Research on Machine Learning for Drug Discovery and Precison Medicine

ZHAO, Kai

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy
in
School of Biomedical Sciences

Supervised by

Prof. So Hon-cheong

The Chinese University of Hong Kong
July 2020

Thesis Assessment Committee

Professor Cheng Sze Lok Alfred (Chair)

Professor So Hon-cheong (Thesis Supervisor)

Professor Chen Yangchao (Committee Member)

Professor Sham Pak Chung (External Examiner)

Abstract of thesis entitled:

Research on Machine Learning for Drug Discovery and Precison Medicine

Submitted by ZHAO, Kai

for the degree of Doctor of Philosophy

at The Chinese University of Hong Kong in July 2020

The recent years have witnessed a rapid advance in machine learning (ML) algorithms and increasing availability of biomedical data. There is great opportunity to integrate the advances both fields to benefit healthcare. One of the long-standing problem in biomedical science is the high cost and high failure rate of drug development. Another related problem is that the same drug or treatment may not have the same effect on every patient, yet it is difficult to precisely estimate the benefit of treatment for each person based on his/her background.

In this thesis, we seek to address these concerns by application of ML methods on biomedical data, especially "omics" data. In this study, we proposed a computational drug repurposing framework using drug expression profiles, leveraging ML methods. Expression data was used as predictors and drug indication was considered as outcomes. We found that the method can 're-discover' known drugs for treatment, demonstrating its usefulness.

Drug repurposing however is not always available, and uncovering valid drug targets for specific diseases remain a major challenge. Drug development suffers from high failure rate largely due to the wrong target pursued. To tackle this issue, we proposed a computational framework to identify promising drug targets for further studies, without using knowledge of the known drug targets. We showed that the identified targets from our method were enriched for targets identified from different lines of evidence.

Precision medicine has been advocated in the past years. One of the major aims is to estimate *individualized* treatment effects (ITE), i.e. the effect of a treatment (or risk factor) for each person based one's clinical/genetic background. Here we em-

ployed tree-based methods to achieve the goal with consideration of each subject's clinical and genetic information. We also proposed an 'imputation' approach to incorporate time-to-event data. Additionally, since there are no well-established methods to evaluate model fitting, we proposed several statistical methods to address this issue. To examine the validity of our framework, we carried out simulation studies, which showed that our proposed statistical methods maintained good type I error control and had strong power in detecting effect heterogeneity. We also applied our approach to GWAS data of COVID-19 to study ITE of clinical variables on the severity of COVID-19, given genetic and other clinical variables as covariates. Taken together, we hope the proposed methods will open new avenues for translating genetic data into clinical applications.

Acknowledgement

I would like to thank my supervisor Prof. So Hon-cheong for offering me the opportunity to work with him. He teaches me how to conduct research works and gave me generous help in my daily life. He is humorous and highly empathic. Truth to be told, it is a wonderful journey to study and work with him, and I am lucky to have the experience.

I also would like to thank my wife. It's nearly ten years since the first meet, and she has become the most important person in my life. Thanks for supporting my decision to pursue a higher degree and bearing all burden from family to free me from distractions. This is not easy to her. You are a brave girl and wonderful mother for the kids. My any achievement is impossible without you.

Thanks my father and mother for never saying NO to the decision of further my education and for taking care of my kids in the past years. I know the hardness for you to bear the burden. I highly appreciate the support.

Thanks my kids for tolerating my absence of parental responsibility for the past year. You let me be your father and share tremendous joy with me. I promise my love to you will alway be the same.

Thanks my lab mates for having some awesome years with you. You are a part of my daily life in those years. The times we spent together will be a precious memory.

Thanks everybody who helped me for their kindness!

Dedicated to those who risked their life to fight against COVID-19.

Contents

List of Figures

List of Tables

Symbols and Acronyms

In general, we denote a scalar by an italic lower case letter, a vector by a roman lower case bold letter, and a matrix by a roman upper case bold letter respectively, e.g., $a \in \mathbb{R}$, $\mathbf{v} \in \mathbb{R}^n$ and $\mathbf{M} \in \mathbb{R}^{p \times q}$, with any exceptions to be mentioned in the context case by case.

An identity matrix is written as $\mathbf{I}$. Specifically, an $n \times n$ identity matrix is written as $\mathbf{I}_n$. A zero matrix or vector is written as $\mathbf{0}$. Specifically, an $m \times n$ zero matrix is written as $\mathbf{0}_{m \times n}$.

Specialized symbols and major acronyms are defined as follows:

$\mu_0(\mathbf{x})$	a function for control group $f(\mathbf{x}, w = 0)$
$\mu_1(\mathbf{x})$	a function for treatment group $f(\mathbf{x}, w = 1)$
$p(\cdot)$	the probability density function (PDF)
$\Pr(\cdot)$	the probability value
$\mathbb{E}(\cdot)$	the expectation
$\mathbf{\Sigma}$	a covariance matrix
$\mathbf{N}(\mu, \mathbf{\Sigma})$	a normal distribution with mean μ and covariance $\mathbf{\Sigma}$
ε	a noise vector
$\mathbf{e}(\cdot)$	an error/residual function
$\mathbf{H}$	Hessian matrix
$\mathrm{tr}(\cdot)$	trace of a matrix
$\det(\cdot)$	determinant of a matrix

DNN	deep neuron network
GBM	gradient boosting machine
SVM	support vector machine
CF	causal forests
RF	risk factors
ML	machine learning
EN	elastic net
ITE	individualized treatment effects
TE	treatment effects
tx	treatment
CM	cardiometabolic
GWAS	genome-wide association studies
CNV	copy number variation
SML	supervised machine learning
MCMC	Markov chain Monte Carlo
GRF	general causal forests
CF	causal forests
CV	cross validation
MSE	mean squared error
CI	confidence interval

Chapter 1

Introduction

Machine learning (ML) is one of the fastest growing fields in science in the past decade. At the same time, the rapid accumulation of 'omics' and other forms of biomedical data have revolutionized the field. One of the most pressing questions to date is how to make use of modern ML methodologies and ever-growing data from biomedical sciences to improve our understanding of human diseases and develop better treatments. In this thesis, we will develop and apply several ML approaches to drug discovery/repositioning and predicting individualized effects of risk factors and treatments.

1.1 Drug Discovery Today

New drug development is a lengthy and costly process, and a recent study reported an average cost of $\sim$ 2558 million US dollars in developing a new drug [48]. Part of the reason of the high cost is due to the high failure rate of preclinical drug candidates, which is largely caused by lack of efficacy of these candidates; this indicates that the wrong target is pursued [185]. Computational drug repositioning and target discovery may serve as a new way to shorten the process of drug development, due to the lower cost and more established safety profile of existing drugs [50]. A number of

in silico approaches have been developed for drug repositioning and target discovery and are reviewed elsewhere [94, 215, 107]. With the rapid rise of machine learning (ML) technologies in the past decade, there has been a rising interest in applying ML methods in drug repositioning or target discovery.

Machine learning refers to a vast number of methods for computers to ”learn” and gain insight into data without human interference. These methods are classified into two categories: supervised and unsupervised. Supervised machine learning methods are models for prediction or estimation based on one or more inputs. They are called supervised methods, as their learning is ”supervised” by known output values. On the other hand, unsupervised machine learning methods can be used to detect relationship or patterns underlying ”unlabeled” data. Here we focus on supervised learning methods for classification, since in most cases studies related to drug discovery using ML are the application of classification algorithms. One approach to drug repurposing is to employ drug expression profiles as predictors (i.e., features) to predict a drug's treatment potential. The outcome variable can be the drug category (e.g., whether it is a cardiovascular or anticancer agent) or whether the drug is indicated for a particular disorder (e.g., whether the drug is indicated for diabetes). In the former case, drugs that are classified into categories other than its own indications may be considered for repositioning. In the latter case, drugs with high predicted probabilities but not indicated for the disorder may serve as candidates for repositioning. Additionally, we may also interested in whether the expression profile induced by genetic perturbation (e.g. over-expression or knock down) showed similar pattern to expression profiles of drugs considered as treatments for specific disease, since in this case the perturbed gene can be considered as promising target for the disease. Note that indications for drugs can easily be obtained from publicly available resources such as the Anatomical Therapeutic Chemical (ATC) Classification System [226]. An important advantage is that ML algorithms are abundant and in rapid development, and any existing or new algorithms can be applied without much modification.

With the growth of availability of biomedical data, especially “omics”, computational methods can offer a fast, cost-efficient and systematic way to priority promising drug target and drug repurposing candidates for various diseases. The approach has several advantages. Specifically, finding new indications for existing drugs, an approach known as drug repositioning or repurposing, can serve as a useful strategy to shorten the development cycle. Repurposed drugs can be brought to the market in a much shorter time-frame and at lower costs. Meanwhile, computational drug target identification also can speed up the drug development by prioritizing the most promising drug target candidates in a short time, greatly reducing the time in seeking for potential drug targets.

There has been increasing interest in computational drug repositioning recently, in view of the rising cost of new drug development. Hodos et al. provided a comprehensive and updated review on drug repositioning [94], and G. Kandoi et. al. briefly reviewed applications of machine learning and system biology on discovery of target proteins [107]. For the purpose of repositioning, similarity-based methods [76, 157, 131, 137, 151, 124] usually were employed to explore repositioning opportunities, but as noted by Hodos et al., the dependence on data in the “nearby pharmacological space” might limit the ability to find medications with novel mechanisms of actions. Another related methodology is the network-based approach [133], which typically requires data on the relationship between drugs, genes and diseases as well as connections within each category (e.g. drug-drug similarities). It still constraints by the focus on a nearby pharmacological space and the choice of tuning parameters in network construction or inference is often ad hoc [58]. The present work is different in that we apply a broad framework for repositioning and we do not focus on one but many different kinds of learning methods. There is comparatively less reliance on known drug mechanisms or the ”nearby pharmacological space” as we let the different algorithms ”learn” the relationship between drugs, genes and disease in their own ways. We note that kernel-based ML methods such as support vector machine

(SVM) are also based on some sort of similarity measures. A related work [151] have also examined SVM as a promising ML approach for drug repositioning and identified several interesting candidates. However, here our focus is different in that we considered a variety of other approaches and SVM is one of the methods which falls under the broader framework of ML for repositioning.

Although computational drug repositioning has attracted increased attention recently, few studies focus on psychiatric disorders, compared to other areas like oncology. Psychiatric disorders are leading causes of disability worldwide [219], however there have been limited advances in the development of new pharmacological agents in the last two decades or so [99]. Development of new therapies is also limited by the difficulty of animal models to fully mimic human psychiatric conditions [153]. Investment by drug companies has in general been declining [99] , and new approaches for drug discoveries are very much needed in this field. We will explore repositioning opportunities for schizophrenia along with depression and anxiety disorders. Here depression and anxiety disorders are analyzed together as they are highly clinically comorbid [110, 160], show significant genetic correlations[160], and share similar pharmacological treatments [16].

Meanwhile, we also have witnessed a rise in the interest of computational target discovery in recent years. G. Kandoi et. al. briefly reviewed applications of machine learning and system biology on discovery of target proteins [107]. These studies explored different biological properties by machine learning methods to identify druggable targets [15, 56, 125]. Biological features of human proteins like amino acid composition and amino acid property group composition were studied by a sequence-based prediction method to identify drug target proteins, and a comprehensive comparison of several machine learning methods was conducted [116]. In another study, eight key properties of human drug target were extracted, and learned by support vector machine (SVM) to discover new targets; similar studies extracted simple physicochemical properties from known drug targets and explored the predictive power of

these properties [125, 54]. Topological features of human protein–protein interaction network also were utilized by network based methods to identify potential drug targets [126]. In a recent study, gene-disease association data from Open Targets was explored by four different machine learning methods, including deep neuron networks, to find novel targets, and a large proportion of new targets identified were supported by previous literatures [58]. Dorothea Emig et. al. proposed an integrated network-based method to predict drug targets based on disease gene expression profiles and a high-quality interaction network, and some novel drug targets for scleroderma and other types of cancer were presented [54]. A most recent study proposed pairwise learning and joint learning methods constructed on chemically and genetically perturbed gene expression profiles, and outcome variable was defined as highly correlated pair given by the direct correlation calculation [179]. These studies aim to discover new targets by making use of structural attributes of proteins or properties of known targets, so targets with similar properties usually are identified, but drug targets with novel mechanisms are difficult to identify using this kind of approaches. Network based methods for target discovery, as mentioned previous, rely on known nearby targets to inference potential relationship, so they suffers the same drawback.

Even though repositioned drugs can be brought into market in a much shorter time, drug repositioning may not always be feasible (for example due to side-effects of existing drugs), and drug repositioning and target discovery can complement each other in drug development and pharmacological research. Additionally, in traditional drug development majority of drugs fail to complete the development process due to lack of efficacy, indicating that the wrong target is pursued [185].Usually, a drug target is selected via analyzing how its function influences the disease. However, this process is time-consuming, because investigating a large number of potential targets is usually necessary for finding an ideal one [168]. Computational methods can be utilized to hasten this process by prioritizing promising drug targets.

1.2 Precision Medicine

Researchers have discovered hundreds of genes that harbor variations contributing to human illness. For example, genome wide association studies have identified 108 schizophrenia-associated genetic loci [173]. Genetic variability leads to patients response differently to dozens of treatments, and molecular causes of some diseases have been targeted in research. Direct-to-consumer (DTC) companies take advantage of DNA tests to provide genetic insights into personal traits and disease risks [154]. The genetic testing can improve disease prevention [87]. Moreover, scientists have been beginning to utilize genetics or other molecular mechanisms to better predict patients' responses to targeted therapy [84]. Particularly, there are successful examples in translation of cancer genomics into therapeutics and diagnostics. For instance, doctors recommend KRAS mutation testing for patients with colon or lung cancer before the initiation of EGFR-targeted therapies; moreover, patients with RAS mutations are more likely to benefit from pharmacological inhibition of the kinases MEK1 and 2 [127]. Another clinically relevant question to individuals is how a risk factor will affect them individually. A more recent example is that the pandemic of coronavirus-19 has been attacking people globally, but only a minority of infected patients, including young adults, have respiratory failure [53]. There are numerous evidences to demonstrate the fact that different individuals response differently to the same risk factor. For example, not all obese persons develop cardiometabolic (CM) diseases though obesity is a well-known risk factor to the diseases [152]. As another example, cigarette smoking is the top one risk factor for chronic obstructive pulmonary disease (COPD), but we still can find elder people with smoking habits do not develop COPD [51]. Conceptually, risk factors or treatments are equivalent, as risk factors can be considered as treatments with adverse effects. Exploration of the heterogeneity in subjects' outcomes in presence of risk factors is of great importance, since it allows us to offer tailored health management to each subjects. This enable us to offer personalized prevention or treatment strategies in a cost-effective way to

benefit them the most. This is also the aim of "personalized medicine".

The cost of genome sequencing have been decreased dramatically in the past decades. This increases the wide availability of "omics" data. On the advent of genomic ara, genetic data can be utilized to customize disease prevention and medical treatment. Clinically, there are examples of success of personalized medicine, especially in cancer treatment. Chemotherapy medications such as trastuzumab and imatinib target specific cancers [69]. Despite all of these advances, only for a few diagnoses and treatment resort to patient's genetics information nowadays, end even if doctors can access to patients' genomes today, only a small percentage of the genome has been utilized [235]. Doctors still exercise 'one drug fits all' approach in the clinic. The key to 'personalized medicine' is to answer the most crucial concerns that how a RF or treatment will affect patients individually given their genetic and clinical information. The answer to this concern can advance our disease prevention and treatment. However, current researches on this issue largely focus on the average treatment effect of RFs in population rather than individualized treatment effects.

1.3 Supervised Machine Learning Methods

In this section we first define a general framework for supervised machine learning methods and then discuss several popular machine learning methods in detail. Let X represent the real-valued input matrix (with dimension $n \times p$), where n and p denote the sample size and the number of features, respectively. Y (with dimension n) denotes a random output vector. The subscript i refers to the input or output of the ith observation. Our aim is to seek a function $f(x)$ for estimating Y given the input X. A loss function L is required to penalize the error made in the prediction. Thus, we choose f that minimizes the following the equation [65]:

$$\mathrm{EPE}(f) = L(f(X,Y)). \tag{1.1}$$

Here EPE stands for the expected prediction error.

Linear Methods

The linear model is a simple and intuitive ML approach for regression and classification. In the case of biological systems, the true relationship underlying the data is often nonlinear. For example, in the current application, different genes may act in a complex and nonlinear manner to affect the potential of treatment. However, it can be regarded as a benchmark for the development of more sophisticated ML models. Basically, it assumes that the function f we seek is linear:

$$f(X) = X\beta \tag{1.2}$$

Here we assume the additional column with all 1 is added as the first column of X for the ease of the equation representation; thus, the dimension of X is $n \times (1+p)$. β is a vector of coefficients, with the first element denoting the intercept. When the output Y is real-valued, the typical loss function used is the squared error loss. This leads to a criterion for finding the optimal β, which minimizes the loss function below:

$$\text{EPE} = (X\beta - Y)^T (X\beta - Y) \tag{1.3}$$

However, the optimal coefficients β chosen in a simple liner model as listed above may not yield the best predictive performance on new datasets, especially when the input is high-dimensional with low signal-to-noise ratio. One reason for this is that some noise may be learned by our model (i.e., the model "overfits"), leading to poor performance when applied in a new dataset. In the present application, transcriptome data is of high dimension, and it is reasonable to suspect that only a portion of the input genes may have significant effects on the potential of disease treatment.

To overcome the above issue, regularized regression models can be used to do feature selection. In essence we penalize large values of β to make the model less

complex and less prone to overfitting, thus leading to prediction based on highly influential genes. The ridge penalty leads to coefficient shrinkage which can reduce the risks of model overfitting [95].

However, some features may have little or no association with the output, and filtering out these features may make the interpretation easier and improve predictive performance. Another method known as LASSO (least absolute shrinkage and selection operator) can shrink coefficients down to zero [208], creating a sparse model with fewer features.

In biomedical applications, some features (such as expression of genes in the same pathway) tend to be highly correlated. LASSO usually select one or several features from a group but ignore the others from the same group. Elastic net, a more advanced penalized regression method, may overcome this problem by combining ridge and LASSO penalties [243]. The function we seek to minimize is as follows:

$$\text{EPE} = (X\beta - Y)^T(X\beta - Y) + \lambda\left[\alpha||\beta||_{l_1} + \frac{1}{2}(1-\alpha)||\beta||_{l_2}\right] \tag{1.4}$$

where $||\beta||_{l_1}$ denotes the L1 norm (i.e., sum of absolute values of β) and $||\beta||_{l_2}$ denotes the L2 norm (i.e., sum of the squared β) and λ and α are tuning parameters. The elastic net regularization is a combination of L1 and L2 shrinkage. Ridge and LASSO regression are special cases of elastic net regression when α is 0 and 1, respectively.

In regression, the output of linear model is a real number, but in classification the output of linear model should be a probability within the interval between 0 and 1. In classification, the loss function to be minimized is usually the cross entropy instead of squared error loss [26]. A logistic regression model is commonly used to predict the probability of a binary outcome y_i:

$$\Pr(y_i = 1) = 1/(1 + e^{f(x_i)}) \tag{1.5}$$

Here $f(x)$ denotes $x_i\beta$, where the coefficient β is a parameter of the prediction

function. The loss function for binary classification is

$$\text{EPE} = -\sum_i \left[y_i \log \Pr(x_i) + (1 - y_i) \log(1 - \Pr(x_i)) \right] \tag{1.6}$$

where i refers to the sample index. The three kinds of regularization methods mentioned above can also be applied in similar manners to logistic regression models. A number of studies have adopted linear models for drug repurposing or discovery. A recent work [230] attempted to discover novel therapeutic properties of drugs from transcriptional responses as a multi-label classification problem and reported that multi-label logistic regression showed superior performance over other methods such as random forest (RF) and convolutional neural networks (CNN). Another study employed logistic regression to predict therapeutic indications and side effects from various drug properties such as chemical structure and protein targets [222].

Linear models are computationally fast and intuitive, and can be readily implemented in various programming languages and statistical software. For example, the R package "glmnet" enables fast implementation of regularized linear models, and a detailed documentation and vignette is available online [64]. Linear models are also easy to interpret, as the importance of features may be judged from the magnitude of regression coefficients, and recently methods for assessing statistical significance have also been developed [132]. However, linear models only capture linear relationship between input features and output variable(s), which may not be the case in many real-life scenarios including biomedical applications. In one of our recent works [240], drug repurposing using elastic net in general performed not as well as other nonlinear classifiers, but the ease of interpretation is an advantage. The selected features and magnitude of regression coefficients provided useful information concerning which genes contributed to the drug actions.

Tree-Based Methods

Classification and Regression Tree (CART) is another important type of ML model for classification and regression [29, 28]. The two most popular applications of tree-based models are random forest and gradient boosting machine, which are ensemble ML models that generally outperformed simple CARTs [28, 63]. We first discuss how to construct a decision tree given input X and output Y.

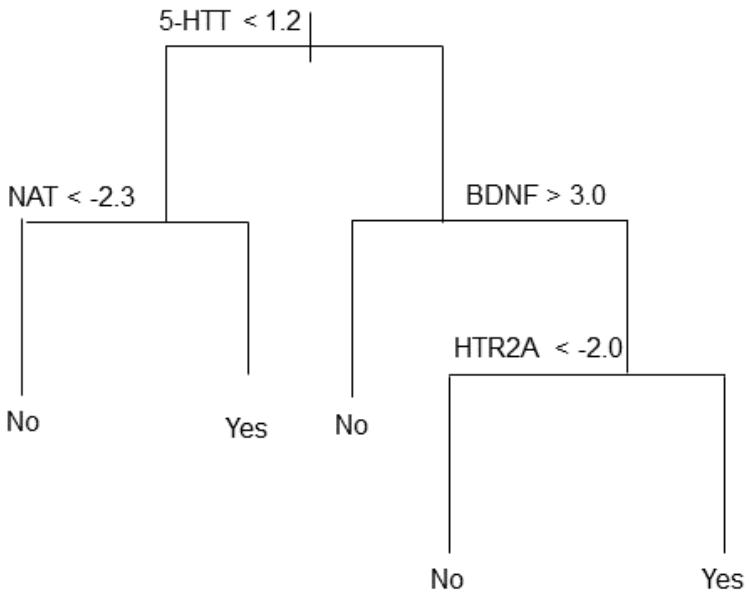

Figure 1.1: A single decision tree in which drug-induced gene expression data are used to predict treatment effects

Figure 1.1 shows a single decision tree on fake drug expression data. In brief, for each time of splitting, we select a variable according to certain criteria and find a cutoff value of that variable to minimize the current loss. To grow a decision tree, the algorithm recursively splits the feature space of training data, and it stops when each leaf node has less than a minimum number of observations or the tree reaches the maximum depth. The *Gini index*, a typical criterion used to make binary splits,

measures the impurity of each node and is defined by [103]

$$G = \sum_{m=1}^{T} \sum_{k=1}^{K} \hat{p_{mk}}(1 - \hat{p_{mk}}), \tag{1.7}$$

which measures total variance of binary classes for each leaf node. Here T refers to the number of leaf nodes of a tree, K denotes the number of classes, and $\hat{p_{mk}}$ represents the fraction of training observations in the mth region that are from the kth class. For binary classification K is 2.

For tree-based regression, the procedure to grow a tree is similar to classification, but the criterion to make a binary split is different, which is [103]

$$\text{EPE} = \sum_{m=1}^{T} \sum_{x_i \in r_m} (y_i - \hat{y_{r_m}})^2, \tag{1.8}$$

Similarly, T is the number of leaf nodes, r_m is the mth leaf node, and $\hat{y_{r_m}}$ is the mean response of training observations in the r_m region. Like linear models, penalty can also be imposed to reduce the complexity of tree to build models with lower variance. One form of regularization is to control the number of leaf nodes [103]:

$$G = \sum_{m=1}^{T} \sum_{k=1}^{K} \hat{p_{mk}}(1 - \hat{p_{mk}}) + \alpha T \tag{1.9}$$

Note that in regression the average output of training observations falling into a leaf node can be regarded as the predicted value; in classification the probability of a class is estimated by the fraction of observations belonging to the class in the leaf node.

A single decision tree usually suffers from high variance which leads to poor predictive performance. Also, some observations may be predicted worse than others. To alleviate the problems of predicting with a single tree, "combining" many trees trained on different subsets of training data might improve predictive performances.Bagging, random forest, and boosting are powerful tools using this idea.

Bagging, or bootstrap aggregation, is a procedure to reduce the variance of tree-based methods by averaging estimations from models trained on a number of training sets sampled by bootstrap. Observations are drawn with replacement in a bootstrap procedure. The prediction from bagging (for regression) is given by [103]

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f^{*b}}(x) \tag{1.10}$$

where B is the number of trees ensembled. For qualitative outputs, a majority vote can be taken to determine the predicted classes, i.e., the most commonly occurring class among the B estimators for each observation.

Random forest (RF) may be considered as a modified version of standard bagging. In random forest, only a subset of features is considered at each candidate split. Usually, features chosen for splitting are a minority of total features, and typically $m \approx \sqrt{p}$ (where p is the total number of features) is chosen in practice [103]. This aims to reduce the correlation between different trees, as aggregating many uncorrelated trees will benefit from a larger variance reduction than aggregating trees that are highly correlated.

Gradient boosting is a general ML approach which aims at combining weak learners to produce an improved prediction model and is most often applied to decision trees [63]. Unlike bagging and random forest, boosting grows trees sequentially. In essence the algorithm tries to improve the model sequentially via fitting a learner to the residuals (or pseudo-residuals) from the previous model. Boosting for classification tree was first proposed by Freund and Schapire in [62], based on the idea of growing new trees by emphasizing more on observations poorly learned by previous trees. Friedman later developed a more general framework for boosting [63].

There are several advantages for tree-based methods. Firstly, decision tree mimics human decision processes and is relatively easy to interpret. For ensemble models, feature importance may be assessed by various means, for example improvement

in the criterion for split (e.g., Gini index) and permutation importance in random forest, and the number of times a feature is used or total gain of splits using the particular feature in boosted trees. Also, tree-based models can handle qualitative and quantitative features and response with ease. In linear models, dummy variables are needed to handle qualitative features, but tree-based methods can absorb qualitative variable directly. Tree-based methods are also robust to outliers and model complex nonlinear relationships well.

Support Vector Machine

Support vector machine (SVM) is a typical maximum margin classifier that aims to separate different classes with a large "gap" [41]. By using the "kernel trick", SVM can map feature space from low dimensions to high, even infinite, dimensions, which makes problems that cannot be solved in low dimensions solvable.

Here we will discuss SVM for classification only. We first assume that the data (X,Y) is linearly separable and $Y \in \{-1,1\}^n$. Intuitively, we can model this problem as follows:

$$\min_{\gamma,w,b} \frac{1}{2} w^T w, s.t. y_i(w^T x_i + b) > 1, i = 1, \cdots, m \tag{1.11}$$

Here w denotes coefficients, b stands for the intercept, and $s.t.$ in the equation is abbreviation of "subject to". This is a typical convex problem with linear restrictions, and it can be solved using convex optimization techniques. In reality, linear separable data is very rare, and SVM can also adapt to inseparable cases with nonlinear decision boundary. The reformulated equation is as follows:

$$\begin{aligned} &\min_{\gamma,w,b} \frac{1}{2} w^T w + C \sum_{i=1}^{m} \xi_i \\ &s.t. y_i(w^T \phi(x_i) + b) > 1 - \xi_i, \xi_i > 0, i = 1, \cdots, m \end{aligned} \tag{1.12}$$

Here, $\phi(xi)$ maps features x_i from low to higher dimensions to capture nonlinear relationships; ξ_i are slack variables that allows some observations to be on the wrong

side; and C controls the penalty of relaxing the functional margin. Figure 1.2 shows a hypothetical classification problem in which two observations fall into wrong sides after the introduction of slack variables.

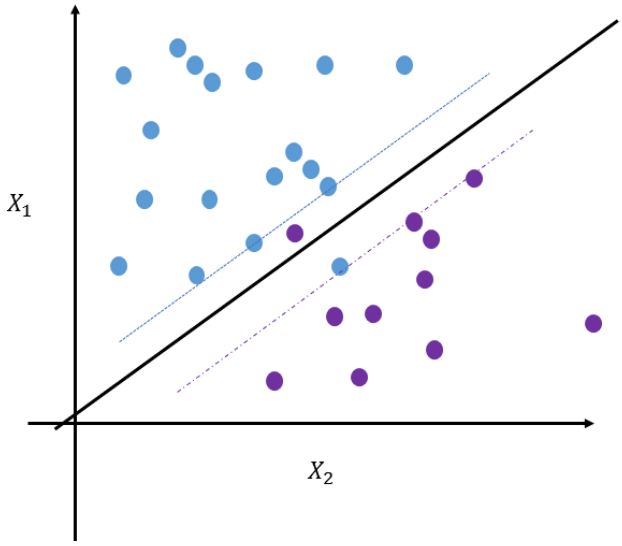

Figure 1.2: A hypothetical classification task using linear SVM. Two observations fall into the wrong sides after the introduction of slack variables

The form of decision boundary can be transformed into the sum of inner product of feature mapped with the form of $\sum_{i=1}^{m} \alpha_i y_i \langle \phi(x_i), \phi(x) \rangle$, and $\langle \phi(x_i), \phi(x) \rangle$ is a kernel that measures the similarity between xi and x. The Gaussian (or radial basis function, RBF) kernel is one of the most widely used kernels to produce complex nonlinear decision boundaries. The Gaussian kernel can be expressed as:

$$K(x,z) = \exp\left(-\frac{||x-z||^2}{2\sigma^2}\right) \tag{1.13}$$

There are serval key characteristics of SVM. First, the decision boundary of SVM is actually determined by observations near the boundary, and thus data points far away from the boundary have little effect on the decision boundary. It models nonlinear relationships well, and various kernels can be applied to make different complex

decision boundaries to satisfy different classification problems [26].

SVM has been employed for drug or target discovery in earlier studies. For example, in a recent work, the authors integrated several layers of drug properties including chemical structures and proximity of targets in an interaction network and expression profiles and used SVM to predict therapeutic classes [151]. Another study adopted SVM trained on molecular structure, molecular activity, and phenotype data to discover new indications for drugs [225]. A study we mentioned earlier [1] employed SVM and DNN to learn drug therapeutic categories from gene expression data. Regarding the application of SVM on drug target discovery, studies [15, 125] utilized SVM on structural or chemical properties of known drug targets to identify promising drug targets.

In our recent application, we used SVM with Gaussian kernel and found SVM in general performed favorably compared to other methods [240]. We used the Python package "scikit-learn" [165, 30] for implementation, but similar packages in R or other programming languages are also available.

Regarding the limitations of this approach, SVM models are often difficult to interpret, and there is a lack of widely used criteria to quantify the importance of individual features. In the case of drug repositioning or target discovery, this may be a limitation given that we are usually interested in identifying which genetic or biological factors contribute to the treatment effects on diseases. As a kernel-based method, drug repositioning with SVM employs a comparable principle to other "similarity-based" approaches (e.g., a drug X with high similarity to a known treatment A may also be able to treat the same disease) [94]. Similarly, perturbation of a gene may induce a pattern of expression profile that is similar to that of a drug, then compounds acting on the gene may have similar mechanisms to the drug. Thus, one limitation is that such an approach may not be very good at revealing candidates with novel mechanisms of actions [94].

Deep Neural Networks

Deep learning has attracted increasing attention in recent years and contributed to significant advances in many fields such as computer vision. Deep neural networks (DNN) are based on the concept of "representation learning" [20] and are very good at capturing nonlinear relationships. Many different network architectures have been developed, but here we only discussed feedforward neural networks with fully connected layers.

By using multiple hidden layers, DNN can handle more complex relationships than a simple single-layered network. The optimal number of hidden layers and neurons will depend on the nature and complexity of the problem as well as the data size. DNN usually requires relatively large sample sizes to achieve good predictive power as the number of parameters is large and overfitting can be a major problem. Dropout is a simple and widely used approach to avoid overfitting by "inactivating" a proportion of neurons randomly during training [194]. Feature selection and shrinkage can also be applied by employing L1 and L2 regularization [243]. There are also numerous other hyper-parameters to choose from, such as the activation function, learning rate, momentum, batch size, etc. For activation function of hidden layers, ReLU is often used. In the output layer, sigmoid function can be used in binary classification problems and softmax in multi-classification problems. The performance of DNN is promising in recent studies of drug repositioning and drug category classification [4, 240]. Nevertheless, DNN models are hard to interpret, and the choice of hyper-parameters is often difficult. The computational and time costs for training a model are relatively high (especially for large datasets); however, the use of graphic processing units (GPUs) can greatly accelerate the computing speed.

With the rapid development of deep learning methodologies, they have been increasingly used for drug repurposing [4, 240, 230] or prediction of various drug properties or toxicities. For example, Klambauer et al. applied deep neural networks on chemical features of compounds to predict their toxicities [143]. Ryu et al. employed

deep learning to improve prediction of drug-drug and drug-food interactions [178]. Deep learning has also been used to predict synergistic effects of drugs in cancer therapy [172]. Readers may also refer to recent reviews on the applications of deep learning in biomedicine and drug discovery [17, 34, 36].

Cross Validation to Assess Predictive Performance

Above we have introduced several common ML algorithms for training a prediction model. Meanwhile, assessing the performance of model is a critical issue, and the performance of the model in train set can dramatically underestimate of the true prediction error.

To avoid overoptimistic estimation of model performance, the prediction error can be estimated in a new dataset independent of the training set, if such data is available. However, data is often limited, and a more popular approach is K-fold cross validation. A typical practice is to firstly split the entire dataset into K folds evenly and then set side onefold of data as testing set and train on the other folds in each loop. There is no fixed rule to determine K, but it is often set at 5 or 10. A very low K (e.g., leave-one-out cross validation) will lead to almost unbiased but high variance of the prediction error estimate, as the training sets are highly similar. Increasing K will reduce the variance but may increase the bias [65].

In practice, one often needs to tune hyper-parameters, and dividing the data into training and testing sets will not be sufficient. In some studies, the authors would train the model in the training set and pick the hyper-parameters that give the best predictive performance in test set and then report the corresponding prediction error. (In case of cross validation, the ”best” prediction error may be averaged over the K folds). However, such an approach still tends to give overoptimistic estimates of the prediction error as one is picking the best-performing parameters each time which may not be generalized to a completely new dataset [217]. To avoid this problem, the testing set should not be involved in parameter tuning. For example, the dataset can

be divided into training, validation, and testing sets, in which the hyper-parameters are chosen based on predictive performance in the validation set. A more advanced approach is nested K-fold cross validation [217]. In this case, inner K-fold cross validation is used to choose the best hyper-parameters, and the performance of the model with the best parameters chosen is evaluated on the testing set.

Criteria for Model Selection

Here we describe criteria for assessing model fit and predictive performance. For regression, the most commonly used criteria are mean squared error loss. Below we discuss the metrics for classification.

Log loss, or cross entropy, measures the negative log-transformed probability of belonging to expected class for each observation. Its equation is

$$\text{EPE} = -\sum_i y_i \log p(x_i) + (1 - y_i) \log(1 - p(x_i)). \tag{1.14}$$

Therefore, the higher the probability an observation belongs to the expected class, the smaller the log loss value of the observation. Cross entropy is a widely used objective function in classification tasks.

If there is a predefined specific threshold to define a positive or negative outcome (e.g., say it is generally agreed that predicted probability $> 30\%$ represents positive outcome), different measures such as sensitivity (aka recall), specificity, precision (aka positive predictive value), and F1 score (harmonic mean of precision and sensitivity) can be computed accordingly. However, often we may not have such a predefined threshold, and we may wish to consider the overall performance of the model under a variety of possible thresholds. In this case we may use the area under the receiver operating characteristic (ROC) curve (AUROC) or area under the precision-recall curve (AUPRC) as metrics of predictive performance.

The ROC curve records the true positive rate (sensitivity) against false positive

rate (1- specificity) at different thresholds. Area under the ROC curve (AUROC) is a commonly used metric to assess predictive performance especially in the medical field. For problems with class imbalance, it has been argued that AUPRC may better reflect the model performance [47].

Common Issues of Machine Learning in Biomedical Studies

Overfitting refers to overlearning of a model on the training data, leading to poor performance when applied in a new situation. Underfitting describes an opposite phenomenon, in which the model fails to capture the complex relationships within the data. A closely related concept is the "bias-variance" tradeoff. In general, models that are complex will have small bias but higher variance, while simpler models enjoy lower variance but have increased bias. Several approaches may be employed to reduce the risk of overfitting. For example, one may reduce the number of features by preselection or some form of dimension reduction, apply heavier regularization penalty to make the model simpler, or switch to less complex ML models. If possible, obtaining larger sample sizes will also alleviate the problem. Underfitting can be overcome by opposite strategies.

However, how do we know whether a model overfits or underfits in practice? A typical strategy is to examine or plot a curve of the training and testing errors. If training error is unacceptably high and the gap between the two errors is small, then the complexity of the model chosen may be too low, or underfitting is present. If the training error is close to 0 but testing error is high, the model might be overfitting.

Imbalanced data is a problem often encountered in biomedical applications in which observations with positive outcome may be uncommon. For example, only a few people may develop a disease or complication, or only a minority of the drugs can treat a specific disorder. There are several common strategies for imbalanced data, such as down-sampling the majority class, up-sampling of the minority class, and constructing new cases by methods such as SMOTE [33]. Here we briefly de-

scribe how to tackle this problem with class weights. If the default weight for each observation is 1 and positive observations are rare, the total weights of positive and negative observations will be imbalanced. To remedy the situation, we can *increase* the weight for each positive observation to balance the total weights of the positive and negative classes. This strategy can also be used in multi-class classification problems. In a recent work of drug repositioning, we did observe obvious improvement in predictive power using the above weighting scheme [240].

1.4 Machine Learning for Individualized Treatment Effects Estimation

Current machine learning (ML) mostly focus on prediction of outcome at the population level. Nevertheless, a more clinically relevant question would be: how could the treatment or risk factor affect someone like me? The same treatment or risk factor can have different effects on different individuals, due to their varying background (e.g. age, sex, comorbid diseases, socioeconomic status etc.). This is a particularly relevant question in the era of precision medicine.

To address this issue, there is a rise in the number of studies using machine learning approach to estimate individualized treatment effect (ITE). Typically, there are two typical steps in the estimation of ITE. We need to identify groups of heterogeneity in treatment effects first, then estimate of individualized treatment effects [180]. Currently. there are mainly two streams in the application of ML approaches to ITE estimation: average treatment effect of subgroups defined by learning algorithms and ensemble approaches based on tree-based methods.

Some recent studies focused on detecting subgroups demonstrating heterogeneity in treatment effects. Specifically, these studies try to to estimate the difference of averaged outcome between treatments and controls in pre-specified subgroups [67] or subgroup defined by learning algorithms [199, 198, 11, 61, 128]. Su et. al. employed

interaction trees to iteratively search for subgroups having similar treatment effects [199, 198]. However, the definition of subgroup is often arbitrary and somewhat algorithm dependent, the generalization ability of subgroup algorithms may be poor, and these subgroup algorithms tend to over-estimate the heterogeneity present. As stated in a recent study [220], the reporting of only the subgroups with extreme treatment effects to highlight heterogeneity may lead to optimistic bias. In high-dimensional settings, as is the case for most genomics data, it is still very challenging to divide subjects into appropriate subgroups [171].

Forests-based methods have gained popularity in ITE estimation in the past years [220, 117, 136, 45, 90, 90, 78], as forests-based methods enjoy several advantages. Most importantly, they have relatively good interpretability since variables contributing to prediction can be extracted, and one can even 'visualize' the 'black-box' of decision-making by plotting representative trees. Forest-based methods are also able to capture complex interactions in data.

A recent study by Basu et al. shows that forest-based methods can discover predictive and stable high-order interactions [19]. They also enjoy other benefits, for instance they are like less likely to overfit and are robust to missing values. Several studies [78, 89, 90] also proposed the use of Bayesian additive regression tree (BART) method [37] to model ITE. One advantage of this kind of approach is that reliable intervals for treatment effects can be readily obtained by MCMC sampling. Other methodologies like meta-learners and deep-learning based ITE estimation could be found in [117, 105, 171]. Powers et al. proposed causal boosting, causal Multivariate adaptive regression splines (MARS) and pollinated transformed outcome (PTO) forests to fit adjusted outcomes [171] for estimation of ITE. Another study employed deep learning to make counterfactual inference on ITE [105]. In practice, these approaches try to fit $\mu_1(\mathbf{x})$ and $\mu_0(\mathbf{x})$ simultaneously / separately and estimate the difference by computing the different between $\mu(\mathbf{x}, w)$ and $\mu(\mathbf{x}, 1-w)$ (Notations here are same defined as in page xi).

Similarly, meta-algorithms decompose estimation of conditional average treatment effects (CATE) into several sub-regression tasks, that can be tackled with any supervised learning methods [117]. A limitation of these studies is a lack of formal statistical inference [220] for ITE. Wager et al. proposed causal forests with honesty to estimate ITE [220]. In causal forests with honesty, training set is split into two parts, in which one partition is used for tree growing, and the other partition is utilized to make honest estimation. The authors have showed that causal forests with honesty have good asymptotic properties [220]. Despite these advances, there are still important study gaps. For example, there is a lack of methods for the statistical inference of the presence of heterogeneity and whether the ITE model outperforms a standard model assuming homogeneous effects. Methods for validating ITE models are also seldom explored. The above questions will be addressed in this thesis.

1.5 Summary

In this section, we introduced the current situation of drug development and difficulties in the process of developing a new medication, reviewed some studies using computational approaches to drug target discovery and purposing, and discussed their limitations. As the technique behind the computational method largely belongs to the field of machine learning, we gave a detailed introduction of machine learning approaches, with some biomedical applications of corresponding ML methods as examples. Then we went into another topic of discovering ITE for risk factors or treatments. For estimation of ITE, we reviewed several ML based approaches and pointed out possible limitations of these studies.

□ End of chapter.

Chapter 2

Drug Repositioning for Psychiatric Disorders: A Machine Learning Approach Leveraging Expression Data

2.1 Background

2.1.1 Motivation

Development of new medications is a very lengthy and costly process. While investment in research and development has been increasing, there is a lack of proportional rise in the number of drugs approved in the past two decades, especially for drugs with novel mechanisms of actions [162]. There is an urgent need for innovative approaches to improve the productivity of drug development. This is particularly true for some areas like psychiatry, for which there has been lack of therapeutic advances for some time [114, 99].

Finding new indications for existing drugs, an approach known as drug reposi-

tioning or repurposing, can serve as a useful strategy to shorten the development cycle [50]. Repurposed drugs can be brought to the market in a much shorter time-frame and at lower costs. With the exponential growth of "omics" and other biomedical data in recent years, computational drug repositioning provides a fast, cost-effective and systematic way to identify promising repositioning opportunities [50].

In this study we describe a general drug repositioning approach by predicting drug indications based on their expression profiles, with a focus on applications in psychiatry. We treat drug repositioning as a supervised learning problem and apply different state-of-the-art machine learning methods for prediction. Drugs that are not originally indicated for the disease but have high predicted probabilities serve as good candidates for repositioning. There are several advantages of this approach. Firstly, the presented approach is a general and broad framework that leverages machine learning (ML) methodologies, a field with very rapid advances in the last decade. This provides great flexibility and opportunities for further improvement in the future as virtually any supervised learning methods can be applied. Newly developed prediction algorithms can also be readily incorporated to improve the detection of useful drug candidates. In addition, the method described here is widely applicable to any chemical or drugs with expression profiles recorded, even if the drug targets or mechanisms of actions are unknown. For example, herbal medicine products may contain a mixture of ingredients with uncertain drug targets; even for many known medications (e.g. lithium [140]), their mechanisms of actions and exact targets are not completely known. If transcriptomic profiling has been performed, they can still be analyzed for therapeutic or repositioning potential under the current approach.

2.1.2 Related Works

There has been increasing interest in computational drug repositioning recently, in view of the rising cost of new drug development. Hodos et al. [94] provided a comprehensive and updated review on this topic. Similarity-based methods (e.g. ref. [76,

157, 131, 137, 151, 124]) represent one common approach, but as noted by Hodos et al., the dependence on data in the "nearby pharmacological space" might limit the ability to find medications with novel mechanisms of actions. Another related methodology is the network-based approach [133], which typically requires data on the relationship between drugs, genes and diseases as well as connections within each category (e.g. drug-drug similarities). It can integrate different sources of information but may still be constrained by the focus on a nearby pharmacological space and the choice of tuning parameters in network construction or inference is often ad hoc [94]. The present work is different in that we apply a broad framework for repositioning and we do not focus on one but many different kinds of learning methods. There is comparatively less reliance on known drug mechanisms or the "nearby pharmacological space" as we let the different algorithms "learn" the relationship between drugs, genes and disease in their own ways. We note that kernel-based ML methods such as support vector machine (SVM) are also based on some sort of similarity measures. A related work [151] have also examined SVM as a promising ML approach for drug repositioning and identified several interesting candidates. However, here our focus is different in that we considered a variety of other approaches and SVM is one of the methods which falls under the broader framework of ML for repositioning. We also employed more in-depth validation strategies, such as assessing enrichment for drugs considered in clinical trials and correlations with the level of literature support. As for other advantages of ML approaches, for high-throughput omics data, often only a subset of genes or input features are relevant, and many ML methods are able to "learn" which features to consider for repositioning. As we shall discuss later, ML approaches may also provide a new avenue to explore the mechanisms of different drug classes, by studying the variable importance of gene features.

We are particularly interested in drug repositioning for psychiatric disorders in view of the lack of novel treatments in the area. Although computational drug repositioning has attracted increased attention recently, relatively few studies focus on psy-

chiatric disorders (except e.g. [231, 106, 170, 191]), compared to other areas like oncology. Psychiatric disorders are leading causes of disability worldwide [219], however there have been limited advances in the development of new pharmacological agents in the last two decades or so [99]. Development of new therapies is also limited by the difficulty of animal models to fully mimic human psychiatric conditions [153]. Investment by drug companies has in general been declining [99], and new approaches for drug discoveries are very much needed in this field. We will explore repositioning opportunities for schizophrenia along with depression and anxiety disorders. Here depression and anxiety disorders are analyzed together as they are highly clinically comorbid [121, 110], show significant genetic correlations [160], and share similar pharmacological treatments [16].

2.1.3 Significances

Contributions of this study are summarized below. Firstly, we presented a general workflow and approach to drug repositioning of a disease based on ML methods, leveraging drug expression profiles as predictors. While previous work [4] has also proposed the use of ML on drug transcriptome profiles for classifying drugs into groups (e.g. anti-cancer drugs, cardiovascular drugs, drugs acting on the central nervous system etc.), we focused on drug repositioning for particular diseases instead of predicting the big therapeutic groups, as disorders in the same group can have diverse treatments. Secondly, we have performed a comparison of the predictive performances of five state-of-the-art and perhaps most commonly employed ML algorithms, including deep neural networks, support vector machines, elastic net, random forest and gradient boosted trees. Thirdly, we identified new repositioning opportunities for schizophrenia and depression/anxiety disorders and validated the relevance of the repositioned drugs by showing their enrichment among drugs considered for clinical trials, as well as support by previous literature. As another means of validation, we also showed that the predicted probabilities of treatment potential

are significantly and positively correlated with the level of literature support (using the number of research articles support as proxy). Finally, we explored which genes and pathways contributed the most to our predictions, hence shedding light on the molecular mechanisms underlying the actions of antipsychotics and antidepressants.

2.2 Datasets and Methods

2.2.1 Datasets

We present a general drug repositioning approach adopting a supervised learning approach. We construct prediction models in which the outcome is defined as whether the drug is a known treatment for the disease, and the predictors are expression profiles of each drug. Drugs that are not originally known to treat the disease but have high predicted probabilities are regarded as good candidates for repositioning.

Drug expression data

The expression data is downloaded from Connectivity Map (CMap), which captures transcriptomic changes when three cell lines (HL60, PC3, MCF7) were treated with a drug or chemical [121]. We downloaded raw expression data from Cmap, and performed normalization with the MAS5 algorithm [166]. Expression levels of genes represented on more than one probe sets were averaged. We employed the limma package [175] to perform analyses on differential expression between treated cell lines and controls. Analyses were performed on each combination of drug and cell line, with a total of 3478 instances. Note the expression of each drug was measured on three different cell lines. Expression measurements were available for 12436 genes. Thus, the scale of our dataset is 3478×12436. Statistical analyses were performed in R3.2.1 with the help of the R package ”longevityTools”.

Defining drug indications

Drug indications were extracted from two known resources, namely the Anatomical Therapeutic Chemical (ATC) classification system and the MEDication Indication Resource high precision subset (MEDI-HPS)[115]. We focus on schizophrenia as well as depression and anxiety disorders in this study. From the ATC classification system, two groups of drugs were extracted, consisting antipsychotics and antidepressants. On the other hand, the MEDI-HPS dataset integrates four public medication resources, including RxNorm, Side Effect Resource 2 (SIDER2) [115], Wikipedia and MedlinePlus. We used the MEDI high-precision subset (MEDI-HPS) which only include drug indications found in RxNorm or in at least 2 out of 3 other sources [226]. This subset achieves a precision of up to 92% according to Wei et al. [226]. To our knowledge, antidepressants from ATC and depression / anxiety from MEDI-HPS roughly fall into the same category, and this is also the same for antipsychosis from ATC and schizophrenia from MEDI-HPS.

2.2.2 Methods

We employed different state-of-the-art machine learning approaches including deep neural networks (DNN) [74], support vector machine (SVM) [41], random forest (RF) [28], gradient boosted machine with trees (GBM) [63] and logistic regression with elastic net regularization (EN) [243] to predict indications with binary classifiers. Our data is imbalanced as only a minority of the drugs are indicated for schizophrenia or depression/anxiety disorders. We performed both unweighted and weighted analyses in this study; in the weighted analysis, class weights are adjusted such that the minority group (drugs indicated for the disorder) will receive higher weight to achieve a balance between positive and negative instances.

In our unweighted model, DNN was implemented in the python package keras. We employed a fully connected feedforward neural network. Hyperparameters were chosen by the ”fmin” optimization algorithm from ”hyperopt”, which employs a se-

quential model-based optimization approach [23]. The tree-structured Parzen estimator (TPE) was used. The more sophisticated hyper-parameter search strategies provided by sequential model-based methods may produce better results than simpler approaches (e.g. grid search) when the number of hyper-parameters is large, such as in deep learning settings [23]. Fifty evaluations were run for each search of optimal hyper-parameters. Dropout and mixed $L1/L2$ penalties were employed to reduce over-fitting. The neural networks consisted of two or three layers, with number of nodes selected uniformly from the range [64, 1024]. Dropout percentage was selected uniformly from [0.25, 0.75], and $L1/L2$ penalty uniformly from [1E-5, 1E-3]. Optimizer was chosen from "adadelta" [238], "adam" [111] and "rmsprop" [92], and the activation function chosen from "relu", "softplus" or "tanh". One hundred epochs were run for each model and we extracted the model weights corresponding to the best epoch.

We also attempted "hyperopt" for the weighted analysis, however the predictive performance was unexpectedly poor due to unclear reasons yet to be revealed. We therefore turned to hyperparameter selection with grid search, with some adjustments in the parameter ranges. A two-layer neural network was constructed with dropout and mixed $L1/L2$ to avoid over-fitting. The number of neurons in the first hidden layer was selected from $\{1000, 1500, 2000\}$, the number in the second layer from $\{500, 1000, 1500\}$, dropout rate from $\{0.4, 0.5, 0.6, 0.7, 0.8\}$, $L1/L2$ penalty from intervals [-13, -3] and [-9, -8] in log space and the number of epochs from [10, 20, 30, 50]. In order to speed up hyperparameter selection, we first chose the number of epochs, the best optimizer and activation function (following the same parameter range as described above) with other parameters fixed, and then used the best parameters chosen in the first step to find the optimal complexity of our neural networks by selecting the number of neurons in each layer, dropout and mixed $L1/L2$ penalties.

SVM, RF and GBM models were implemented in "scikit-learn" (sklearn) in python. Hyper-parameter selection was performed by the built-in function GridSearchCV

in sklearn. For SVM, we chose radial basis function as the kernel. The two hyper-parameters C and γ were selected from [-5, 2] and [-6, 2] in log10-space respectively.

For RF, the number of bagged trees was set to 1000, the maximum number of features used for splitting was selected from {800, 1000, 1500, 2000, 3000, 5000 } and min_samples_leaf (the minimum number of samples required at a leaf node) was selected from {1, 3, 5, 10, 30, 50, 80 }. For GBM, the number of boosting iterations was selected from a sequence of 100 to 1001 with step size 50, learning rate from {0.005, 0.01, 0.015, 0.02, 0.03, 0.05 }, subsampling proportion from {0.8, 1 }, maximum depth of each estimator from {2, 3, 5, 10} and maximum number of features from {10, 30, 50, 100, 500, 1000 }. Finally, the EN model was implemented by the R package "glmnet" [64]. The elastic-net penalty parameter α was chosen from seq(0, 1, by=0.1), with other settings following the default.

Nested Cross-validation

We adopted nested three-fold cross validation (CV) to choose hyper-parameters and evaluate model performances. It has been observed that optimistic bias will result if one uses simple CV to compute an error estimate for a prediction algorithm that itself is tuned using CV [217]. Nested CV avoids this problem and is able to give an almost unbiased estimate of prediction accuracy [217]. The inner loop CVs were used to choose the parameters that optimized predictive performance. In each outer loop CV we made predictions on the corresponding test set using the best model trained from the inner CV loops. To achieve maximum consistency in our comparisons, we compared different methods on the same test set in each loop. Note that the test sets were not involved in model training or parameter tuning.

Predictive performance measures

The performances of the machine learning methods were evaluated in the test sets using three metrics, including log loss, area under the receiver operating character-

istic curve (ROC-AUC) and area under the precision recall curve (PR-AUC). Log loss compares the predicted probabilities against the true labels. The receiver operating characteristic curve, which plots the sensitivity (i.e. recall) against (1- specificity), is a very widely used approach to evaluate predictive performances in biomedical applications. The precision-recall curve on the hand plots the precision (i.e. positive predictive value) against the sensitivity (recall). Since precision depends on the overall proportion of positive labels, the PR-AUC is also dependent on such proportions. Davies et al. [47] that the PR curve may give more informative comparisons when working with imbalanced data.

Identifying important genes and pathways

We also performed analyses to reveal the genes which contribute the most to the prediction model. For elastic net, we extracted the genes with non-zero coefficient in at least one cross-validation fold, and the resulting genes were subject to an over-representation analysis (ORA) (using hypergeometric tests) to reveal the pathways involved. For RF and GBM, feature importance was computed using built-in functions in sklearn based on Gini importance (i.e. the average decrease in node impurity). We then performed a gene-set enrichment analysis (GSEA [201]) based on the genes together with their respective importance scores (the highest score across three folds was taken). For SVM and DNN, there is a lack of widely adopted importance measures, so we focused on the rest of ML methods in this part. Pathway analyses were conducted by the web-based program WebGestalt [223]. Four pathway databases were considered in our analyses, including KEGG, PANTHER, Reactome and Wikipathways.

External validation by testing for enrichment of psychiatric drugs considered for clinical trials

We then performed additional analyses to assess if the drugs with high predicted probabilities from our machine learning models are indeed good candidates for repositioning. Briefly, we tested whether the drugs with no known indication for the disease but high predicted probabilities are more likely to be included in clinical trials.

In the first step, we filtered off drugs that are known to be indicated for the disease as derived from ATC and MEDI-HPS. This is because we are mainly interested in repositioning other drugs of unknown therapeutic potential, and that the labels of drug indication (from ATC or MEDI-HPS) have already been utilized in the ML prediction steps. Including known indications will lead to over-optimistic estimates of significance of enrichment. Next, we extracted a list of drugs that were included in clinical trials for schizophrenia as well as depression and anxiety disorders. The list was derived from clincialTrial.gov and we downloaded a pre-compiled version from.

We then tested for enrichment of those drugs listed in clinicalTrial.gov among the top repositioning results. We performed an enrichment analysis of "drug-sets", similar to a gene-set analysis approach widely used in bioinformatics [106]. We performed one-tailed t-tests to assess if the predicted probabilities (derived from machine learning models) are significantly higher for psychiatric drugs considered in clinical trials.

Searching for literature support

We manually search for literature support for the top 15 repositioning hits for each method in PubMed and Google scholar. The search strategy is given in Supplementary text. A limitation of manual search is that it is extremely time-consuming to perform such a search for all drugs. It should be noted that publication bias is likely to be present (as negative studies are less likely to be reported), although it is difficult to

exactly quantify the extent of bias. As we shall discuss later, we did find literature support for a number of top repositioning candidates, but it is still possible that similar evidence may be found for drugs ranked in the middle or lower. Also, one may wish to assess whether drugs with less (or no) support by prior studies would indeed have lower predicted probabilities (similar to having 'negative controls' in an experiment).

To ensure an unbiased and more comprehensive comparison, we conducted an analysis with "automated" literature mining on all drugs in PubMed. We extracted the number of research articles supporting each drug's association with the corresponding psychiatric disorder (schizophrenia or depression/anxiety). We then examined the correlation between the number of research articles support and the predicted probabilities of treatment potential from ML models. Similar approaches for validating repositioning candidates based on literature mining have also been used in other studies [98]. As the number of articles is typically skewed and not normally distributed, we employed Spearman and Kendall correlation measures. We also compared the predicted probabilities of drugs with no article support versus those with at least one article support. We hypothesized that drugs without any literature support would have lower predicted probabilities from ML models, and vice versa. We used the Wilcoxon rank-sum test for such comparison. All tests were carried out in R 3.2.3 and the tests were one-tailed. Similar to the enrichment test discussed above, drugs that are known to be indicated for the disorder (from ATC or MEDI-HPS) were excluded from this analysis. This is because the indication label of these drugs have been used in the ML model, which may lead to over-optimistic results, and that we wish to focus on repositioning potential of other drugs of unknown significance.

Table 2.1: Average predictive performance of different ML models across four datasets in unweighted (top) and weighted analysis (bottom)

Unweighted analysis	Average Log Loss			
	MEDI-HPS DEP/ANX	ATC ATD	MEDI-HPS SCZ	ATC ATP
SVM	**0.1188**	0.0943	**0.1018**	**0.0895**
EN	0.1249	**0.0916**	0.1097	0.0954
DNN	0.124	0.0948	0.1111	0.0992
GBM	0.1293	0.1018	0.1157	0.1039
RF	0.1294	0.1002	0.1155	0.1013
	Average ROC-AUC			
SVM	**0.7141**	0.7619	**0.7705**	**0.7755**
EN	0.725	**0.779**	0.7515	0.7681
DNN	0.6952	0.7456	0.7533	0.7604
GBM	0.6536	0.6042	0.7172	0.7433
RF	0.6315	0.639	0.7036	0.7501
	Average PR-AUC			
SVM	**0.2026**	**0.1485**	**0.2973**	**0.3379**
EN	0.1372	0.1008	0.1586	0.2087
DNN	0.1447	0.0877	0.1577	0.2156
GBM	0.091	0.0417	0.1426	0.1528
RF	0.1193	0.0639	0.1677	0.1703
weighted analysis	Average Log Loss			
	MEDI-HPS DEP/ANX	ATC ATD	MEDI-HPS SCZ	ATC ATP
SVM	**0.1189**	**0.0934**	**0.1022**	**0.0898**
EN	0.5803	0.5344	0.5028	0.5112
DNN	0.1309	0.099	0.1308	0.1098
GBM	0.1281	0.1032	0.1114	0.0981
RF	0.1234	0.0943	0.106	0.0939
	Average ROC-AUC			
SVM	0.7198	0.7718	0.7731	0.7765
EN	0.661	0.7394	0.7494	**0.7997**
DNN	**0.7424**	**0.7979**	0.741	0.7576
GBM	0.7155	0.7578	0.7584	0.7794
RF	0.689	0.7355	**0.7843**	0.7801
	Average PR-AUC			
SVM	**0.2017**	**0.151**	**0.298**	**0.3361**
EN	0.0751	0.0896	0.152	0.203
DNN	0.1796	0.1107	0.2278	0.2641
GBM	0.18	0.1168	0.2697	0.278
RF	0.1771	0.1165	0.2721	0.2707

1. Values of evaluation metrics for algorithms with the best performance in each dataset (for each evaluation metrics) are marked in bold.
2. SVM: support vector machines; EN: logistic regression with elastic net regularization; DNN: deep neural networks; RF: random forest; GBM, gradient boosted machines.
3. MEDI-HPS: MEDication Indication - High Precision Subset; ATC: Anatomical Therapeutic Chemical classification.

2.3 Experiment Results

2.3.1 Predictive performance comparison

Unweighted analysis

The average predictive performances (averaged over three folds) of different machine learning methods are listed in Table 2.3.1. When considering log loss as the criteria of interest, SVM gave the best result overall, though EN showed the best performance in one of the four datasets. DNN and EN showed quite similar predictive performances. RF and GBM were slightly worse than other methods, but the difference was small. When ROC-AUC was considered as the performance metric, SVM and EN gave similar performances. SVM outperformed EN in the schizophrenia datasets, while EN showed better results in the other two datasets. The performance of DNN was worse than that of SVM and EN, although the differences were not large. The two tree-based methods performed worse especially in the depression/anxiety datasets. We then considered PR-AUC, which is more sensitive to imbalanced data, as the measure of predictive performance. SVM was the best-performing method. EN and DNN followed with very similar performances. Consistent with other performance measures, GBM and RF did not perform as well in the depression/anxiety datasets, but the performance was more comparable for the schizophrenia datasets.

Weighted analysis

Compared with unweighted analysis, we observed improvements in predictive performance for several methods including GBM, RF and deep learning. SVM and EN performed similarly in general. Considering ROC-AUC, deep learning performed the best for depression and anxiety disorders, while RF and EN showed highest ROC-AUC for schizophrenia. SVM achieved the best PR-AUC and log-loss compared to other ML approaches.

Table 2.2: Enrichment for psychiatric drugs included in clinical trials among the repositioning hits

	Dataset	unweighted analysis	weighted analysis
		P-value	P-value
SVM	MEDI-HPS DEP/ANX	**0.0014**	**0.0011**
	ATC ATD	**0.018**	**0.0039**
	MEDI-HPS SCZ	**0.0205**	**0.0264**
	ATC ATP	**0.0098**	**0.0084**
EN	MEDI-HPS DEP/ANX	**0.0022**	**0.0023**
	ATC ATD	**0.0032**	**0.0087**
	MEDI-HPS SCZ	**0.0294**	**0.0315**
	ATC ATP	**0.0104**	**0.0033**
DNN	MEDI-HPS DEP/ANX	**0.0105**	**0.0009**
	ATC ATD	0.1369	**0.0017**
	MEDI-HPS SCZ	0.0908	**0.019**
	ATC ATP	**0.0237**	**0.0021**
GBM	MEDI-HPS DEP/ANX	**0.0494**	**0.0003**
	ATC ATD	**0.0433**	**0.0002**
	MEDI-HPS SCZ	0.2283	**0.0269**
	ATC ATP	0.2482	**0.0005**
RF	MEDI-HPS DEP/ANX	0.0651	**0.0005**
	ATC ATD	0.2518	**0.0007**
	MEDI-HPS SCZ	0.1299	**0.0427**
	ATC ATP	0.5232	**0.0063**

P-values < 0.05 and with FDR less than 0.05 are in bold.

2.3.2 Enrichment for psychiatric drugs considered clinical trials

We further tested whether the top repositioning results are enriched for drugs included in clinical trials for psychiatric disorders. As shown in Table 2.2, we observed significant enrichment of such drugs for both schizophrenia and depression/anxiety disorders across all methods in the weighted analysis. In addition, most results in the unweighted analysis were also significant. This external validation provides further support to the usefulness of our approach in identifying new repositioning opportunities.

2.3.3 Correlation of predicted probabilities with degree of literature support

We also examined Spearman and Kendall correlations between predicted probabilities from ML models and the number of PubMed articles retrieved, which serves as a proxy for the level of literature support 2.3. We focused on results from the weighted analysis as they have better predictive performances in general. We found significant and positive correlations for all ML methods across all four analyses. DNN was the best performing method (in terms of the correlation metric and level of significance) in two out of four tasks (ATC antipsychotics and ATC antidepressants), and was relatively close to the best ones for the other two analyses. SVM performed the best in these analyses, but its deficit when compared to DNN was proportionately large for the two ATC tasks. The results of Wilcoxon rank-sum test were generally concordant with those from correlation tests.

2.3.4 Identifying contributing genes and pathways

Supplementary Tables [1] 1-4 show the top genes as identified by variable importance measures (for RF and GBM) and regression coefficients (for EN). The enriched pathways are shown in Table IV and Supplementary tables 5-12. Since the number of

[1] Available at https://drive.google.com/open?id=1YDtk-uTVX5gsnZvWM7q3PRz3_x8yWrpQ

Table 2.3: Correlations between predicted probability of treatment potential with number of research articles supporting association with schizophrenia or depression/anxiety

	MEDI-HPS DEP/ANX	ATC ATD	MEDI-HPS SCZ	ATC ATP
		Spearman's rho		
SVM	**0.078**	0.049	**0.098**	0.088
EN	0.031	0.043	0.091	0.073
DNN	0.075	**0.065**	0.085	**0.11**
GBM	0.065	0.05	0.082	0.102
RF	0.066	0.051	0.052	0.055
		Kendall's tau		
SVM	**0.06**	0.037	**0.077**	0.068
EN	0.024	0.033	0.071	0.058
DNN	0.058	**0.05**	0.067	**0.086**
GBM	0.05	0.038	0.064	0.08
RF	0.052	0.04	0.041	0.044
		Spearman correlation p-value		
SVM	**2.09E-05**	4.91E-03	**6.46E-09**	1.47E-07
EN	4.89E-02	4.91E-03	5.96E-08	9.16E-06
DNN	3.45E-05	**2.64E-04**	3.37E-07	**7.53E-11**
GBM	2.91E-04	3.87E-03	9.64E-07	1.52E-09
RF	2.35E-04	3.51E-03	1.35E-03	6.86E-04
		Kendall correlation p-value		
SVM	**2.27E-05**	5.46E-03	**6.81E-09**	1.81E-07
EN	5.00E-02	1.10E-02	5.81E-08	9.05E-06
DNN	3.56E-05	**2.83E-04**	3.33E-07	**8.52E-11**
GBM	2.90E-04	4.22E-03	1.00E-06	1.62E-09
RF	2.33E-04	3.63E-03	1.37E-03	6.93E-04
		Wilcoxon rank-sum test p-value		
SVM	**1.94E-04**	2.10E-02	**1.49E-07**	3.27E-06
EN	1.09E-01	2.50E-02	3.07E-07	3.40E-05
DNN	3.85E-04	**1.30E-03**	8.63E-07	**8.76E-10**
GBM	1.43E-03	2.11E-02	1.61E-05	1.91E-08
RF	1.00E-03	1.08E-02	4.49E-03	2.98E-03

The highest correlation coefficient or lowest p-value in each analysis is marked in bold.

Table 2.4: Selected enriched pathways based on variable importance of genes in ML models with FDR < 0.2

Method	Name	#Gene	FDR
ATC antidepressants and MEDI-HPS depression/anxiety (weighted analysis)			
eNet-ORA_Reactome	Cholesterol biosynthesis	7	1.78E-06
eNet-ORA_Wikipathway	Sterol Regulatory Element-Binding Proteins (SREBP) signalling	8	1.73E-04
eNet-ORA_KEGG	Steroid biosynthesis - Homo sapiens (human)	5	2.38E-04
gbm_KEGG	Fat digestion and absorption - Homo sapiens (human)	34	8.95E-03
gbm_Panther	Insulin/IGF pathway-protein kinase B signaling cascade	34	1.05E-02
rf_Wikipathway	Mismatch repair	9	1.38E-01
rf_Wikipathway	ID signaling pathway	16	1.58E-01
rf_Wikipathway	Statin Pathway	25	1.61E-01
rf_Wikipathway	Photodynamic therapy-induced HIF-1 survival signaling	37	1.63E-01
eNet-ORA_Panther	TGF-beta signaling pathway	4	1.70E-01
ATC antipsychotics and MEDI-HPS schizophrenia (weighted analysis)			
rf_Wikipathway	Sterol Regulatory Element-Binding Proteins (SREBP) signalling	64	0.00E+00
eNet-ORA_Reactome	Cholesterol biosynthesis	7	9.33E-05
eNet-ORA_KEGG	Steroid biosynthesis - Homo sapiens (human)	4	2.01E-03
rf_Wikipathway	Statin Pathway	25	2.51E-02
eNet-ORA_Wikipathway	Fatty Acid Beta Oxidation	5	2.98E-02
eNet-ORA_Reactome	Asparagine N-linked glycosylation	15	4.36E-02
eNet-ORA_KEGG	Metabolic pathways - Homo sapiens (human)	37	4.87E-02
eNet-ORA_Reactome	Synthesis of UDP-N-acetyl-glucosamine	3	8.11E-02
gbm_Panther	5HT3 type receptor mediated signaling pathway	14	8.58E-02
eNet-ORA_KEGG	Citrate cycle (TCA cycle) - Homo sapiens (human)	3	8.96E-02
gbm_Reactome	G1/S-Specific Transcription	18	1.32E-01
eNet-ORA_Reactome	Antigen Presentation: Folding, assembly and peptide loading of class I MHC	4	1.32E-01
gbm_Panther	Androgen/estrogene/progesterone biosynthesis	9	1.40E-01
rf_Wikipathway	Photodynamic therapy-induced unfolded protein response	23	1.42E-01
eNet-ORA_Reactome	COPII (Coat Protein 2) Mediated Vesicle Transport	6	1.44E-01

We aggregated pathway analysis results from 4 databases (KEGG, Reactome, Panther and Wikipathways). Pathways that were highly similar were filtered. Only results from weighted analysis are included here.

genes involved is large, we only highlighted a few top enriched pathways here. Interestingly, steroid and cholesterol biosynthesis are among the most significantly enriched pathways for drugs against schizophrenia and depression/anxiety. Notably, abnormalities in the hypothalamic-pituitary-adrenal (HPA) axis have long been suggested as one of the key pathological mechanisms underlying depression [216]. The steroid (cortisol) synthesis inhibitor metyrapone has been shown to be effective for depression in a double-blind randomized controlled trial (RCT) [102] and other studies [139], although another trial failed to show any benefits [144]. Antidepressants have also been shown to regulate glucocorticoid receptor functioning in vivo. On the other hand, neuroactive steroids may be implicated in the pathophysiology of schizophrenia [186]. Cholesterol biosynthesis, including regulation by sterol regulatory element-binding protein (SREBP), was frequently top-listed in our pathway analysis. Antipsychotics and some antidepressants are associated with metabolic syndrome and weight gain, and previous in vitro and in vivo studies have shown lipogenic effects of these drugs as controlled by SREBP transcription factors [174, 57]. Interestingly, some studies showed lower cholesterol may be associated with suicidality [77], depressive symptoms [167, 236], and poorer cognition in schizophrenia [113], but these findings are controversial. Whether pathways related to cholesterol synthesis may play a role in the therapeutic effects of psychotropic drugs remain a topic for further investigation. Some other pathways are also worth mentioning. For example, IGF signaling pathway was significantly enriched under antidepressants. IGF-I has been reported to improve depression and anxiety symptoms in clinical samples [205], and showed antidepressant-like effects in animal models [31, 80]. The 5-HT3 signaling pathway was also top-listed under antipsychotics. 5-HT3 has been proposed as a new drug target and improvements in negative and cognitive symptoms have been reported in clinical trials [52].

2.3.5 Top repositioning hits and literature support from previous studies

Table 2.5 show some of the selected top repositioning candidates with literature support, which will also be discussed below. More detailed tables showing the top 100 hits for each ML method in both unweighted and weighted analyses are presented in Supplementary Tables 13-16 [2]. Note that drugs that are known to be indicated for these disorders by ATC or MEDI-HPS were excluded from the lists. We noted overlap in top hits derived from different machine learning methods, but some repositioning candidates are unique to one or few ML approaches. This suggests that employing a diverse set of ML methods may be advantageous in "learning" different potential repositioning candidates. We will chiefly focus on the top 15 hits for each ML method in the exposition below.

Repositioning candidates for depression/anxiety disorders

Regarding depression and anxiety disorders, many of the top results are antipsychotics, such as trifluoperazine, perphenazine, fluphenazine and thioridazine, among others. Antipsychotics have long been used for the treatment for depression [223]. In earlier studies, phenothiazines (a class of antipsychotic to which many of our top hits belong) was observed to produce similar anti-depressive effects as tricyclic antidepressants [176]. Due to the risk of extra-pyramidal side-effects, typical antipsychotics are less commonly used these days and second-generation (atypical) antipsychotics are more often prescribed. Meta-analyses have shown that atypical antipsychotics are effective as adjunctive or primary treatment for depression [176, 193]. Antipsychotics are also commonly prescribed for severe depressive episodes with psychotic symptoms.

A few other drugs on the lists are also worth mentioning. Cyproheptadine (top-listed by SVM, RF, GBM and EN) is a 5-HT2 receptor antagonist and was shown to improve depression in a small cross-over trial [78]. It was also reported that the drug

[2] Available at https://drive.google.com/open?id=1YDtk-uTVX5gsnZvWM7q3PRz3_x8yWrpQ

Table 2.5: Some literature-supported candidates selected from top hits derived from machine learning methods (known antipsychotics and antidepressants are not included in this list)

Drug	Method	Relationship with disease
Depression/anxiety		
Cyproheptadine	SVM, RF, GBM, EN	5-HT2 receptor antagonist, improve depression in a small cross-over trial
Chlorcyclizine	DNN, RF, GBM, SVM	phenylpiperazine group to which many other antidepressants and antipsychotics belong
Pizotifen	EN	5-HT2A/2C antagonist, positive result in an RCT
TrichostatinA/ Vorinostat	DNN, EN	HDAC inhibitors may have antidepressant effects as shown in animal models
Tetrandrine	DNN, RF, GBM	CCB; antidepressant-like effects in mice; may increase 5-HT, NE and BDNF concentrations
Apigenin	GBM	Antidepressant and anxiolytic effects in animal models and in an RCT
Metformin	EN	may reduce depression risk among DM subjects
Schizophrenia		
Valproate	DNN, SVM	open RCTs reported symptom improvement when used as adjunctive treatment
Raloxifene	DNN, EN	improve SCZ symptoms in an RCT of post-menopausal women
Nordihydroguaiaretic acid	DNN, GBM, SVM	antioxidant; oxidative stress implicated in SCZ
Pioglitazone	DNN	Another drug in the same class (pioglitazone) improved SCZ symptoms in RCT
Tretinoin	DNN	Retinoid; dysfunction in retinoid signaling may be implicated in SCZ
Felodipine	GBM	CCB; CCB added to antipsychotics may be beneficial
Aspirin	SVM	NSAID; may improve SCZ symptoms as shown in RC
Genistein	GBM	Phytoestrogen; animal model shows possible anti-dopaminergic effects

As a number of top results were known antipsychotics or antidepressants (please refer to the main text for details), these were not presented in the above table. RCT, randomized controlled trial; HDAC, Histone deacetylases; CCB, calcium channel blocker; 5-HT, serotonin; NE, norepinephrine; BDNF, brain-derived neurotrophic factor; NSAID, non-steroidal anti-inflammatory drugs.

reduced the neuropsychiatric side-effects of the antiviral therapy efavirenz, including depressive and anxiety symptoms [44]. Chlorcyclizine belongs to the phenylpiperazine class and numerous antidepressants and antipsychotics also belong to this class [139]. Pizotifen, listed by EN, is a 5-HT2A/2C antagonist which was shown to possess antidepressant effects in a double-blind RCT [195]. DNN and EN have identified histone deacetylase (HDAC) inhibitors including trichostatin A and vorinostat as top repositioning hits for depression/anxiety and schizophrenia. HDAC have been implicated in the pathogenesis of psychiatric disorders including depression, as reviewed by Fuchikami et al. [66]. HDAC inhibitors have been reported to produce antidepressant-like effects in animal models [93, 43], although no clinical trials on psychiatric disorders were available. Interestingly, in a recent study which employed gene-set analysis on de novo mutations to uncover repositioning opportunities, HDAC inhibitors were highlighted as candidates for schizophrenia and other neurodevelopmental disorders [190].

Another candidate was tetrandrine, a calcium channel blocker top-listed by DNN, RF, GBM and SVM. Tetrandrine demonstrated antidepressant-like effects in mice [70] in forced swimming and tail suspension tests. The drug also increased the concentration of 5-hydroytrytamine (5-HT) and norepinephrine in mice treated with reserpine or chromic mild stress, and raised the levels of brain-derived neurotrophic factor (BDNF) in the latter case [70].

Repositioning candidates for schizophrenia

With regards to repositioning results for schizophrenia, some of the hits are antidepressants, such as protriptyline, maprotiline and clomipramine, among others. Antidepressants are frequently prescribed in schizophrenia patients due to possibility of comorbid depression or obsessive-compulsive disorder [141]. In meta-analyses antidepressants were also found to improve negative symptoms of schizophrenia [187, 183]. For the antidepressants on the list, maprotiline (listed by RF, GBM, EN, SVM)

has been reported to improve negative symptoms in chronic schizophrenia patients [233] as an adjunctive treatment. Other drug clomipramine (listed by DNN, GBM, EN) has been shown to ameliorate not only obsessive-compulsive but also overall schizophrenic symptoms in patients with comorbid disorders [24]. Interestingly, the mood stabilizer valproate was also listed among the top (by DNN and SVM). Valproate may improve clinical response when added to antipsychotics, although the evidence is mainly based on open RCTs [181]. The EN algorithm also "re-discovered" spiperone, an antipsychotic not listed in ATC or MEDI-HPS, as one of the top repositioning hits.

Several other drugs less well-known for psychiatric disorders are also worth mentioning. The selective estrogen receptor modulator raloxifene (listed by DNN and EN) was shown to improve schizophrenia symptom scores in double-blind RCTs of postmenopausal women [211, 210]. Another drug nordihydroguaiaretic acid (listed by DNN, GBM, SVM) has antioxidant properties [135] and may be useful in combating oxidative stress in schizophrenia [228]. Pioglitazone, top-ranked by DNN, belongs to the class of thiazolidinediones and has anti-diabetic and anti-inflammatory properties. Although this drug was withdrawn due to unexpected adverse effects on the liver, our finding suggested that other thiazolidinediones may be useful for schizophrenia. Indeed, another drug of the same class known as pioglitazone has been shown to improve negative symptoms in schizophrenia patients in a double-blind RCT [101]. Another RCT also showed improvements in depressive symptoms [189]. Tretinoin (listed by DNN) is a retinoid and retinoid dysfunction has been linked to schizophrenia [75, 123] Clinical trials with another retinoid (bexarotene) showed some benefits of the drug as an add-on agent in schizophrenia. Again retinoid signaling was implicated for schizophrenia in a recent study on drug repositioning leveraging de novo mutations [190]. Felodipine (listed by GBM) is a calcium channel blocker and GWAS on schizophrenia and bipolar disorder have revealed many genes related to calcium channels [88, 39]; a recent study also suggested concomitant use of CCB and antipsy-

chotics may be more beneficial than antipsychotics alone [213].

Some hits from the unweighted analysis

The top repositioning candidates from unweighted analysis for each ML method are listed in Supplementary Tables 13-16. There were a number of overlaps with the candidates from the weighted analysis. Here we highlight a few prioritized drugs (that have not been mentioned above) with literature support. Aspirin (acetylsalicylic acid) is a non-steroidal anti-inflammatory agent (listed by SVM), which has been shown to improve schizophrenia symptoms in a recent meta-analysis of RCTs [192]. Genistein is a phytoestrogen and can bind to estrogen receptors [224]. An animal study showed that genistein may possess anti-dopaminergic actions [202]; interestingly, clinical studies have shown potential therapeutic benefits of estrogens on schizophrenia [192].

The EN algorithm identified metformin as one of the top repositioning hits for depression/anxiety disorders. A study in Taiwan reported that the risk of depression in diabetic patients was reduced by 60% for those given metformin with sulfonylurea [221]. Another study reported improved depressive symptoms and cognitive functions for patients with comorbid diabetes and depression [82]. Another drug apigenin, top-listed by GBM, was supported by a number of in vitro and animal studies for possible antidepressant-like and anxiolytic effects [164]. A clinical trial of oral chamomile (which was standardized to contain 1.2% of apigenin) showed benefits for anxiety and depression [6, 7].

2.4 Discussion

In this study, we have presented a repositioning approach by predicting drug indications based on expression profiles. We employed and compared five state-of-the-art ML methods to perform predictions. We also observed that the top repositioning hits

are enriched for psychiatric drugs considered for clinical trials and that many hits are backed up by evidence form animal or clinical studies, supporting the validity of our approach.

Concerning the performance of different machine learning classifiers, we have employed five methods in total, and all but one (EN) are non-linear classifiers. SVM is a kernel-based learning approach that is widely used in bioinformatics. On the other hand, deep learning methods (such as DNN) that are based on the principles of representation learning [20] have witnessed rapid advances in the last decade, especially in the field of computer vision. While potentially powerful, current successful applications typically require very large datasets for training, and we suspect that the relatively modest sample size (N = 3478) of our dataset may have limited DNN to achieve the optimal predictive ability. We have used at most two hidden layers in view of the moderate sample size, and the complexity of the network may be increased with larger samples, although larger samples would lead to greater computational costs. This study shows that deep learning can achieve reasonable performance in drug repurposing, and indeed DNN achieved the best ROC-AUC for depression/anxiety disorders in the weighted analysis. Given the rapid growth in the area, deep learning approaches might be worthy of further investigations. While logistic regression with EN is a linear classifier, it performed well overall though lagging behind SVM. The performances of the two tree-based methods (RF and GBM) were largely comparable with other methods in the weighted analysis, although they were less satisfactory without weighting. Notwithstanding the differences in predictive performances, different algorithms are based on diverse model assumptions and principles, and as shown above, methods with slightly lower predictive accuracy may still reveal useful repositioning candidates that are of different mechanisms of actions.

To the best of our knowledge, this is the first study to employ a comprehensive array of machine learning methods on drug expression profiles for drug repositioning of any particular disease; it is also the first application in psychopharmacology.

This is also the first work to suggest an ML approach to explore the molecular mechanisms underlying drug actions. In a related work, Aliper et al. made use of the drug transcriptome to predict large drug classes e.g. drugs for neurological diseases, drugs for cardiovascular diseases, anti-cancer agents etc [4]. Here our focus is different and clinically more relevant in that we directly identify repositioning opportunities for a particular disorder. It should be noted that drugs for the same body system can have diverse (or sometimes even opposite) effects. For example, antipsychotics like haloperidol are used to treat schizophrenia but they also cause Parkinsonism [142]. Statins reduces low-density lipoprotein levels and coronary heart disease risk [38], but can cause weight gain and increased diabetic risk [203]. In addition, we concentrated on the study of psychiatric disorders, which was not explicitly considered in Aliper et al. [4] or other previous works. Interestingly, DNN was reported to be the best performing method in their study. However their study [4] and the present work are not directly comparable as the outcomes studied are different and the evaluation metrics also differ. F1 score was used in Aliper et al. [4] (although the choice of a cut-off probability for classification was not explicitly stated) while we used ROC-AUC, PR-AUC and log loss as performance indicators.

Here we aim to provide a proof-of-concept example showing that the application of machine learning methodologies on drug expression profiles may help to identify candidates for repositioning, particularly for psychiatric disorders. The approach is intuitive and also highly flexible. Nevertheless, there is still room for improvement. Firstly, we only consider drug indications and the drug-induced transcriptomic changes in our prediction model. This makes the method very flexible and widely applicable to any compounds or drugs for which an expression profile is available. The use of drug transcriptome evades the need of specifying targets and knowing the mechanisms of actions, and the approach may even be applicable to a mixture of chemical ingredients as may be the case for herbal medicines. However, it is possible that our methods may be further improved by incorporating other information

such as drug targets and chemical structure, if such information is available. For example, dopaminergic and monoaminergic pathways have known importance in SCZ and depression treatment respectively, and incorporating such information into our ML framework may further improve prediction accuracy.

As for the prediction algorithm, in our dataset the number of positive labels is small. We tackled this problem by adjusting the weighting of positive and negative instances and indeed found improvements for several ML approaches. Other methodologies to account for imbalanced data are also possible [85], and this may be a topic for further explorations. We covered five commonly used algorithms here but this coverage is obviously not complete; further studies may benefit from the use of more advanced or recently developed learning methods. We also notice that there is an ongoing effort to expand the coverage of CMap [200], and that the study with updated full data and documentations have just been released. We are planning to further explore the current framework in the expanded dataset.

It is reassuring to observe that many repositioning hits are supported by previous studies, the predicted probabilities from our model significantly correlate with the degree of literature support, and that the results are enriched for psychiatric drugs considered in clinical trials. However, we stress that further well-designed pre-clinical and clinical studies are necessary before the any results can be brought into clinical practice. The analytic validation we employed in this study aims to provide evidence for the overall validity of the presented repositioning approach. Validation of individual candidates via detailed animal and clinical studies are essential before a drug can be brought into practice; however it should also be noted that detailed experimental/clinical validation of one or two candidates is less suited to provide evidence on whether the approach as a whole works or not, since most drugs cannot be covered and chance findings are possible.

We have also made use of ML methods to explore potentially important genes and pathways that may contribute to treatment effects. Nevertheless, the results again re-

quire further experimental validations. Computational approaches for repositioning and explorations of drug mechanisms, such as ML-based methods, provide a cost-effective and systematic way to assess and prioritize drug candidates. While we believe the current approach can improve the prioritization of drug candidates, not all top-ranked drugs will be effective and we do not expect to uncover all potential new treatments. However, given the rising cost in developing a new drug (up to USD 2.6 billion [108]), if a method can reduce the failure rate by even a tiny margin, it will already result in large savings in absolute terms. Further work might involve combining the current approach with other computational and experimental methods to further improve the accuracy of repositioning. As an example, ketamine is one of the most promising new therapies for depression, but the current method did not reveal any similar drugs (ketamine is not listed in CMap). However, another computational repositioning method which compared drug and disease transcriptomes suggested several NMDA antagonists for depression [108], highlighting the potential of integrating different methods in future studies.

2.5 Conclusion

In this work we have presented and applied a machine learning to drug repositioning for schizophrenia and depression/anxiety disorders. We found the candidates were enriched for psychiatric drugs considered in clinical trials, and that numerous top hits were supported by previous studies; the degree of literature support also showed a significant correlation with predicted treatment probabilities from ML models.

It is widely acknowledged that drug development in psychiatry has become stagnant for some years, and that traditional approaches to drug discovery has not been as successful as anticipated. On the other hand, the past few years have seen an extremely rapid development in ML methods and applications; in this regard, we hope that this study will open a new avenue for drug repositioning/discovery, and stimulate

further research to bridge the gap between ML and biomedical applications especially drug development. The list of repositioning candidates might also serve as a useful resource for researchers and clinicians working on schizophrenia as well as depression and anxiety disorders, which are illnesses very much in need of new therapies.

□ **End of chapter.**

Chapter 3

A Machine Learning Approach to Prioritizing Candidate Drug Targets for Complex Diseases

3.1 Introduction

3.1.1 Motivation

Traditionally, drug discovery involves five steps: target identification, target validation, lead identification, lead optimization and introduction of the new drugs to the public [168]. Nevertheless, the speed of new drug development has been slower than anticipated, despite increasing investment [162]. It is estimated that the cost of developing a new drug is USD 2.6 billion [212]. One of the main reasons for the enormous cost of drug discovery is due to the high failure rate.

Success of drug development largely depends on the validity of targets. However, the majority of drugs fail to complete the development process due to lack of efficacy, and this is often due to the wrong target being pursued [185]. Traditionally, drug targets are often identified from hypothesis-driven preclinical models, yet preclini-

cal models may not always translate well to clinical applications. For some diseases such as psychiatric disorders, current animal or cell models are still far from capturing the complexity of the human disorder [153]. In addition, some have hypothesized the hypothesis-driven nature of many studies may have led to ”filtering” of findings and publication bias, exacerbating the reliability and reproducibility issues of some research findings. On the other hand, the recent decade has witnessed a remarkable growth in “omics” and other forms of big data. As increasing amount of biomedical data has been made available, computational methods can offer a fast, cost-effective and unbiased way to prioritize promising drug targets. Given the limitation of current approaches and the urgent need to develop therapies for diseases, addressing the problem of target identification and drug development from different angles is essential. We believe that computational and experimental approaches can complement each other to improve the efficiency and reliability of drug target finding.

In the study, we present a flexible novel computational target discovery framework, in which various machine learning (ML) methods can be adopted. It is a data-driven approach to prioritize drug targets for specific diseases, and is independent from most other kinds of evidences e.g. animal models, top genes from GWAS or sequencing studies etc., which are listed in OpenTargets [112], one of the largest drug target databases to date. Specifically, we employ ML methods to drug-induced expression profiles with indication as the outcome variable to *learn the pattern of gene expression contributing to treatment potential*; we then applied the fitted models to transcriptome data derived from gene perturbations (i.e. over-expression [OE] or knock-down [KD] of specific genes) to predict ”treatment potential” of OE/KD of specific genes. We could then prioritize drug targets based on the predicted probabilities.

Intuitively, for example, over-expression (OE) of gene X leads to an expression profile similar to that of five other drugs that are known to treat diabetes. Then an agonist targeted at X (or other drugs that activate X and related pathways) may also be useful for treating diabetes. In this case we expect the ML model (trained on drugs

but applied to gene perturbation data) would output a *high* predicted probability for gene *X*, which can be prioritized for further studies. Let us consider an opposite scenario in which over-expression of gene *Y* *increases* the disease risk. In this case we may observe a *lower*-than-expected predicted probability of 'treatment potential' from the ML model. In this case down-regulation of gene Y may be beneficial for treatment.

3.1.2 Related Works

Kandoi et al. has reviewed applications of ML and system biology on the discovery of target proteins [107]. In these applications, different kinds of biological properties have been explored using ML methods to identify druggable targets [15, 56, 126, 116, 125]. A sequence-based prediction method was proposed to identify drug target proteins based on biological features like amino acid composition, and a comprehensive comparison of several machine learning methods was conducted [116]. In another study [15], eight key properties of human drug target were summarized, and support vector machine (SVM) was employed to build a classifier on these properties to predict probabilities of potential targets. In a similar study, the authors extracted physicochemical properties from known drug targets, trained a classifier with these properties, and listed possible drug targets by predicted probabilities from the classifier [125]. Network-based methods also were employed to identify potential drug targets using topological features of human protein–protein interaction network [126]. These studies aimed to discover new targets by making use of structural attributes, but gene-disease association data such as gene expression profiles may also be used to identify target genes [54, 58, 179, 42]. In a recent study, gene-disease association data from Open Targets was explored by employing four different ML methods to find novel targets [58]. Emig et. al. proposed an integrated network-based method to predict drug targets based on disease gene expression profiles and a high-quality interaction network, and some novel drug targets for scleroderma and other types

of cancer were presented [54]. A most recent study constructed pairwise learning and joint learning methods on chemically and genetically perturbed gene expression profiles to predict drug targets[179].

However, our study is different from the previous studies in several aspects. Some of the previous works (e.g.[58, 15, 116, 125]) aimed to predict general therapeutic targets, instead of targets for specific diseases. Some employed network-based methods (e.g. [179, 126, 54]) for target prediction, which is powerful approach. However, they are relatively dependent on similarity between entities, hence less capable of discovering novel drug targets. Here we employed an ML approach to predict potential drug targets. An advantage is that the method is general and highly flexible, and any ML methodologies including newly developed ones may be used. A recent study [179] also employed gene perturbation to predict drug targets. However, the methodologies and aims of our study and [179] are different. The present work proposed the use of ML methods to assess how the expression profiles from gene perturbations are related to those of drugs. [179] mainly employed Pearson correlation and linear models to assess the similarity between transcriptomic changes from gene perturbation and those from drugs. An advantage of our approach is that different kinds of ML methods (e.g. SVM, random forests, boosted trees) may be used, which may accommodate complex non-linear relationships and possibly interactions between features. [179] used transcriptomic data from gene perturbations mainly to predict drug-protein interactions; prediction of *disease-specific* drug targets was performed in another analysis using networks. Here we proposed integrating transcriptomic data with ML approaches in a unified framework to predict drug targets *for specific diseases*. We also note a previous study of ours has employed an ML approach for drug repositioning (comment: cite our study); however here our aim is to uncover new *drug targets*. Drug repositioning may not always be feasible (for example due to side-effects of existing drugs), and revealing new targets remains an important goal in drug development and pharmacological research. Besides, unlike the previous work, we have covered diseases

other than psychiatric disorders.

Validation of drug-disease or drug-target predictions from computational methods has always been a difficult task. As reported by [81], a cross-validation approach may over-estimate predictive accuracy, as the training and testing set may have overlapping drugs. Also, drugs that are highly similar may be split into train and test sets, hence the similarity of training and testing set may be higher than anticipated in real-life scenarios. Some studies evaluate validity of results using performance evaluation metrics (e.g. AUC-ROC) under the framework of cross-validation, which may lead to overoptimistic results. To avoid this problem, we utilized an independent resource to examine whether our approach can 'rediscover' known drug targets for diseases. Briefly, we performed validation of our results by assessing for enrichment of targets listed by Open Targets [112], a platform for systematic drug target identification and prioritization. The platform integrates data from genetics, somatic mutations, expression analysis, drugs, animal models and the literature through robust pipelines and uses an aggregate score to indicate the association of a target with disease [112].

In summary, we first proposed a general framework for identifying drug targets of specific diseases, based on ML using expression profiles. The methodology was applied to a number of diseases including type 2 diabetes mellitus (DM), hypertension (HT), schizophrenia (SCZ), bipolar disorder (BP) and rheumatoid arthritis (RA). We then validated the approach by assessing its ability to 'rediscover' drug targets based on an external established database. We also found that many candidate targets are supported by the literature and are functionally relevant.

3.2 Datasets and Methods

3.2.1 Datasets

The gene expression profiles were downloaded from the website[1], and consensus transcriptional signatures have been computed for the expression profiles from the Library of Integrated Network Based Cellular Signatures (LINCS) L1000 perturbations [200]. The data includes expression profiles induced by drugs and by over-expression (OE) or knockdown (KD) of specific genes.

We kept genes with expression data present in both data-sets. The LINCS drug expression dataset consisted of 1158 observations, with expression data of 7467 genes. OE and KD expression datasets consisted of 2413 and 4326 samples respectively, with expression data for 7467 genes.

Drug indications were extracted from Anatomical Therapeutic Chemical (ATC) classification system and the MEDication Indication Resource high precision subset (MEDI-HPS) [226]. The MEDI-HPS indication resource includes indications extracted from RxNorm, SIDER Side Effect Resource, MedlinePlus, and Wikipedia [226]. The high-precision subset (HPS) only considers medications indicated by RxNorm or those that appear in two out of three resources.HPS contains 13,304 unique indications for 2,136 medications [226]. Further validation of MEDI-HPS was also provided in a further study [227].

In the study, the code N05A of ATC was considered as the category of treatments for SCZ, and we used the same category for both SCZ and BP. Drugs under C02 and A10 of ATC were considered as treatments for HT and DM respectively. Drugs in MEDI-HPS are classified by ICD9, and then we used drugs under 714.0 for treatments of rheumatoid arthritis.

[1] https://github.com/dhimmel/lincs.

3.2.2 Methods

Here we proposed a general computational framework for prioritizing drug targets for further study, in which any classification algorithms are applicable. In this study, several ML methods, including support vector machine (SVM) [41], gradient boosting machine (GBM) with trees [63], random forest (RF) [28] and logistic regression with elastic net penalty (EN) [243] were employed for prediction modelling.

In the first step, we utilized the above ML methods to predict the 'treatment potential' of each drug for each disease under study. The gene expression were considered as features (predictors), while the indication (whether the drug indicated for the studied disease; coded 0 or 1) was considered as the outcome. The prediction model was then applied to expression profiles resulted from OE/KD to predict the probabilities of 'treatment potential' by over-expressing or knocking down corresponding genes. As explained in the introduction, particularly high or low predicted probabilities may indicate the gene as a potential drug target.

In this study we studied five kinds of diseases: hypertension (HT), type 2 diabetes mellitus (DM), rheumatoid arthritis (RA), bipolar disorder (BP), and schizophrenia (SCZ). Details of model specification, model evaluation and external validation are detailed below.

Model Specifications

As the number of drugs known for treat the disease is small, classes for positive outcomes and for negative outcomes are imbalanced. In practice, the strategy of balanced class weights was adopted, which places more emphasis on the minority class to balance the importance of the positive and negative classes, following a similar strategy as our previous study [240].

We implemented SVM, RF and GBM models using Python package "scikit-learn" [165]. Following the strategy recommended by [97], we adopted two-step hyper-parameter tuning with gridsearchCV provided by the package [165]. Specifically, we

first defined a broad hyper-parameter grid with a large step size along the axis of each parameter, a and then we refined the parameter grid based on predictive performance. Optimistic bias due to hyper-parameter tuning was avoided by nested cross-validation (see below).

In the study, we considered three hyperparameters for SVM, namely the kernel type , regularization parameter C and kernel coefficient γ. Here we used radial basis function (rbf) as the kernel, and C and γ were chosen from (-5, 15) and (-15, 3) in log-2 space respectively. For RF, the number of features considered for each splitting (*max_features*) and the minimum number of samples required at a leaf node (*min_samples_leaf*) were used to restrict the complexity of RF. We fixed the number of tree to 1000, and selected *max_features* and *min_samples_leaf* from {800, 1000, 1500, 2000, 3000, 5000} and {1, 3, 5, 10, 30, 50, 80} respectively. Like RF, gradient boosting machine (GBM) is an ensemble method, but in a given iteration GBM gives more emphasis to observations that are misclassified in previous iterations. For GBM, learning rate was chosen from {0.005, 0.01, 0.015, 0.02, 0.03, 0.05}, the number of boosting iterations from the sequence from 100 to 1001 with step size 50, maximum depth of each estimator from {2, 3, 5, 10} and maximum number of features from {10, 30, 50, 100, 500, 1000}. Subsampling proportion was fixed to 1. Finally, logistic regression with elastic net regularization (EN) was implemented using the R package "glmnet", with the mixing parameter α ranging from 0 to 1 (step size 0.1) and λ using the default range by glmnet. In the model, α regulates the sparsity (balance between L1 and L2 penalty), and λ is responsible for overall regularization.

In the study, a nested 5-fold cross validation (CV) was employed to choose the best hyperparameters and evaluate performance for each ML algorithm. Here we repeatedly split the data into three pieces, namely training, validation and test sets. Learning algorithms were trained on the training set and hyperparameters were chosen based on the validation set. The performance of ML models were evaluated on the test set, which is independent from the the dataset for model training and validation.

Compared to simple CV, nested CV can evaluate model performance more accurately [217]. The splitting of datasets in the nested CV is the same for different ML models by setting the same random seed.

Predictive Performance Evaluation

The performance of different ML models was evaluated by log Loss, area under the receiver operating characteristic curve (ROC-AUC) and area under the precision recall curve (PR-AUC). Log loss, related to cross-entropy, computes the negative log-likelihood of the true labels predicted by classification models, and the smaller values of log loss indicates that the model performs better. ROC-AUC measures the area under the curve of true positive rate (TPR) against the false positive rate (FPR). On the other hand, PR-AUC measures the area under the curve precision against recall. PR-AUC may be useful in model evaluation for imbalanced datasets[47].

External Validation of drug targets

We performed validation of our approach by testing if it can 'rediscover' known or potential drug targets for diseases based on other lines of evidence. Drug target data was downloaded from Open Targets [112] to validate our results. Note that our approach is independent of all kinds of evidence used to defined targets in the database. Intuitively, we are interested in whether the predicted probabilities of known targets from our ML models are significantly different from those of other genes.

Open Targets provides a continuous score ranging from 0 to 1 to indicate the association strength between targets and disease. We used a sequence of cutoff ranging from 0 to 1 with step size 0.2. For each cutoff, the targets with scores greater than the cutoff were considered as drug targets for the disease. We conducted t-tests to examine if the ML-predicted probabilities of genes listed by Open Targets were significantly different from those of other unmatched genes. The false discovery rate (FDR) approach was used to control for multiple testing [22].

3.3 Results

3.3.1 Model Performance

Average predictive performance of different ML models, measured in log loss, AUC-ROC and AUC-PR, is presented in Table 3.1. It shows that SVM performed the best cross four datasets in term of log loss, and the performance of RF and GBM were similar, slight worse than that of SVM, but the difference is small.

Table 3.1: Average predictive performance of different machine learning methods across four datasets

	ATC DM	ATC HT	MEDI-HPS RA	ATC SCZ
		Average Log Loss		
SVM	**0.08**	**0.2366**	**0.1209**	**0.141**
RF	0.0836	0.2471	0.1261	0.1462
GBM	0.0875	0.2523	0.1308	0.1555
EN	0.5752	0.6781	0.6312	0.5114
		Average AUC-ROC		
SVM	0.6232	**0.5433**	0.4972	**0.7582**
RF	0.6024	0.5488	0.5706	0.7377
GBM	0.5404	0.5516	0.5244	0.7474
EN	**0.6485**	0.5506	**0.5788**	0.7496
		Average AUC-PR		
SVM	**0.0834**	0.0804	0.0649	**0.2402**
RF	0.0616	0.0884	0.0471	0.2113
GBM	0.0578	**0.0937**	0.0485	0.2106
EN	0.0338	0.0792	**0.0706**	0.2362

1. The figure for best performance of learning algorithms for each dataset for each evaluation metric is in bold.
2. MEDI-HPS: MEDication Indication-High Precision Subset; ATC: Anatomical Therapeutic Chemical classification.
3. Abbreviations: DM stands for diabetes mellitus, HT for hypertension, SCZ for schizophrenia, RA for rheumatoid arthritis, BP for bipolar disorders.

When considering AUC-ROC as the performance evaluation metric, we find that SVM and EN had the best performance in two datasets. Specifically, SVM outperforms

other methods in ATC-HT and ATC-SCZ, while EN achieves the best performance in the other two datasets. All ML models performed better in ATC-SCZ data than in the other three datasets. In term of AUC-PR, the performance of ML methods varied. SVM outperformed other methods in ATC-DM and ATC-SCZ datasets, but GBM and EN showed the best performance in other two datasets. Even though in one of our previous studies [240] we have done a similar analysis on SCZ, this study on SCZ is an independent analysis from the previous one.

3.3.2 External Validation

Table 3.2: enrichment for target genes of HT by results on ATC-HT dataset

	SVM		RF		GBM		EN	
thresholds	*P-value*	*q-value*	*P-value*	*q-value*	*P-value*	*q-value*	*P-value*	*q-value*
1	4.81E-03	5.77E-03	3.31E-02	3.97E-02	9.81E-02	1.18E-01	2.09E-02	**2.60E-02**
0.8	4.32E-03	5.77E-03	2.56E-02	3.84E-02	8.29E-02	1.18E-01	1.61E-02	**2.60E-02**
0.6	**4.41E-04**	**2.48E-03**	4.94E-03	**1.48E-02**	1.76E-02	**5.28E-02**	**5.68E-03**	**2.60E-02**
0.4	8.26E-04	2.48E-03	**4.86E-03**	**1.48E-02**	**1.60E-02**	5.28E-02	1.12E-02	**2.60E-02**
0.2	2.04E-03	4.08E-03	1.19E-02	2.38E-02	2.91E-02	5.82E-02	2.17E-02	**2.60E-02**
0	1.56E-01	1.56E-01	6.84E-01	6.84E-01	1.96E-01	1.96E-01	1.12E-01	1.12E-01

1. Figures in the table are p-values calculated by two tailed t-test with alternative hypothesis that the mean predicted probability of genes listed by Open Targets is different than those of other genes.
2. The lowest p-values/q-values for every ML model in each dataset are in bold.
3. Abbreviations are defined the same as Table 3.1.

Table 3.3: enrichment for target genes of DM by results on ATC-DM dataset

	SVM		RF		GBM		EN	
thresholds	*P-value*	*q-value*	*P-value*	*q-value*	*P-value*	*q-value*	*P-value*	*q-value*
1	5.78E-02	8.67E-02	6.53E-02	9.80E-02	3.28E-05	8.62E-05	2.32E-03	4.64E-03
0.8	2.33E-02	4.66E-02	2.13E-02	6.39E-02	4.31E-05	8.62E-05	1.67E-03	4.64E-03
0.6	**1.26E-02**	**4.66E-02**	**1.37E-02**	**6.39E-02**	**1.61E-05**	**8.62E-05**	**1.57E-03**	**4.64E-03**
0.4	2.32E-02	4.66E-02	4.32E-02	8.64E-02	1.72E-03	2.58E-03	5.22E-03	7.83E-03
0.2	6.43E-01	6.43E-01	3.67E-01	4.40E-01	1.71E-02	2.05E-02	3.55E-02	4.26E-02
0	3.00E-01	3.60E-01	6.24E-01	6.24E-01	8.10E-01	8.10E-01	8.21E-02	8.21E-02

1. Illustrations of the table are the same as Table 3.2

Table 3.4: enrichment for target genes of RA by results on MEDI-HPS RA dataset

	SVM	RF	RF		GBM		EN	
thresholds	*P-value*	*q-value*	*P-value*	*q-value*	*P-value*	*q-value*	*P-value*	*q-value*
1	1.18E-01	2.72E-01	6.23E-04	1.87E-03	2.01E-02	4.50E-02	9.22E-01	9.94E-01
0.8	1.36E-01	**2.72E-01**	**3.93E-04**	**1.87E-03**	**8.53E-03**	**4.50E-02**	9.94E-01	9.94E-01
0.6	**1.18E-01**	2.72E-01	1.41E-01	1.69E-01	**3.67E-01**	3.67E-01	9.20E-01	9.94E-01
0.4	6.44E-01	6.44E-01	1.14E-02	2.28E-02	2.25E-02	4.50E-02	2.69E-01	5.38E-01
0.2	3.12E-01	4.45E-01	8.47E-02	1.27E-01	4.15E-02	6.23E-02	8.48E-02	**5.09E-01**
0	3.71E-01	4.45E-01	7.01E-01	7.01E-01	1.96E-01	2.35E-01	**2.56E-01**	5.38E-01

1. Illustrations of the table are the same as Table 3.2

Table 3.5: enrichment for target genes of DM by results on ATC SCZ dataset

	SVM	RF	RF		GBM		EN	
thresholds	*P-value*	*q-value*	*P-value*	*q-value*	*P-value*	*q-value*	*P-value*	*q-value*
1	3.32E-01	4.98E-01	2.84E-01	5.68E-01	**2.57E-01**	**7.92E-01**	3.56E-01	**5.76E-01**
0.8	4.66E-01	5.59E-01	2.47E-01	**5.68E-01**	2.64E-01	7.92E-01	**2.80E-01**	5.76E-01
0.6	2.18E-02	6.54E-02	7.78E-01	8.97E-01	9.94E-01	9.94E-01	7.47E-01	7.47E-01
0.4	**1.91E-02**	6.54E-02	8.62E-01	8.97E-01	7.97E-01	9.94E-01	3.84E-01	5.76E-01
0.2	7.00E-02	**1.40E-01**	8.97E-01	8.97E-01	5.42E-01	9.94E-01	7.18E-01	7.47E-01
0	7.11E-01	7.11E-01	**1.85E-01**	5.68E-01	9.01E-01	9.94E-01	3.14E-01	5.76E-01

1. Illustrations of the table are the same as Table 3.2

Table 3.6: enrichment for target genes of BP by results on ATC SCZ dataset

	SVM	RF	RF		GBM		EN	
thresholds	*P-value*	*q-value*	*P-value*	*q-value*	*P-value*	*q-value*	*P-value*	*q-value*
1	4.14E-01	6.99E-01	7.24E-01	7.97E-01	5.88E-01	7.06E-01	5.59E-01	6.71E-01
0.8	4.66E-01	6.99E-01	7.58E-01	**7.97E-01**	9.43E-01	9.43E-01	9.43E-01	9.43E-01
0.6	8.57E-01	8.57E-01	6.84E-01	7.97E-01	2.56E-01	3.84E-01	2.90E-02	5.80E-02
0.4	8.57E-01	8.57E-01	6.84E-01	7.97E-01	2.56E-01	3.84E-01	2.90E-02	5.80E-02
0.2	4.13E-01	6.99E-01	**3.17E-01**	7.97E-01	**6.03E-02**	**3.62E-01**	**1.45E-03**	**8.70E-03**
0	**1.31E-02**	**7.86E-02**	7.97E-01	7.97E-01	1.91E-01	3.84E-01	4.76E-01	6.71E-01

1. Illustrations of the table are the same as Table 3.2

The results of enrichment test (for validation of drug targets) for over-expressed genes are shown in Tables 3.2, 3.3, 3.4, 3.5, 3.6. No statistically significant enrichment was observed for KD genes, and the results are shown in supplementary tables.

For DM and HT (3.2, 3.3), we observed significant enrichment across multiple thresholds with FDR < 0.05, indicating the proposed method indeed 're-discovered' known targets more than expected by chance. For RA, significant enrichment was mainly observed for prediction models based on RF or GBM. 3.4). For SCZ and BP which shared anti-psychotics as treatment, the enrichment was not as strong, but suggestive enrichment (FDR < 0.2) were observed for SVM in SCZ and EN in BP (3.5 3.6)

3.3.3 Literature Support

Our model generated several potential druggable targets for different diseases. We extracted the top candidates (30 with the highest and lowest predicted probabilities respectively) and we highlight several targets that are supported by previous studies here and list them in Table 3.7.

Some of the top potential DM drug targets suggested by our model included SGK1, ESR1, CISH, MAPK4K4, UGCG. Our model suggests that inhibition of SGK1 is may be therapeutically useful on DM. In diabetic animal models, SGK1 expressions were up-regulated [91, 232, 32], which is likely caused by production of advanced glycation products (AGE) [91, 32]. SGK1 is critical to the development of diabetic neuropathy, and inhibition of SGK1 by fluvastatin has shown favorable effects in ameliorating progression of DM [232]. CISH (Cytokine-inducible SH2) containing Cish protein is involved in the signaling pathway of human growth hormone (hGC), and induction of this pathway by hGC minigene has been shown to promote beta-islet cell proliferation in murine model [14]. MAPK4K4 is also a drug target suggested by our model for DM. In a cell culture experiment, inhibition of MAPK4K4 by RNA interference was able to completely reverse the insulin resisting effect of TNF-alpha [27]. UGCG

Table 3.7: Some literature-supported candidates selected from top hits derived from machine learning methods

Disease	Gene	Description
DM	SGK1	inhibition of SGK1 by fluvastatin has shown favorable effects in ameliorating progression of DM
	CISH	CISH is involved in the signaling pathway of hGC, and induction of this pathway by hGC minigene has been shown to promote beta-islet cell proliferation in murine model
	MAPK4K4	In a cell culture experiment, inhibition of MAPK4K4 by RNA interference is able to completely reverse the insulin resisting effect of TNF-alpha
	UGCG	UGCG is known as an inhibitor of insulin sensitivity
RA	TFF2	TFF2 was significantly upregulated in RA samples.
	SP	SP has been shown to improve actions of dexamethasone for the treatment of arthritis in rats
	CTNNBIP-1	It encodes β-catenin, a critical player of the Wnt signaling pathway that initiates arthritic progression and development in joints.
	PSMB8	PSMB8 encodes the β5i subunit, known as LMP7, and findings suggest that selective inhibitors for LMP7 may be repositioned to treat RA.
	NR4A2	an orphan nuclear receptor responsible for proinflammatory responses particularly to IL-8 in RA
	TNF	Currently, much of the DMARD developed and in clinical use are against TNF.
HT	Angiotensin II	Antagonizing angiotensin II and glucocorticoid is a well-known popular therapeutic approach to treat hypertension.
	Glucocorticoid	
	SERPINA6	SERPINA6 deficiency usually leads low blood pressure
	ANP	ANP regulates renal sodium excretion and reduces extracellular fluid volume
SCZ	PIP5K2A	PIP5K2A acts as an agonist of amino acid transporter EAAT3, which facilities glutamate reuptake.
		Glutamatergic transmission has been implicated in SCZ as an important mechanism of disease
	DRD1	there is emerging evidence that D1 receptors may contribute to negative symptoms
	HTR2C	5-HT2c receptor is the strongest candidate for contributing to their antipsychotic action.
BPD	JUN	JUN is a subunit of AP-1, and the potential beneficial effect of AP-1 is due to its neurotrophic and neuroprotective effects
	TH	BIP is associated with an increased amount of tyrosine hydroxylase, and that chronic treatment of valproic acid has been shown to reduce the mRNA level of TH

Abbr.: DM stands for diabetes mellitus, RA for rheumatoid arthritis, HT for hypertension, SCZ for schizophrenia, and BPD for bipolar disorders.

(Ceramide and its metabolites) is known as an inhibitor of insulin sensitivity. Aerts et al. demonstrated that both ceramide and its downstream metabolites could reduce insulin sensitivity in *vivo* and that treatment with AMP-DNM could reverse insulin insensitivity [1].

Our approach also identified several promising targets for rheumatoid arthritis. it has long been shown that substance P inhibition has an important role in the pathophysiology of RA [129, 71, 79, 109, 134, 158, 120, 119]. It serves as a pain transmitter, exacerbates the inflammatory process in arthritic joints and worsens disease progression. Our study identifies SP inhibition as a favorable drug target, and this has already been proven in previous studies. For example, substance P has been shown to improve actions of dexamethasone for the treatment of arthritis in rats [118]. CTNNBIP-1, which encodes for β-catenin, is identified as a promising therapeutic target by our model, and it has been studied extensively. β-catenin is a critical player of the Wnt signaling pathway that initiates arthritic progression and development in joints [182, 241], and dysfunction of this pathway was considered a disease model for RA [229, 241].

LMP7 is another top target we found. It is a subunit of the immunoproteosome important for generating peptide fragments on the MHC class I receptor [104], implicating the role it plays in autoimmune diseases like RA. Basler et al. found that LMP7 alone was not sufficient, and that LMP2 must also be co-inhibited in order to block immunity [18]. These findings suggest that selective inhibitors for LMP7, such as ONX 0914, which is already used for other autoimmune diseases [5, 130, 218], may be repositioned to treat RA.

Our model also suggests that antagonizing NR4A2 could help treat RA, which is consistent with previous findings. NR4A2 is an orphan nuclear receptor that is responsible for proinflammatory responses particularly to IL-8 in RA [3]. It was also found that it plays as a downstream response element of TNF-alpha [148] as well as a transcription factor for the immunomodulatory peptide hormone prolactin [145].

Currently, many of the Disease-modifying anti-rheumatic drugs (DMARDs) in clinical use were developed against TNF. Our SVM model shows this receptor may bring about therapeutic effects.

Known popular targets discovered by our approach for hypertension include those related to angiotensin and glucocorticoid regulation. For example, our model uncovers AGT (angiotensinogen) as a potential target and suggest inhibition may lead to therapeutic effects. AGT is well-known as a component in the renin-angiotensin system, a system critical in regulating blood pressure. Angiotensinogen can be converted to angiotensin I then angiotensin II, which causes vasoconstriction. Our model also suggests that inhibiting the corticosteroid-binding protein SERPINA6 will reduce blood pressure. This is also a biological plausible candidate, and it has been shown that SERPINA6 deficiency usually leads low blood pressure [209]. Another target we found is atrial natriuretic peptide, which regulates renal sodium excretion and reduces extracellular fluid volume. The direction of effect also agrees with that predicted from the model (up-regulation beneficial for treating HT).

For SCZ, we also observed a number of gene candidates with evidence in the pathophsyiology of SCZ. For example, PIP5K2A acts as an agonist of amino acid transporter EAAT3, which facilities glutamate reuptake. Glutamatergic transmission has been implicated in SCZ as an important mechanism of disease.The dopaminergic pathways in the brain play crucial roles in the pathophysiology of SCZ. Our model suggests that a potential treatment approach is to agonize the D1 receptor. Despite the dominant role of D2 in the nigrostriatal pathway [96] which makes D2 antagonists the primary anti-psychotic agents, there is emerging evidence that D1 receptors may contribute to negative symptoms [72]. Non-human primates models demonstrate that administration of D1 selective agonist enhance spatial working memory, but it is not the case for D2 agonist [72]. Another possible candidate to deal with negative symptoms of SCZ include 5-HT2C receptor agonists, which is also implicated by our analysis. Although the first line of treatment for SCZ is usually atypical

antipsychotics which are agonists to 5-HT2A, pharmacological studies on 5-HT2C agonists has shown promising effects in animal models [60, 86, 169]. It is suggested that activation of this receptor will lead to a selective suppression of DA release in ventral tegmental area (VTA) and nucleus accumbens (Nac) but not the nigrostriatal DA neuron firing [147].

Like SCZ, our approach discovered several drug target candidates for bipolar disorder (BPD) for which genetic associations were previously reported (e.g. [122, 239, 73]). One of the top targets from our analysis included the JUN (a subunit of AP-1) gene. Interestingly, AP-1 can be increased by lithium in human neuroblastoma cell lines and rat brain tissues and a similar effect of valproic acid on AP-1 has also been observed [237, 9, 161]. The potential beneficial effect of AP-1 is due to its neurotrophic and neuroprotective effects [138]. Another target we found was TH (tyrosine hydroxylase) and the model suggests inhibition of TH may be a useful treatment option. Some previous studies have shown that BIP is associated with an increased amount of tyrosine hydroxylase, and that chronic treatment of valproic acid has been shown to reduce the mRNA level of TH [163, 149].

3.4 Discussion

In this study, we presented a novel computational approach to identify promising drug targets by incorporating gene expression profiles. This approach is general as it can widely incorporate any classification algorithms. Four state-of-the art machine learning methods were used to demonstrate the potential of our approach, and their performances were compared by three different evaluation metrics. Results of enrichment test show that top genes from our models are enriched for targets from Open Targets for diabetes mellitus, hypertension, schizophrenia, bipolar disorder, and rheumatoid arthritis. Some of the top potential drug targets identified by our approach are also supported by previous studies. As far as we know, this work is the first

to directly employ machine learning methods on both drug-induced and genetically-perturbed expression data to discover potential drug targets for specific diseases.

The study adopted four ML methods, which are capable of learning either linear or nonlinear relationship present in data. Because the number of variables is much larger than the number of observations ($p \gg n$), certain regularization methods are required. Logistic regression with elastic net penalty (EN) can automatically select features that contribute the most to prediction. In our case, the performance of EN is quite reasonable, even though it only models linear relationship. SVM uses kernel tricks to map the feature into higher dimension feature space. On the other hand, random forests (RF) and gradient boosting machine (GBM) are well-known tree-based methods. The basic idea behind them is to assemble a number of 'weak' (tree-based) learners to make accurate predictions. A main difference between the two is that GBM is a sequential tree model, which adjusts the importance of observations in learning according to the performance of previously fitted trees. For GBM and RF, their performances are comparable overall across the four datasets. In this study, we do not put our emphasis on particular algorithms, since we are interested in demonstrating the potential of our framework in prioritizing drug target candidates. Although a number of targets listed are supported by previous studies, further experimental validation is necessary to confirm the findings. The presented algorithm aims to *prioritize* drug targets, which may be useful as an independent source of evidence to support further experimental studies of certain candidates.

We have employed enrichment tests to validate the overall validity of our approach. While animal studies or other experimental validations may provide stronger evidence for individual targets, such validation cannot be easily carried out on a large number of targets. Also, validation on a few targets cannot provide evidence for the *overall validity* of the presented framework, because there may be chance findings.

Our approach is general and highly flexible. However, there are several limitations. One limitation is that our dataset for ML prediction model building is highly

imbalanced, as usually only a small number of drugs are indicated for each disease. In order to address this issue we increased class weight of the minority group. There are also other strategies to address issue such as SMOTE (Synthetic Minority Oversampling Technique) [33], but whether strategies like SMOTE can address this issue in high dimensional settings is still unclear. This will be a topic further investigation. Further, we may also resort to more advanced or recently developed machine learning methods, such as deep neural networks, to uncover new targets. However, such methods may be more useful in larger samples.

Another important aspect is that we observed significant enrichment in OE datasets but not in KD datasets. One hypothesis is that some off-target effects may interfere with the expression profiles in KD experiments, leading to greater difficulties in finding relevant drug targets. How to overcome or reduce the influence of off-target effects remain an area for further studies.

3.5 Conclusion

This study presented a general computational framework to prioritize drug targets for various diseases. Under the framework, different kinds of ML methods can be utilized. We applied four ML methods to identify potential drug target of five disorders. External validation shows that the top candidates are enriched for targets selected by independent lines of evidence from a large external database(Open Targets). Some top target genes were also supported by previous studies.

Finding promising drug targets for diseases is crucial to drug development. However, it is impractical to perform in-depth experimental studies on every possible target for each disease. Computational methods offer a cheap, fast and systematic high-throughput approach to guide prioritization of drug targets. We hope our presented framework will provide an additional way to prioritize drug targets for development, which is independent of and may be combined with other existing sources of data.

□ **End of chapter.**

Chapter 4

Evaluating Individualized Treatment Effects (ITE) of Risk Factors on Patient Outcomes

4.1 Motivation

Traditional biomedical or clinical studies in the area of estimating treatment effect mainly focus on the average effect of risk factors (RFs) or treatment (tx) in population level. However, in the clinical environment we can easily find that the same risk factor may affect patients differently. Thus, patients may pay more attention to how a risk factor will affect them in an individual level rather than in a population level, given their clinical backgrounds and genetic characteristics. The main aim of this study is to resolve this concern by estimating the ITEs for each patient, with consideration of their unique genetic and clinical information. In this study we treat the two terms ”risk factor” and ”treatment” conceptually equivalent, since a risk factor can be considered as a ”treatment” with adverse effects. The approach for estimating the ITEs for each patient allows us to offer tailored management to individual patients. This enables us to deliver more cost-effective prevention or treatment strategies to benefit

them the most. This idea is also in the line with ”personalized medicine”, which has been advocated in recent years.

In spite of an increasing number of studies in this area, current studies in ITE especially in biomedical research are rather limited. Some critical limitations include a lack of well-established validation methods for treatment effect estimations and the contribution of key features in ITE estimation, and failure in incorporating censored data. Even though genetic factors may determine heterogeneous response to tx/RFs, especially to cancer treatments [59], current studies on ITE have not included genomic features. Here we proposed several methodologies to address the above limitations and applied the ITE framework to genomic data. In our approach genomic features were considered as risk factors or covariates that contribute to the heterogeneity of treatment effects.

4.2 Background

It has been well-known that different individuals response differently to the same risk factor (RF) or treatment (tx). For example, even though obesity is a risk factor for cardiometabolic (CM) diseases, there still are obese subjects who don't develop related complications [152]. The type and severity of such CM complications can also show heterogeneity among subjects [152]. Another evidence is that not all people suffering stressful life events are affected by depression, even if stressful events are risk factors for depression [234]. This fact can also be applied to other RFs or treatments. The heterogeneous effect can be contributed by different genetic and/or environmental factors of subjects, and these factors affect them differently. Here we would like to investigate the different treatment effect contributed by variants or mutations instead of clinical factors, since studies have shown that same variant/mutation can have varying outcomes on different subjects [2, 156, 214].

There are dramatic advances in omics technology and a sharp rising availability

of biomedical data. However, current studies in cancer still mainly focus on one clinical/genetic RF at a time, without the consideration of presence of complex interactions among the subject's genetic and/or clinical factors.

One of most crucial concerns to patients is how a RF or treatment will affect them given their genetic and clinical information. However, current researches on this issue largely focus on the average treatment effect of RFs in population rather than individualized treatment effects.

Here we built a computational framework to unravel the individualized effects of RFs/treatment so that we can estimate the treatment effect for each individual with the incorporation of his/her genetic and/or clinical background. We also developed methods to discover genetic and/or clinical features contributing the most to the estimation. We employed our approaches to cancer data to estimate treatment effects of genetic changes (e.g. changes of expression level of risk genes, mutations, CNVs etc.) and other RFs on each individual's survival.

4.3 Overview of Related Work

4.3.1 Background Methods

Here we define notations for the clarification of following presentation. Let $\mathbf{X}^{n\times m}$ denotes the covariate variable, $\mathbf{Y}^n$ denotes outcome variable, and $\mathbf{W}^n$ denotes treatment variable. Given an observation i we denote its covariates as $\mathbf{x_i}^m$, the risk factor/tx status as w_i and outcome y_i. Here we restate that since a risk factor may also be considered as a "treatment" with adverse effects, methods for ITE estimation can also be employed to RFs. Assume that the outcome $\mathbf{Y}$ satisfy

$$\mathbf{Y} = f(\mathbf{X}, \mathbf{W}) + \varepsilon, \tag{4.1}$$

Where ε follows $N(\mathbf{0}, \mathbf{\Sigma})$. $\mathbf{\Sigma}$ is the covariance matrix.

Assumption assume $\mathbf{X}, \mathbf{W}, \mathbf{Y}$ fulfill unconfoundedness assumption (randomiza-

tion conditional on the covariates),

$$\big[\mathbf{Y}_i(1), \mathbf{Y}_i(0)\big] \perp\!\!\!\perp \mathbf{W}_i \,|\, \mathbf{X}_i. \tag{4.2}$$

Under the unconfoundedness assumption 4.2 the key is to estimate the expected difference, the estimation of ITE, for each individual in response between treatment and control. The ITE for subject i is formulated as

$$\tau(\mathbf{x_i}) = \mu_1(\mathbf{x_i}) - \mu_0(\mathbf{x_i}), \tag{4.3}$$

with $\mu_1(x_i)$ and $\mu_0(x_i)$ defined as

$$\begin{aligned} \mu_1(x_i) &= \mathbb{E}(\mathbf{Y} = y_i | \mathbf{X} = \mathbf{x_i}, \mathbf{W} = w_i) = f(x_i, 1) \\ \mu_0(x_i) &= \mathbb{E}(\mathbf{Y} = y_i | \mathbf{X} = \mathbf{x_i}, \mathbf{W} = 1 - w_i) = f(x_i, 0) \end{aligned} \tag{4.4}$$

respectively, where $w_i = 1$ without the loss of generality. For a given subject i, y_i and w_i are scalars, and $\mathbf{x_i}$ is a vector of length m. Traditional machine learning method cannot handle this situation since they cannot capture the difference of ITE when the outcomes for RF/tx were absent or present.

A traditional solution to measure ITE is to estimate the difference of averaged outcome between treatments and controls in pre-specified subgroups [67] or subgroup defined by learning algorithms [199, 198, 11, 61]. Su et. al. employed interaction tree to iteratively searching subgroups based on treatment effect [199, 198]. Similarly, causal trees proposed by Athey and Imbens estimate the treatment effect at the leaves of the tree [11]. However, main drawbacks of the approach are that there is no ground truth for subgroup definition, and that the impediment of iteratively searching for subgroups present obvious treatment effect and reporting only the results for subgroups with extreme treatment effects to highlight heterogeneity may be highly spurious [10, 40]. In the high dimensional setting, it's still very challenging to divide subjects into appropriate subgroups [171]. It's the same case for genomic data.

Alternatively, a feasible approach is to use any supervised machine learning (SML) methods to fit $\mu_1(\mathbf{x})$ and $\mu_0(\mathbf{x})$ simultaneously / separately and estimate the difference by putting them together. Specifically, one may fit a single model $\mu(\mathbf{x}, w)$ or separate models for the treated and control groups, and compute the different between $\mu(\mathbf{x}, w)$ and $\mu(\mathbf{x}, 1-w)$. Studies [136, 45] utilize different counterfactual random forests algorithms to estimates treatment effects by fitting separate random forests models to treatment and control groups. Several literature, including Green and Kern [78], Hill [89], and Hill and Su [90], has employed bayesian forest-based machine learning methods to estimate heterogeneous treatment effects. These studies utilize Bayesian additive regression tree (BART) method [37], and can obtain reliable intervals for treatment effects by MCMC sampling. For lasso-like methods for causal inference [100, 206], it's difficult to capture interactions, which may be naturally present in genomic data, in high-dimensional setting, in split of its simplicity and good ability in feature selection. A limitation of these studies is lack of formal statistical inference results [220]. Some other methods, like Meta learners and deep learning based, for ITE estimation could be found in [117, 105].

Here we are interested in methods with following characteristics: automatically select important features, well capture high interactions present, and have good asymptotic properties. Thus, methods such as causal forests [220] or GRF [12] are much more preferred. Causal forests [220] have been proposed with an objective to maximize the heterogeneity of $\tau(\mathbf{x})$. Recently, Athey et al. proposed an extension of causal forests, GRF [12], inspired by the R-learner proposed in [155]. Both of the two methods [220, 12] inherits the excellent capability of random forest in capturing complex interactions.

However, there are still substantial research gaps, including relative lack of methods for result validation, evaluation of key features contributing to ITE estimation and handling censored data. Here we proposed methods to address these key issues and pioneer new applications to genomic data, which is the first of its kind.

4.3.2 Causal Forests

In this section, we will explain causal forests (CF) technically, a basis of our ITE framework. CF [220] originate from random forests [28], which are related to kernel or nearest neighborhood methods. However, random forests differ in that they determine weights received by nearby observations in a data-driven way, and this characteristics is critical in high dimensional environment or the present of high order interactions among covariates [220]. This is the same case for CF.

Here we begin with causal trees (CT) [11] since CF are made of a number of CT. In this part we follow a similar notations as appeared in [11]. A tree can be considered as a partitioning of the feature space $\mathbb{X}$, denoting as Π. A partition Π with a number of elements $\#(\Pi)$ can be written as

$$\Pi = \{l_1, l_2, \ldots, l_{\#(\Pi)}\},$$

and a union of all elements in partition is the whole feature space $\mathbb{X}$.

Let $\mathbb{P}$ denote the space of partitions, and $\mathbb{S}$ be the space of samples from a population of observations. We seek for a algorithm $\pi : \mathbb{S} \to \mathbb{P}$ that splits sample space $\mathbb{S}$ into partition Π.

Given a partition Π and sample S, the estimated conditional mean for observation $\mathbf{x}$ is

$$\hat{\mu}(\mathbf{x};\Pi) \equiv \frac{1}{\#(i \in \mathrm{S} : \mathbf{x}_i \in l(\mathbf{x};\Pi))} \sum_{i \in \mathrm{S}:\mathbf{x}_i \in l(\mathbf{x};\Pi)} y_i,$$

which is an unbiased estimator for $\mu(\mathbf{x};\Pi)$. Here $l(\mathbf{x};\Pi)$ is the leaf to which $\mathbf{x}$ belongs.

A adjusted MSE criteria including $\mathbb{E}[\mathbf{Y}_i^2]$, a term that does not depend on the estimator, is defined as

$$\mathrm{MSE}_\mu(\mathrm{S}^{\mathrm{te}}, \mathrm{S}^{\mathrm{est}}, \Pi) \equiv \frac{1}{\#(\mathrm{S}^{\mathrm{te}})} \sum_{i \in \mathrm{S}^{\mathrm{te}}} \left\{ (y_i - \hat{\mu}(\mathbf{x}_i; \mathrm{S}^{\mathrm{est}}, \Pi))^2 - y_i^2 \right\}. \tag{4.5}$$

The expectation of the modified MSE is

$$\mathrm{EMSE}_{\mu}(\Pi) \equiv \mathbb{E}_{\mathrm{S}^{\mathrm{te}},\mathrm{S}^{\mathrm{est}}}\left[\mathrm{MSE}_{\mu}(\mathrm{S}^{\mathrm{te}},\mathrm{S}^{\mathrm{est}},\Pi)\right].$$

The objective of CT is to maximize the criteria

$$Q^{\mathrm{H}}(\pi) \equiv -\mathbb{E}_{\mathrm{S}^{\mathrm{te}},\mathrm{S}^{\mathrm{est}},\mathrm{S}^{\mathrm{tr}}}\left[\mathrm{MSE}_{\mu}(\mathrm{S}^{\mathrm{te}},\mathrm{S}^{\mathrm{est}},\pi(\mathrm{S}^{\mathrm{tr}}))\right]. \tag{4.6}$$

This criteria shows better convergence properties of confidence intervals, compared with conventional practice that $\mathrm{S}^{\mathrm{est}}$ and S^{tr} are the same sample for both tree construction and estimation [220].

With the above setup for treatment effect estimation, a similar definition to 4.5, is defined as

$$\mathrm{MSE}_{\tau}(\mathrm{S}^{\mathrm{te}},\mathrm{S}^{\mathrm{est}},\Pi) \equiv \frac{1}{\#(\mathrm{S}^{\mathrm{te}})}\sum_{i\in\mathrm{S}^{\mathrm{te}}}\left\{(\tau_i-\hat{\tau}(\mathbf{x}_i;\mathrm{S}^{\mathrm{est}},\Pi))^2-\tau_i^2\right\}. \tag{4.7}$$

In reality, τ_i cannot be directly observed. The estimated counterparts are defined as

$$\hat{\tau}(\mathbf{x};\mathrm{S},\Pi) \equiv \hat{\mu}(1,\mathbf{x},\mathrm{S},\Pi)-\hat{\mu}(1,\mathbf{x};\mathrm{S},\Pi), \tag{4.8}$$

where

$$\hat{\mu}(w,\mathbf{x},\mathrm{S},\Pi) \equiv \frac{1}{\#(i\in\mathrm{S}_{\mathrm{w}}:\mathbf{x}_i\in l(\mathbf{x};\Pi))}\sum_{i\in\mathrm{S}_{\mathrm{w}}:\mathbf{x}_i\in l(\mathbf{x};\Pi)} y_i^{\mathrm{obs}}.$$

With the fact that $\hat{\tau}$ is constant within each leaf and the fact that

$$\mathbb{E}_{\mathrm{S}^{\mathrm{te}}}[\tau_i|i\in\mathrm{S}^{\mathrm{te}}:i\in l(\mathbf{x},\Pi)] = \mathbb{E}_{\mathrm{S}^{\mathrm{te}}}[\hat{\tau}(\mathbf{x};\mathrm{S}^{\mathrm{te}},\Pi)],$$

A crucial estimator for the infeasible in-sample goodness-of-fit criterion is derived as

$$-\mathrm{MSE}_{\tau}(\mathrm{S}^{\mathrm{tr}},\mathrm{S}^{\mathrm{tr}},\Pi) = \frac{1}{N^{\mathrm{tr}}}\sum_{i\in\mathrm{S}^{\mathrm{tr}}}\hat{\tau}^2(\mathbf{x}_i;\mathrm{S}^{\mathrm{tr}},\Pi). \tag{4.9}$$

Then this leads to an estimator for the criterion relying only on S^{tr} and N^{est}

$$-\mathrm{E}\hat{\mathrm{M}}\mathrm{SE}_{\tau}(\mathrm{S}^{\mathrm{tr}}, N^{\mathrm{est}}, \Pi) \equiv \frac{1}{N^{\mathrm{tr}}} \sum_{i \in \mathrm{S}^{\mathrm{tr}}} \hat{\tau}^2(\mathbf{x}_i; \mathrm{S}^{\mathrm{tr}}, \Pi) - \left(\frac{1}{N^{\mathrm{tr}}} + \frac{1}{N^{\mathrm{est}}}\right) \sum_{l \in \Pi} \left(\frac{\mathrm{S}^2_{\mathrm{S}^{\mathrm{tr}}_{\mathrm{tx}}}(l)}{p} + \frac{\mathrm{S}^2_{\mathrm{S}^{\mathrm{tr}}_{\mathrm{ctl}}}(l)}{1-p} \right) \tag{4.10}$$

where the last term in the second line of 4.10 is pooled within-leaf variance. Definitely the splits of the tree are chosen to maximize the variance of $\tau(\mathbf{x}_i)$. For more detailed derivations, refer to original publication [11].

Briefly, the objective for splitting for the adaptive version of CT, denoted CT-A, uses $-\mathrm{MSE}_{\tau}(\mathrm{S}^{\mathrm{tr}}, \mathrm{S}^{\mathrm{tr}}, \Pi)$. The same objective function also is applicable to CV version of CT-A but evaluated at the samples $\mathrm{S}^{\mathrm{tr,cv}}$ and $\mathrm{S}^{\mathrm{tr,cv}}$. The splitting objective function for the honest CTS, CT-H, is $-\mathrm{MSE}_{\tau}(\mathrm{S}^{\mathrm{tr}}, N^{\mathrm{est}}, \Pi)$. The same objective function also can be applied to a CV version of CT-H, but evaluated at the cross-validation sample $\mathrm{S}^{\mathrm{tr,cv}}$ with known $N^{\mathrm{est,cv}}$.

Procedure for causal forests with honesty and subsampling is proposed as follows: In above table, CT stands for causal tree algorithm defined above in the section.

Algorithm 1 Causal forests with honesty and subsampling

Require: a samples $\mathrm{S}^{n \times m}$ and pre-specified parameters, including the number of trees B, *mtry* and the sub-sampling rate s

1: **for all** i such that $0 \leq i \leq \mathrm{B}$ **do**
2: samples for model construction $\mathrm{S}^{\mathrm{tr,est}}_{\mathrm{i}} \leftarrow$ SUBSAMPLE (S, s), with remaining as test set $\mathrm{S}^{\mathrm{te}}_{\mathrm{i}}$
3: training samples $\mathrm{S}^{\mathrm{tr}}_{\mathrm{i}}, \mathrm{S}^{\mathrm{est}}_{\mathrm{i}} \leftarrow$ SUBSAMPLE ($\mathrm{S}^{\mathrm{tr,est}}_{\mathrm{i}}$)
4: causal tree $\mathrm{T}_{\mathrm{i}} \leftarrow \mathrm{CT}(\mathrm{S}^{\mathrm{tr}}_{\mathrm{i}}, \mathrm{S}^{\mathrm{est}}_{\mathrm{i}})$
5: make out-of-bag prediction $\hat{\tau}(\mathbf{x}^i_j)$ for $\mathbf{x}^i_j \in \mathrm{S}^{\mathrm{te}}_{\mathrm{i}}$
6: **end for**
7: **return** CTs $\mathrm{T}_1, \mathrm{T}_2, \ldots, \mathrm{T}_{\mathrm{i}}, \ldots, \mathrm{T}_{\mathrm{B}}$ and $\hat{\tau}(\mathbf{x}_j)$ by averaging all available out-of-bag predictions $\tau^{(.)}_j$ for $\mathbf{x}_j \in S$.

4.4 ITE Framework

In order to assess the ITE of RFs on disease outcome and discover key features that contribute to estimation of the heterogeneity we proposed an analytic framework to estimate the ITE of RFs/tx, with genetic features as primary RFs and/or covariates. The term 'individualized treatment effect' (ITE) will be used regardless of an RF or treatment being considered, since we have declared that the two are conceptually equivalent entities in this study.

The estimated ITE may be formulated as 4.4 under a counterfactual outcomes framework [177]. In reality the true value of $\tau(\mathbf{x})$ cannot be directly observed, as we only have one of the two potential outcomes. In observational studies, the tx assignment may be associated with potential outcomes due to confounding variables. If the unconfoundedness assumption 4.2 satisfies then the causal ITE can still be captured. Then, in most observational studies, the study is still of significance in despite of the presence of residual confounding.

In that case we may still gain insights into TE heterogeneity at an association level, and covariates responsible for heterogeneity may still deserve further studies, and in practice a continuous variable $\mathbf{W}$ is also allowed [12]. When $\mathbf{W}$ is continuous, an average partial effect is estimated $Cov[\mathbf{Y}, \mathbf{W} \mid \mathbf{X} = \mathbf{x}]/Var[\mathbf{W} \mid \mathbf{X} = \mathbf{x}]$, which may be considered as the increase in Y given a unit increase in W, conditional on the covariates.

Our experiment studies rely on the framework of GRF, as it is a state-of-the-art approach which directly optimizes an objective function for ITE estimation, rather than fit conditional means for treatment and control observations. However, most of the methodologies and extensions presented in this study is data-driven such that it can be widely applied to any other ITE estimation models.

4.4.1 Novel Tests for the presence of heterogeneity

Unlike a SML model, an ITE model is not straightforward since the actual TE is not directly observed. There is no either empirically or theoretically well-established methods in ITE validation. We notice that the function `test_calibration` provided in the R package `grf` [207], and the idea behind it is borrowed from [35]. Thus, we would compare our methods with it in simulations in future section 4.5.

We would propose several novel statistical tests for ITE model evaluation:

Split-half correlation with multiple splits This method borrows the idea of cross validation (CV) in SML. Briefly, we proposed to split the dataset into 2 halves. Specifically, an ITE model is first fitted on the 1st half, then applied to the 2nd half to predict $\tau(\mathbf{x})$. The process can be repeated by reversing the training and testing sets. Then for each half, we have out-of-bag predictions $\tau_{\text{oob}}(\mathbf{x})$ and predictions $\tau_{\text{pred}}(\mathbf{x})$ by applying models fitted on the other half to it. We can assess the model fitting by examining the correlation between the two $\tau(\mathbf{x})$s.

The rational behind this statistical test for the presence of heterogeneity is that we expect the ITE to vary randomly around a constant in the absence of heterogeneous treatment effects that can be explained by covariates, and hence the correlation between the two estimated τ values should be close to 0; otherwise, the stability and generality of ITE models can be examined by checking the replicability on an independent dataset using split-half correlation.

In order to reduce random variations that may be introduced by a single split, we performed split-half correlation test with multiple splits on the data in practice and combined the results together. Standard Simes test [188] may be an options for the combination, since it is robust to positive dependency of p-values. Its alternative hypothesis (H1) is that at least one of the hypotheses is non-null (at least one out of n splits yields a positive significant correlation), so it may be relatively loose for our problem. To increase stringency, we introduced a partial conjunction test (partial Simes test) based on the work from Benjamini et al. [21], whose H1 assumes that at

least r out of the n splits yield a positive significant correlation.

The threshold r defines the stringency/level of consistency for a finding that deserves further study. In experiment we set r to be 10% of n, which can give adequate type I error control, and employed 3 correlation measures (Pearson, Spearman and Kendall) to evaluate the split-half correlation.

A new permutation framework We proposed a novel permutation statistical test to assess the presence of heterogeneity. That is, it can be used to assess whether the predicted ITE are significantly better than predictions assuming a constant TE (which is the norm in most studies).

The objective of CF is to maximize $\widehat{\text{Var}}(\tau)$, as stated in section 4.3.2. When there is no heterogeneity that can be explained by covariates, $\widehat{\text{Var}}(\tau)$ should be low and close to 0. Equivalently, in this situation $\widehat{\text{Var}}(\tau)$ is roughly equal to $\widehat{\text{Var}}(\tau)$ yielded by model fitted on arbitrary covariates. Thus, a permutation approach we proposed is based on this rational to test the significance of $\widehat{\text{Var}}(\tau)$ observed.

To model the null hypothesis we shuffled the covariates for each permutation, such that there is no heterogeneity that can be explained by covariates, and then computed the $\widehat{\text{Var}}_{\text{observed}}$ for the permuted data.

If we repeat this process N times, then a probability for the null hypothesis of $\widehat{\text{Var}}(\tau)$ statistical test is defined as

$$\Pr(\text{null}_{\text{var}}) = \frac{\#(\widehat{\text{Var}}_{\text{perm}}(\hat{\tau}) \geq \widehat{\text{Var}}_{\text{observed}}(\hat{\tau})}{N}, \tag{4.11}$$

A related but clinically relevant question is: does the model that allows ITE outperform the 'standard' model predicting a constant ITE? That is, whether patients benefit more with the introduction of individualized treatments than a conventional treatments with a consideration of averaged treatment effects only. In ordinary regression problem, the goodness-of-fit of model is assessed by the mean squared error (MSE) between the expected outcome and predictions, so it's preferred to compute

the mean squared error (MSE) between the true and estimated τ for model assessment. If ITE model has a lower MSE than a constant model, which assumes that the ITE is the same for every subject, then ITE model outperforms the constant one. In reality, the true $\tau(x)$ cannot be directly observed, but Nie et al. [155] proposed that the MSE between the true and estimated $\tau(\mathbf{x})$, or $\tau_{\text{risk}}(\mathbf{x})$ can be defined as

$$\begin{aligned}\hat{\tau}_{\text{risk}} &= \sum_i \left((y_i - \hat{y}(\mathbf{x}_i) - \big(w_i - \hat{w}(\mathbf{x}_i)\big)\,\hat{\tau}(\mathbf{x}_i) \right)^2 \\ &= \sum_i \left(\tilde{y}_i - \tilde{w}_i \hat{\tau}(\mathbf{x}_i) \right)^2.\end{aligned} \tag{4.12}$$

Here $\hat{y}(\mathbf{x}_i)$ is an estimate of $\mathbb{E}(y_i|\mathbf{X} = \mathbf{x}_i)$ and $\hat{w}(\mathbf{x}_i)$ is an estimate of $\Pr(w_i = 1 \,|\, \mathbf{X} = \mathbf{x}_i)$ by any SML models. For simplicity, $\tilde{y}_i$ stands for $y_i - \hat{y}(\mathbf{x}_i)$, and $\tilde{w}_i(\mathbf{x}_i)$ for $w_i - \hat{w}(\mathbf{x}_i)$. The two terms can be regarded as residualized outcome and treatment respectively. These quantities are out-of-bag estimations. We can then compute the τ_{risk} assuming an unrestricted $\hat{\tau}(\mathbf{x})$ estimated from an ITE model and a constant effect based on the **average treatment effect (ATE)** $\bar{\tau}(\mathbf{x})$. We proposed the following definition to assess the improvement in τ_{risk} due to the incorporation of ITE versus a constant treatment effect

$$\hat{\tau}_{\text{improve}} = \sum_i (\tilde{y} - \tilde{w}\tau(\hat{x}_i))^2 - \sum_i (\tilde{y} - \tilde{w}\tau(\bar{x}_i))^2. \tag{4.13}$$

To model the null distribution of $\hat{\tau}_{\text{improve}}$, we propose a permutation approach in which permuted $\hat{\tau}_{\text{improve}}$s are obtained by shuffling the covariates for a predefined number of times. The null hypothesis of above test is that there is no statistical difference between the observed $\hat{\tau}_{\text{improve}}$ and permuted $\hat{\tau}_{\text{improve}}$s. Note that the test does not require any distributional assumptions.

An adaptive permutation strategy with early stopping was adopted to reduce the computing time. Permutations will be stopped earlier if the result is unlikely to be significant in future runs. Specifically, we will calculate a 99% CI for the permutation

p-value after each $k(k \ll N)$ runs. Here N is the total number permutations. If the lower CI > 0.05, the permutation will be terminated early.

4.4.2 Modeling Survival Outcomes

Time-to-event data are common in biomedical research, and standard ITE estimation methods may not work on survival data due to censoring or lost to follow-up. Usually some subjects have not experienced the event at the end of follow-up, that is, their records of survival time are unavailable (right censored), so the actual survival time for them is unknown. We proposed a flexible approach that can incorporate survival data into GRF and any other ITE models, an approach based on weighted 'mean imputation'.

Given a subject, denote its actual survival time as T_i and its censor time as c_i. In reality we observe y_i which is $\min(T_i, c_i)$ due to censoring. Let $t_{(1)} < t_{(2)} < \cdots < t_{(j)}$ be the censored survival times in ascending order, and $\hat{K}$ be the Kaplan-Meier (KP) estimator function of survival. Given $T_i > c_i$ for subject i, its log of censored survival times can be estimated by

$$\log(y_i^*) = \sum_{t(j) > T_i^c} \log t(j) \frac{\Delta \hat{T}(t(j))}{\hat{T}(T_i^c)}, \tag{4.14}$$

where $\Delta \hat{T}(t(j))$ refers to to the jump size of $\hat{K}$ at $t_{(j)}$ [46]. Under the assumption of the log-normal distribution of survival time, the imputed survival times can be included in ITE estimation.

4.5 Experiment Results

4.5.1 Simulations Studies

In design of simulations to evaluate the performance of our ITE framework and to compare the power and type I error rate of our proposed statistical tests with that of

`test_calibration` provided in the R package `grf`, we adopted similar strategies in the generation of synthetic data as introduced in [171]. Here we assume the log survival time is normally distributed without loss of generality. We considered the following six elements in simulations design:

1. *Sample size* The number of observations in the data is n, and p is the number of covariates.

2. *Distributions of covariates* Across all simulation scenarios, we drawn samples from standard normal distribution for features with odd column number, and for features with even column number we sampled from a Bernoulli distribution with $p = 1/2$. For simplicity, we denotes the distribution for the policy as D_x.

3. *Key functions* we denote propensity function for observations receiving a treatment $\pi(\cdot)$, average treatment effect $\mu(\cdot)$, and treatment effect $\tau(\cdot)$. The conditional mean effect for treatments and controls can be designed to be $\mu_1(\cdot) = \mu(\cdot) + \tau(\cdot)/2$ and $\mu_0(\cdot) = \mu(\cdot) - \tau(\cdot)/2$ respectively.

4. *Generation of survival time* Under the log-normal distribution of survival time, the survival time of observation can be generated by taking the natural exponentiation of the mean effect using $\exp(\cdot)$.

5. *Censor* We considered a censor rate of roughly 20% for all our scenarios. Given the uncensored simulated survival time, we found the cutoff r for the 80% quantile, and then generated censor time $T(\cdot)$ using the exponential distribution with rate parameter $1/r$. If the simulated $\log Y_i > T(\mathbf{x}_i)$, then $T(\mathbf{x}_i)$ should be used as outcome; otherwise $\log Y_i$ was used.

6. *Noise levels* The noise level $\sigma^2_{\log(Y)}$ was introduced in the generation of log survival time $\log Y_i$, which sampled from a normal distribution with mean $\mu(\cdot) + (w - 1/2)\tau(\cdot)$ and variance $\sigma^2_{\log(Y)}$, where $w \sim \text{Bernoulli}(\pi(\cdot))$.

Given the above predefined components, our data generation for observations i is modeled as

$$\begin{aligned}
\mathbf{x}_i &\sim D_x, \\
W_i &\sim \text{Bernoulli}\big(\pi(\mathbf{x}_i)\big), \\
\log Y_i &\sim \text{Normal}\big(\mu(\mathbf{x}_i) + (W_i - 1/2)\tau(\mathbf{x}_i), \sigma^2_{\log(Y)}\big), \\
T_i &\sim \text{Exp}(1/r), \\
c_i &= \begin{cases} 1, & \text{if } \log Y_i \leq T_i \\ 0, & \text{otherwise} \end{cases}, \\
Y_i &= \exp(\min(\log Y_i, T_i)),
\end{aligned}$$

where r is a cutoff corresponding to a specific quantile of $\log Y_i$, and $\pi(\mathbf{x}_i)$ is defined as

$$\pi(\mathbf{x}_i) = \frac{\exp(\mu(\mathbf{x}_i) - \tau(\mathbf{x})/2)}{1 + \exp(\mu(\mathbf{x}_i) - \tau(\mathbf{x})/2)} \tag{4.15}$$

for observational studies; for randomized studies $\pi(\mathbf{x}_i) = 1/2$ for all $\mathbf{x}_i$, which is the same as defined in [171]. In practice, the value for 0.8 quantile was used. c_i is an indicator for censor. If there is an event ($\log Y_i \leq T_i$) for subject i, then c_i is 1; otherwise 0.

We used functions with minor changes from [171] for propensity probability of receiving a treatment $\pi(\cdot)$, average treatment effect $\mu(\cdot)$, and treatment effect $\tau(\cdot)$. Within the simulation experiments, both randomized and observational studies are included, and 8 different functions of mean and treatment effects are made here to represent a combination of univariate/multivariate, additive/ interactive, and linear

and piecewise constant relationships. They are defined as follows:

$$
\begin{aligned}
&f_1(x) = 0, f_2(x) = 5\mathbb{I}(x_1 > 1) - 5 * pnorm(-1), f_3(x) = 5x_1, \\
&f_4(x) = x_2x_4x_6 + 2x_2x_4(1-x_6) + 3x_2(1-x_4)x_6 \\
&\quad + 4x_2(1-x_4)(1-x_6) + 5(1-x_2)x_4x_6 + 6(1-x_2)x_4(1-x_6) \\
&\quad + 7(1-x_2)(1-x_4)x_6 + 8(1-x_2)(1-x_4)(1-x_6) - 4.5, \\
&f_5(x) = x_1 + x_3 + x_5 + x_7 + x_8 + x_9, \\
&f_6(x) = 4\mathbb{I}(x_1 > 1)\mathbb{I}(x_3 > 0) + 4\mathbb{I}(x_5 > 1)\mathbb{I}(x_7 > 0) + 2x_8x_9 - 4 * pnorm(-1), \\
&f_7(x) = \frac{1}{\sqrt{2}}(x_1^3 + x_2 + x_3^3 + x_4 + x_5^3 + x_6 + x_7^3 + x_8 + x_9^3 - 7) \\
&f_8(x) = \frac{1}{2}(f_4(x) + f_5(x)),
\end{aligned}
$$

where $pnorm(x)$ is a function that calculates the cumulative distribution probability $F(x) = \Pr(X \leq x)$. Here X is with standard normal distribution.

Table 4.1: Specifications for simulation scenarios

	Scenarios							
	1,9	2,10	3,11	4,12	5,13	6,14	7,15	8,16
n	300	300	200	600	400	300	450	700
p	400	400	300	300	200	200	100	100
$\mu(x)$	$f_8(x)$	$f_5(x)$	$f_4(x)$	$f_7(x)$	$f_3(x)$	$f_1(x)$	$f_2(x)$	$f_6(x)$
$\tau(x)$	$f_1(x)$	$f_2(x)$	$f_3(x)$	$f_4(x)$	$f_5(x)$	$f_6(x)$	$f_3(x)$	$f_8(x)$
$\sigma^2_{\log Y}$	1	1/4	1	1/4	1	1	4	4

The 8 functions listed are centered and scaled to have mean close to 0 and roughly the same variance. Table 4.1 gives the specifications for simulation scenarios, including sample size n, number of features p, functions for mean and treatment effect $\mu(\cdot)$ and $\tau(\cdot)$, and variance of noise $\sigma^2_{\log Y}$. Specifications for sample size have been adjusted to accommodate our simulated survival data, compared with those in study [171]. Scenarios of odd number are randomized experiments, with $\pi(\mathbf{x}_i) = 1/2$ for all $\mathbf{x}_i$, but scenarios of even number are observational studies, in which $\pi(\mathbf{x}_i)$ is de-

fined by equation 4.15 for each subject $\mathbf{x}_i$. Note there is no heterogeneity in treatment effects in scenario 1 and 9.

Table 4.2: Comparison of power/type I error rate of different tests for the presence of heterogeneity

Scenarios	SHC-P	SHC-K	SHC-S	TC	$\widehat{\text{Var}}_\tau$	τ_{improve}
1*	0.002	0.002	0.002	0.002	0.058	0.054
2	0.266	0.268	0.26	0.896	0.976	0.98
3	1	1	1	1	1	1
4	0.656	0.652	0.652	0.782	0.924	0.924
5	0.548	0.548	0.546	0.752	0.926	0.934
6	0.888	0.88	0.882	0.944	0.992	0.992
7	0.862	0.844	0.85	0.886	0.956	0.974
8	0.752	0.732	0.73	0.42	0.592	0.664
9*	0.004	0.004	0.004	0.002	0.054	0.068
10	0.162	0.164	0.164	0.652	0.856	0.88
11	0.672	0.666	0.672	1	1	1
12	0.904	0.908	0.908	0.92	0.968	0.986
13	0.636	0.628	0.632	0.742	0.924	0.934
14	0.6	0.6	0.598	0.74	0.938	0.95
15	0.784	0.774	0.774	0.604	0.85	0.87
16	0.99	0.986	0.988	0.956	0.978	0.988

1. SHC-P stands for split-half correlation with pearson method, SHC-K for split-half correlation with kendall method, and SHC-S for split-half correlation with kendall method method for correlation testing.
2. TC represents `test_calibration` from the R package `GRF`.
3. $\widehat{\text{Var}}_\tau$ and τ_{improve} stand for two statistical tests for the presence of heterogeneity using them, proposed in section 4.4.1.
4. Scenario 1 and 9 are marked with stars, since they are two scenarios without heterogeneity.

To explore the reliability of our developed statistical tests and compare them with the `test_calibration` (TC) from GRF, we repeated application of our ITE framework with different random seed to the above simulation scenarios 500 times and examined the fitting of our approach with the statistical tests and test_calibration for each run. We calculated the proportion of repeats with p-values ≤ 0.05 for every statistical test. Because of no heterogeneity in scenarios 1 and 9, the proportion of tests returning significant findings (p ¡0.05) in scenarios 1 and 9 represents the type I

error rates. The same proportion represent the power of statistical methods for other simulation scenarios. Simulation results are shown in table 4.2.

The statistical tests $\widehat{\text{Var}}_\tau$ and τ_{improve} and TC showed good validity in our simulations. They all had strong power in capturing the presence of heterogeneity, and notably our methods with $\widehat{\text{Var}}_\tau$ and τ_{improve} completely dominated TC provided in package GRF across all simulation scenarios except 1 and 9. In scenarios 1 and 9 they all showed relatively low type I error rate, with roughly 0.05 for statistical tests using $\widehat{\text{Var}}_\tau$ and τ_{improve} and 0.002 for TC. Type I error rates for our methods were also within an acceptable range, and TC had lower type I error rate than our methods. However, our methods with $\widehat{\text{Var}}_\tau$ and τ_{improve} were more powerful in detecting the presence of heterogeneity.

Split-half correlation (SHC) approaches with three different correlation evaluation methods showed good type I error rates in scenario 1 and 9. However, their powers varied greatly across all other scenarios. By investigating their performance in scenario 6, 7, 12, and 16, we found that they also can maintain good performance if there is strong heterogeneity. Surprisingly, they apparently outperformed the other three methods in Scenario 8. In short, we can still consider SHC approaches as good supplements to the other three methods.

4.5.2 Applications to Real Data

COVID-19 has become a major public health burden and the latest report by WHO showed that COVID-19 has infected over 15 million people and caused more 600,000 deaths throughout the world [49]. Most infected people experience mild to moderate symptoms and recover without special treatment, but a subgroup of patients develop severe complications, including acute respiratory distress syndrome (ARDS) [53]. Many clinical risk factors may contribute to higher severity of the disease, such as diabetes and other chronic illnesses. Nevertheless, there is substantial heterogeneity of outcomes among people with risk factors. Here we aim to identify clinical/genetic

risk factors that demonstrate effect heterogeneity (on severity of illness).

Table 4.3: Results for test of model fitting by SHC and permutation for selected clinical variables with genetic expression from blood

Clinical variables	P-values				
	Pearson	Kendall	Spearman	Perm var	Perm risk
Alcohol intake frequency	1.111E-05	8.412E-03	9.016E-03	**0.076**	**0.024**
Diabetes diagnosed	5.985E-32	6.818E-12	6.203E-12	**0.000**	**0.070**
Alcohol drinker status	5.680E-04	3.154E-03	3.489E-03	**0.000**	0.966
Age at recruitment	1.236E-05	6.283E-04	6.292E-04	**0.076**	**0.070**
Cholesterol	9.323E-09	1.585E-04	1.823E-04	**0.010**	0.472
HbA1c	1.131E-05	2.361E-02	2.238E-02	**0.008**	0.402
Ethnic background	3.672E-08	2.468E-04	1.813E-04	**0.000**	0.144
LDL direct	1.201E-05	5.942E-03	5.890E-03	**0.094**	0.678

1. Perm var stands for permutation variance test, and Perm risk for permutation τ-risk test.
2. P-values < 0.1 for Perm var and Perm risk are in bold.

We applied our ITE framework to GWAS data of patients with COVID-19. The data was extracted from UK Biobank, a large and long-term biobank study in the United Kingdom, which aims to investigate genetic and environmental determinants of diseases [25]. We considered clinical factors shown to affect COVID-19 susceptibility in Atkins et al.[13] as risk factors and covariates in our model. As effect heterogeneity of the risk factors may be contributed by the varied genetic background of subjects, we also considered gene expression profiles as covariates. Genetic expression of each subject was imputed by PrediXcan [68]. The imputation of gene expression by PrediXcan is tissue-specific; here we examined the imputed expression of blood and lung.

We considered COVID-19 positive patients only. The severity of disease was considered as the outcome; based on relatively limited data provided by the UK Biobank, we labeled 'severe' disease if a patient received inpatient treatment. A total of 1550 patients were include, among which 1023 required hospitalization and were labelled as 'severe'. The rest (N=527) were not hospitalized and were assumed to be having a milder disease. Since there were missing values for some clinical variables in the data, we imputed missing data using missForest [197] with default settings. The program

iteratively employs random forests to impute the missing values.

In practice, we considered one clinical variable as treatment, severity as outcome, and all other clinical and genetic factors as covariates for each run. If there were at least two out of three p-values < 0.05 from the split-half correlation (SHC) test, then permutation tests would also be carried out to further validate the ITE model. Imputed expression for lung and blood were modeled separately with clinical variables. Results for blood and lung are summarized in Table 4.3 and 4.4 respectively.

Table 4.4: Results for test of model fitting by SHC and permutation for selected clinical variables with genetic expression from lung

Clinical variables	**P-values**				
	Pearson	**Kendall**	**Spearman**	**Perm var**	**Perm risk**
Alcohol intake frequency	4.639E-08	2.127E-04	1.420E-04	1.000	**0.092**
Diabetes (type 2) diagnosed	3.736E-54	1.575E-17	2.063E-17	**0.000**	**0.000**
Age HBP diagnosed	5.351E-04	1.803E-02	1.757E-02	1.000	0.200
Alcohol drinker status	3.331E-06	1.251E-02	1.317E-02	**0.010**	**0.038**
Age at recruitment	4.973E-11	8.332E-05	1.005E-04	1.000	0.200
GPC	1.966E-02	4.935E-02	5.539E-02	**0.050**	0.296
Cholesterol	1.298E-18	5.319E-06	5.150E-06	1.000	0.117
HbA1c	3.778E-18	2.461E-07	1.778E-07	**0.044**	**0.020**
HDL cholesterol	3.136E-04	2.598E-02	2.456E-02	1.000	0.400
Ethnic background	1.651E-17	2.640E-08	1.776E-08	**0.000**	0.310
LDL direct	4.337E-18	9.382E-05	8.628E-05	1.000	**0.048**

1. Perm var stands for permutation variance test, and Perm risk for permutation τ-risk test.
2. P-values < 0.1 for Perm var and Perm risk are in bold.
3. HBP stands for high blood pressure, and GPC for Genetic principal components.

Clinically, HbA1c is a blood test that is used to diagnose and monitor glycemic control in patients with diabetes. Our results indicate that type 2 diabetes and HbA1c demonstrate effect heterogeneity on the severity of COVID-19 infection, with p-values of SHC $<$ 5E-17 on lung data and $<$ 1E-11 on blood data and of permutation test $<$ 0.001 on lung data and $<$ 0.07 on blood data. The results indicate that the risk conferred by diabetes or high HbA1c is likely different for different individuals. Previous studies have suggested diabetes or poor glycemic control as risk factors for COVID-19 or more severe disease [146, 55, 242], but our results also suggest that the risk

conferred can differ across subjects with the same risk factor, which may have implications in treatment and prevention. Similarly, our results also reveal cholesterol, HDL-cholesterol and LDL-cholesterol are possible risk factors with effect heterogeneity. These clinical variables are known risk factors for cardiometabolic diseases, and studies have demonstrated that cardiometabolic comorbidities are related to a more severe course of COVID-19 [196, 83, 184].

Interestingly, previous studies suggested that alcohol consumption is associated with the amount of ACE2 present in the body and particularly in lung [159, 204], which in turn may lead to increased susceptibility to infection. The present findings further suggested the effect of alcohol drinking may be heterogeneous across individuals. We note, however that the link between alcohol drinking and infection might be subject to confounding variables like sharing alcoholic drinks [150] and propensity to social gathering [8]. Our study also shows ethnic background and age may be risk factors demonstrating effect heterogeneity. A recent study has shown that black ethnicity are at higher risk and South Asians and other ethnicities have intermediate risks compared with the white, and that people aged 80+ years are more likely to have severe infection [13]. However, heterogeneity of the racial effects on people with different background is unknown.

4.6 Conclusion

In the chapter, we proposed a computational framework to estimate individualized treatment effects (ITE) of risk factors on patient outcome. In order to assess model fitting, we propose three different criteria to investigate the model fitting. In order to improve the applicability of ITE framework across different scenarios, we proposed an approach based on weighted "mean imputation" to incorporate survival times as outcomes. We carried out simulations to validate our framework on survival outcomes. Our method showed good power and valid type I error control across different

simulation settings. In real data analysis, we applied our approach to GWAS data of patients with COVID-19, and showed that certain risk factors (e.g. type 2 diabetes) is associated with heterogeneous effect across different subjects.

Our approach may have following limitations. In practice, we found that split-half correlation method has better power than permutation-based methods. We conjecture that this is due to the characteristics of GWAS data, in which each covariate has very small effect and lots of covariates contribute to the outcome. Further work will include a larger panel of simulations with different distribution of effects. The application to COVID-19 is a preliminary study although the results were interesting. Further work is required to further characterize the variables interacting with the risk factors, which may contribute to the heterogeneity. The current sample size is relatively small and power to detect ITE may not be sufficient. Hospitalization is a rough proxy for severity and more detailed clinical data will enable better delineation of outcomes. Finally, we expect some unknown confounders may be missed so the estimated effect cannot be considered as entirely a 'causal' effect. Potential extension also include modeling survival among infected patients in the UK Biobank data.

□ **End of chapter.**

Chapter 5

Conclusions

The number of drugs approved in the last few decades is disproportionate to the increasing amount of investment. This indicates that traditional approaches to drug discovery has not been as successful as anticipated. There is an urgent need of new therapies, especially in some areas such as psychiatry. Moreover, drugs of novel mechanism of actions are decreasing in number, suggesting limitations of the traditional drug development approach. On the other hand, the past few years have seen an extremely rapid development in ML methods and applications. Hence computational approaches based on ML methods may be utilized to address this issue, given the wide availability of omics data.

In chapter 2 we have presented and applied a machine learning to drug repositioning for schizophrenia and depression/anxiety disorders. We found the candidates were enriched for psychiatric drugs considered in clinical trials, and that numerous top hits were supported by previous studies. A systematic literature support analysis showed that the number of article supporting the association between the drug and disease is significantly correlated to predicted treatment probabilities from ML models. The list of repositioning candidates might serve as a useful resource for researchers and clinicians working on schizophrenia as well as depression and anxiety disorders, which are illnesses very much in need of new therapies. In addition, this

study may shed light on molecular mechanism of actions of drugs by examining the variable importance provided by ML methods. On the other hand, our approach still has room for further improvements. In the regard of imbalanced data, more advanced approaches may be employed, such as down-sampling on the class of majority to balance the class size. Meanwhile, increasing the sample size is another direction for further improvement.

However, drug repositioning may not be always be available in practice. Identifying promising drug target is another direction to hasten drug development, but traditional approach for drug target discovery suffers from high failure rate. In reality, it is impractical to perform in-depth experimental studies on every possible target for each disease. Computational methods offer a cheap, fast and systematic high-throughput approach to guide prioritization of drug targets. Moreover, we have witnessed a rise in the number of studies using computational approaches to discover potential drug targets for further investigation.

In chapter 3, we presented a general computational framework to prioritize drug targets for various diseases. Under the framework, different kinds of ML methods can be utilized. We applied four ML methods to identify potential drug target of five disorders. External validation shows that the top candidates were enriched for targets selected by independent lines of evidence from a large external database (Open Targets). Some top target genes were also supported by previous studies. We hope our presented framework can provide an additional way to prioritize drug targets for development, which is independent of and may be combined with other existing sources of data. In addition, a direction for further improvement is that high quality of expression data will dramatically improve the validity of our result, since in practice part of our result is interfered by potential off-target effects in knock-down experiments.

In this thesis, we also explored the estimation of individualized treatment effects. It has been widely acknowledged that patients response differently to the same treat-

ment. This heterogeneity may be caused by their different clinical and/or genetic backgrounds. Advances in this area can directly benefit disease treatment at an individual level. Here we have employed forest-based machine learning methods to estimate ITE. Such methods can capture high-order interactions commonly present in biomedical data.

In the chapter 4, we proposed a computational framework to estimate ITE of risk factors/treatments on patient outcome. In order to assess model fitting, we proposed different statistical methods to investigate the model fitting. In order to increase the generality of ITE framework, we proposed a weighted "mean imputation" approach to incorporate survival times as outcomes. We carried out simulations to validate our framework on survival outcomes. We also illustrated the usefulness of this approach by application in real clinical datasets.

In this thesis, we studied translational bioinformatic approaches ranging from drug development to estimation of personalized treatment effect. We believe this work will provide a new angle to explore address some key issue in these fields, with the hope to benefit patient care ultimately.

□ **End of chapter.**

Appendix A

Proof of Propositions

□ **End of chapter.**