in data science. Springer, New York, pp 125–142

Hyafil L, Rivest RL (1976) Constructing optimal binary decision trees is NP-complete. Inf Process Lett 5(1):15–17

Imbens GW, Rubin DB (2015) Causal inference for statistics, social, and biomedical sciences: an introduction. Cambridge University Press, Cambridge

Karimi A-H, Barthe G, Schölkopf B, Valera I (2020) A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. arXiv:2010.04050

Karimi A-H, von Kügelgen J, Schölkopf B, Valera I (2020) Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. Adv Neural Inf Process Syst 33:265–277

Karoui NE, Purdom E (2018) Can we trust the bootstrap in high-dimensions? The case of linear models. J Mach Learn Res 19(5):1–66

Kavvas ES, Yang L, Monk JM, Heckmann D, Palsson BO (2020) A biochemically-interpretable machine learning classifier for microbial GWAS. Nat Commun 11(1):2580

Kavvas E (2020) Biologically-interpretable machine learning for microbial genomics. Doctoral dissertation, UC San Diego

Lage I, Chen E, He J, Narayanan M, Gershman S, Kim B, Doshi-Velez F (2018) An evaluation of the human-interpretability of explanation. In: NeurIPS workshop on correcting and critiquing trends in machine learning

Lakkaraju H, Kamar E, Caruana R, Leskovec J (2019) Faithful and customizable explanations of black box models. In: Conference on AI, ethics and society, pp 131–138

Law CW, Chen Y, Shi W, Smyth GK (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol 15(2):R29

Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L (2018) Distribution-free predictive inference for regression. J Am Stat Assoc 113(523):1094–1111

Letham B, Rudin C, McCormick TH, Madigan D (2015) Interpretable classifiers using rules and Bayesian analysis: building a better stroke prediction model. Ann Appl Stat 9(3):1350–1371

Levy JJ, Titus AJ, Petersen CL, Chen Y, Salas LA, Christensen BC (2020) MethylNet: an automated and modular deep learning approach for DNA methylation analysis. BMC Bioinform 21(1):108

Li S, Sesia M, Romano Y, Candès E, Sabatti C (2021) Searching for consistent associations with a multi-environment knockoff filter. arXiv:2106.04118

Linardatos P, Papastefanopoulos V, Kotsiantis S (2021) Explainable AI: a review of machine learning interpretability methods. Entropy 23(1):18

Locatello F, Bauer S, Lucic M, Raetsch G, Gelly S, Schölkopf B, Bachem O (2019) Challenging common assumptions in the unsupervised learning of disentangled representations. In: International conference on machine learning

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15(12):550

Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30:4765–4774

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Lee S-I (2020) From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2(1):56–67

Marcinkevičs R, Vogt JE (2020). Interpretability and explainability: a machine learning zoo mini-tour. arXiv:2012.01805

Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ (2019) Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet 51(4):584–591

Mazurowski MA, Buda M, Saha A, Bashir MR (2019) Deep learning in radiology: an overview of the concepts and a survey of the state of the art with focus on MRI. J Mag Reson Imaging 49(4):939–954

McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Shetty S (2020) International evaluation of an AI system for breast cancer screening. Nature 577(7788):89–94

Meinshausen N, Hauser A, Mooij J, Peters J, Versteeg P, Bühlmann P (2016) Methods for causal inference from gene perturbation experiments and validation. Proc Natl Acad Sci 113(27):7361–7368

Merrick L, Taly A (2020) The explanation game: explaining machine learning models using Shapley values. In: Machine learning and knowledge extraction, pp 17–38

Milioli HH, Tishchenko I, Riveros C, Berretta R, Moscato P (2017) Basal-like breast cancer: molecular profiles, clinical features and survival outcomes. BMC Med Genom 10(1):19

Mohseni S, Zarei N, Ragan ED (2020) A multidisciplinary survey and framework for design and evaluation of explainable AI systems. arXiv:1811.11839

Molnar C (2019) Interpretable machine learning: A guide for making black box models explainable. https://christophm.github.io/interpretable-ml-book/

Montavon G, Lapuschkin S, Binder A, Samek W, Müller KR (2017) Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognit 65(May 2016):211–222

Mueller ST, Hoffman RR, Clancey W, Emrey A, Klein G (2019). Explanation in human-AI systems: a literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. arXiv:1902.01876

Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B (2019) Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci 116(44):22071–22080

Nalenz M, Villani M (2018) Tree ensembles with rule structured horseshoe regularization. Ann Appl Stat 12(4):2379–2408

Nembrini S, König IR, Wright MN (2018) The revival of the Gini importance? Bioinformatics 34(21):3711–3718

Nicholls HL, John CR, Watson DS, Munroe PB, Barnes MR, Cabrera CP (2020) Reaching the end-game for GWAS: machine learning approaches for the prioritization of complex disease loci. Front Genet 11:350

Nikumbh S (2019) Interpretable machine learning methods for prediction and analysis of genome regulation in 3D. Doctoral dissertation, Saarland University

Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464):447–453

Panagiotou OA, Ioannidis JPA (2012) What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. Int J Epidemiol 41(1):273–286

Pearl J (2000) Causality: models, reasoning, and inference. Cambridge University Press, New York

Pearl J, Bareinboim E (2014) External validity: from do-calculus to transportability across populations. Stat Sci 29(4):579–595

Peters J, Bühlmann P, Meinshausen N (2016) Causal inference by using invariant prediction: identification and confidence intervals. J R Stat Soc Ser B Stat Methodol 78(5):947–1012

Peters J, Janzing D, Schölkopf B (2017) The elements of causal inference: foundations and learning algorithms. The MIT Press, Cambridge

Ploenzke M (2020) Interpretable machine learning methods with applications in genomics. Doctoral dissertation, Harvard University

Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?": explaining the predictions of any classifier. In: International conference on knowledge discovery and data mining, pp 1135–1144

Ribeiro MT, Singh S, Guestrin C (2018) Anchors: high-precision model-agnostic explanations. In: Association for the advancement of artificial intelligence, pp 1527–1535

Rinaldo A, Wasserman L, G'Sell M (2019) Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. Ann Stat 47(6):3438–3469

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215

Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C (2021) Interpretable machine learning: fundamental principles and 10 grand challenges. Stat Surv

Sarkar JP, Saha I, Sarkar A, Maulik U (2021) Machine learning integrated ensemble of feature selection methods followed by survival analysis for predicting breast cancer subtype specific miRNA biomarkers. Comput Biol Med 131:104244

Schölkopf B, Tsuda K, Vert J-P (eds) (2004) Kernel methods in computational biology. The MIT Press, Cambridge

Schölkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, Goyal A, Bengio Y (2021) Toward causal representation learning. Proc IEEE 109(5):612–634

Schwab P, Karlen W (2019) CXPlain: causal explanations for model interpretation under uncertainty. Adv Neural Inf Process Syst 32:10220–10230

Scott JG, Berger JO (2010) Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. Ann Stat 38(5):2587–2619

Selbst A, Powles J (2017) Meaningful information and the right to explanation. Int Data Priv Law 7(4):233–242

Sesia M, Sabatti C, Candès E (2019) Gene hunting with hidden Markov model knockoffs. Biometrika 106(1):1–18

Sesia M, Bates S, Candès E, Sabatti C (2020) Multi-resolution localization of causal variants across the genome. Nat Commun 11(1):1093

Shah RD, Meinshausen N (2014) Random intersection trees. J Mach Learn Res 15(20):629–654

Shapley L (1953) A value for n-person games. In: Contributions to the theory of games, pp 307–317

Shrikumar A, Greenside P, Kundaje A (2017) Learning important features through propagating activation differences. In: International conference on machine learning

Shrikumar A (2020) Interpretable machine learning for scientific discovery in regulatory genomics. Doctoral dissertation, Stanford University

Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3(1)

Sonnenburg S, Zien A, Philips P, Rätsch G (2008) POIMs: positional oligomer importance matrices—understanding support vector machine-based signal detectors. Bioinform 24(13):i6–i14

Sørlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Botstein D (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci 100(14):8418 LP-8423 LP

Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, Collins JJ (2020) A deep learning approach to antibiotic discovery. Cell 180(4):688-702.e13

Storey JD (2003) The positive false discovery rate: a Bayesian interpretation and the q -value. Ann Stat 31(6):2013–2035

Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A (2008) Conditional variable importance for random forests. BMC Bioinform 9(1):307

Sundararajan M, Najmi A (2019) The many Shapley values for model explanation. In: Proceedings of ACM conference. ACM, New York

Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. In: International conference on machine learning

Talukder A, Barham C, Li X, Hu H (2021) Interpretation of deep learning in genomics and epigenomics. Brief Bioinform 22(3):177

Tasaki S, Gaiteri C, Mostafavi S, Wang Y (2020) Deep learning decodes the principles of differential gene expression. Nat Mach Intell 2(7):376–386

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Stat Methodol 58(1):267–288

Tideman LEM, Migas LG, Djambazova KV, Patterson NH, Caprioli RM, Spraggins JM, Van de Plas R (2021) Automated biomarker candidate discovery in imaging mass spectrometry data through spatially localized Shapley additive explanations. Anal Chim Acta 1177:338522

Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25(1):44–56

Treppner M, Binder H, Hess M (2021) Interpretable generative deep learning: an illustration with single cell gene expression data. Hum Genet

Turner NC, Reis-Filho JS (2006) Basal-like breast cancer and the BRCA1 phenotype. Oncogene 25:5846

Vapnik V (1995) The nature of statistical learning theory. Springer, New York

Vidovic MM-C, Görnitz N, Müller K-R, Rätsch G, Kloft M (2015) SVM2Motif: reconstructing overlapping DNA sequence motifs by mimicking an SVM predictor. PLoS ONE 10(12):e0144782–e0144782

Vidovic MM-C, Kloft M, Müller K-R, Görnitz N (2017) ML2Motif: reliable extraction of discriminative sequence motifs from learning machines. PLoS ONE 12(3):e0174392–e0174392

Vilone G, Longo L (2020) Explainable artificial intelligence: a systematic review. arXiv:2006.00093

Wachter S, Mittelstadt B, Floridi L (2017) Why a right to explanation of automated decision-making does not exist in the general data protection regulation. Int Data Priv Law 7(2):76–99

Wachter S, Mittelstadt B, Russell C (2018) Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harvard J Law Technol 31(2):841–887

Waldmann P, Mészáros G, Gredler B, Fürst C, Sölkner J (2013) Evaluation of the lasso and the elastic net in genome-wide association studies. Front Genet 4:270

Watson D, Krutzinna J, Bruce IN, Griffiths CEM, McInnes IB, Barnes MR, Floridi L (2019) Clinical applications of machine learning algorithms: beyond the black box. BMJ 364:446–448

Watson D, Floridi L (2021) The explanation game: a formal framework for interpretable machine learning. Synthese 198(10):9211–9242

Watson D, Wright M (2021) Testing conditional independence in supervised learning algorithms. Mach Learn 110(8):2107–2129

Williamson BD, Gilbert PB, Carone M, Simon N (2021) Nonparametric variable importance assessment using machine learning techniques. Biometrics 77(1):9–22

Woodward J (2019) Scientific Explanation. In: Zalta EN (ed) The Stanford encyclopedia of philosophy (winter 201). Metaphysics Research Lab, Stanford University, Stanford

Xie YR, Castro DC, Bell SE, Rubakhin SS, Sweedler JV (2020) Single-cell classification using mass spectrometry through interpretable machine learning. Anal Chem 92(13):9338–9347

Xu G, Duong TD, Li Q, Liu S, Wang X (2020) Causality learning: a new perspective for interpretable machine learning. arXiv:2006.16789

Yang H, Rudin C, Seltzer M (2017) Scalable Bayesian rule lists. In: International conference on machine learning

Yap M, Johnston RL, Foley H, MacDonald S, Kondrashova O, Tran KA, Waddell N (2021) Verifying explainability of a deep learning tissue classifier trained on RNA-seq data. Sci Rep 11(1):2641

Zhang Z, Chen P, McGough M, Xing F, Wang C, Bui M, Yang L (2019) Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. Nat Mach Intell 1(5):236–245

Zhao Q, Hastie T (2021) Causal interpretations of black-box models. J Bus Econ Stat 39(1):272–281

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol 67(2):301–320