Journal of the American Statistical Association

Taylor &Francis
Til)'lof&frM'lCi C,oup

Journal of the American Statistical Association

ISSN: (Print) (Online)Journal homepage: www.tandfonline.com/journals/uasa20

Operator-Induced Structural Variable Selection for Identifying Materials Genes

Shengbin Ye, Thomas P. Senftle & Meng Li

To cite this article: Shengbin Ye, Thomas P. Senftle & Meng Li (2024) Operator-Induced Structural Variable Selection for Identifying Materials Genes,Journal of the American Statistical Association, 119:545, 81-94, DOI: 10.1080/01621459.2023.2294527

To link to this article: https://doi.org/10.1080/01621459.2023.2294527

JOURNAL OFTHE AMERICAN STATISTICALASSOCIATION
2024, VOL 119, NO.545,81-94: ApplicationsandCaseStudies
https://doi.org/10.1080/01621459.2023.2294527

Taylor & Francis
Taylor&Franci.sGroup

Operator-Induced Structural Variable Selection for Identifying Materials Genes

Shengbin Ye•e, Thomas P. Senftlebe, and Meng u•e

•Department of Statistics, Rice University, Houston, TX; bDepartment of Chemical and Biomolecular Engineering, Rice University, Houston, TX

ABSTRACT

In the emerging field of materials informatics, a fundamental task is to identify physicochemically meaningful descriptors, or materials genes, which are engineered from primary features and a set of elementary algebraic operators through compositions. Standard practice directly analyzes the high-dimensional candidate predictor space in a linear model; statistical analyses are then substantially hampered by the daunting challenge posed by the astronomically large number of correlated predictors with limited sample size. We formulate this problem as variable selection with operator-induced structure (015) and propose a new method to achieve unconventional dimension reduction by using the geometry embedded in OIS. Although the model remains linear, we iterate nonparametric variable selection for effective dimension reduction. This enables variable selection based on ab initio primary features, leading to a method that is orders of magnitude faster than existing methods, with improved accuracy. To select the nonparametric module, we discuss a desired performance criterion that isuniquely induced by variable selection with OIS; in particular, we propose to employ a Bayesian Additive Regression Trees (BART)-based variable selection method. Numerical studies show superiority of the proposed method, which continues to exhibit robust performance when the input dimension is out of reach of existing methods. Our analysis of single-atom catalysis identifies physical descriptors that explain the binding energy of metal-support pairs with high explanatory power, leading to interpretable insights to guide the prevention of a notorious problem called sintering and aid catalysisdesign. Supplementary materials for thisarticle are available on line.

ARTICLE HISTORY
Received November 2021
Accepted November 2023

KEYWORDS
BART; Bayesian nonparametrics; Feature engineering; Materials genomes; Nonparametric dimension reduction

1. Introduction

The Materials Genome Initiative set up by the White House is a large-scale effort concerning the utilization of computational tools to accelerate the pace of discovery and deployment of advanced material systems. Since its inception in 2011, there has been a surge of interest in data-driven materials design and understanding (Zhong et al. 2020; Hart et al. 2021; Keith et al. 2021; Lin et al. 2021; Liu et al. 2021). In this nascent area called materials informatics, computational methods that account for physical and chemical mechanisms of a material system play a central role in aiding, augmenting, or even replacing the time-consuming trial and error experimentation. A fundamental task is to identify physicochemically meaningful *descriptors,* or *materials genes* (Ghiringhelli et al. 2015; Poppa et al. 2021). These descriptors, for example, are key to modeling single-atom catalysis and finding or developing more efficient catalytically active materials. In statistical terms, descriptors are high-dimensional predictors but with strong structure in that they are functional transformations of a set of primary features $X = (x_1,\ldots,x_p)$. For instance, a simple example of descriptors is $f(X) = \{\exp(x_1) - \exp(x_2)\}^2$, which can be constructed using exponential and squared functions in combination with subtraction.

Suppose the response vector y measures the material property of interest, and the primary features matrix $X = (x_1,\ldots,x_p)$ collects physical or chemical properties of the materials such as atomic radii, ionization energies, etc. Then the space of engineered predictors (or descriptors) up to order M is $(\cdot)(M)(X)$, which consists of nonlinear predictors with explicit functional form resulting from M-order compositions of operators $(\cdot)$ on X:

$$o<M)(X) = (\cdot)_0\ o<M-l)(X) = \underbrace{(\cdot)_0 \cdots_0 C'.J(X)}_{\text{Mtimes}}.$$

For example, some commonly used operators in materials genome are

$$(\cdot) = \{+, -, x, /, 1\text{-}1, l, \exp, \log, |\cdot|, \ ,-^{1}, ^{2}, \sin(rr\bullet), \cos(rr\bullet)\}, \quad (1)$$

andtheaforementioneddescriptorf(X) $= \{\exp(x_1)\text{-}\exp(x_{2)}\}^2$ belongs to $o<^3>(X)$. We refer to this distinctive geometry encoded in $(')(M)(X)$ as *operator-induced structure* {01S). The aforementioned descriptors in materials genome are thus the predictors in a linear model with 01S. Henceforth, we will use *descriptor selection* and variable selection in the presence of 01S interchangeably. Note that the specification of $(\cdot)$ depends on domain knowledge, and we intentionally include the absolute difference operator $|-|$ in $(\cdot)$ because it is directly interpretable in materials science, and it often provides clear intuition on many physical phenomena, such as the metal-oxide binding

CONTACT MengLie meng@rice.eduQ Department of Statistics, Rice University, Houston, TX 77005.

4) Supplementary materials for thisarticle areavailable online.Please go to www.tandfonline.com/r/JA5A.

energy (Liu et al. 2022). Treating it as a single operator reduces the required number of iterations to generate related descriptors.

A common practice in materials genome (O'Connor et al. 2018; Ouyang et al. 2018; Liu et al. 2020) is to employ modern statistical variable selection developed for linear models. However, the geometry of OIS defined by operators O and high-order compositions induces high correlation and ultra-high dimension to the feature space. As detailed in Section 2, the dimension of o<M)(X) increases *double exponentially* with Mand the number of binary operators in *0.* For example, with $p = 59$ in our real data application, enumerating o<^{3}\X) gives 1.01 x 10^{17} predictors while only a handful of them are associated with the response. Moreover, these predictors are highly correlated as a result of iteratively applying unary operators. This along with a small size such as $n = 91$ in our real data application substantially hurdles the performance of existing methods that rely on linear variable selection methods. Indeed, materials genomes are an analog concept to genomes, but the dimension of predictors and inherent strongcorrelation in materials genome-wide association studies, or *materials GWAS,*pose unprecedented challenges to statistical analysis.

In this article, we aim to develop a powerful method for materials GWAS in which we effectively identify materials genes that are associated with the response of interest. To achieve dimension reduction in materials GWAS, we consider an iterativeapproach by applying a small set of operators and immediatelyidentifying the relevant descriptors,*'D,* before constructing more complex descriptors. This step is iterated in light of the composition structure in OIS, in striking contrast to existing literature that aims to exhaustively generate o<M)(X). In each iteration, O('D) is typically substantially smaller than o<M)(X), and suchsparsity achieves dimension reduction and tackles the daunting computational challenges posed by materials GWAS.

Iterative dimension reduction, however, faces two intertwined challenges. First, the constructed descriptors in intermediate steps, unlike the astronomically large$_{Q(M)}$ (X),arenot necessarily linearly associated with the response. To address this, we propose to use *nonparametric variable selection* for dimension reduction to ensure selection accuracy under the geometry of OIS. That is, while the model is assumed to be linear, we employ nonparametric variable selection to achieve dimension reduction while maintaining highselection accuracy. We refer to this key novelty of our proposed method as "parametrics assisted by nonparametrics': or PAN.

The second challenge pertains to the selection of the nonparametric module. Unlike traditional nonparametric variable selection, OIS variable selection calls for new performance criteria for the nonparametric module to ensure OIS selection accuracy (see Section 2.3 for more details). We introduce a *PAN criterion* to reassess nonparametric selection methods, which elucidates an asymmetric effect between false positives and false negatives and highlights a desired invariance property to unary transformations. In particular, we propose to use a Bayesian additive regression tree variable selection method, BART-G.SE (Bleich et al. 2014), as the nonparametric module, which we show is well suited to satisfy the PAN criterion.

Coupling the PAN strategy with BART-G.SE, together with additional considerations to address the complexities of materials GWAS,leads to a new method for materials GWAS, which we call iterative BART, or iBART. The iterative framework of iBART reduces the size of the effective descriptor space significantly, mitigating collinearity in the process, and the use of nonparametric variable selection accounts for structural model misspecification in intermediate variable selection steps. Our extensive experiments show that iBART gives excellent performance with accuracy and scalability that are not seen in existing methods. Note that iBART is not a new BART variant for nonparametric regression, but rather an iterative use of BART within the PAN framework specifically tailored for materials GWAS.

The outline of the article is as follows. Section 1.1 reviews related work in materials genome. In Section 2, we introduce the OIS framework, describe the PAN selection procedure, and discuss how to choose the nonparametric module in PAN and somepracticalconsiderations regarding PAN.Section 3 contains a simulation study that shows superior performance of iBART relative to existing methods. In Section 4, we apply iBART to a single-atom catalysis dataset and it identifies physical descriptors that explain the binding energy of metal-support pairs with high explanatory power, leading to interpretable insights to guide the prevention of a notorious problem called sintering. We closein Section 5 with a discussion. All proofs, detailsof variants of iBART, and additional simulation results and discussion are deferred to the supplementary material.

1.1. Related Work

Descriptor selection has attracted growingattention in materials science. Recent methods often build on a *one-shot* descriptor generation and selection scheme followed by modern statistical variable selection approaches (O'Connor et al. 2018; Ouyang et al. 2018; Liu et al. 2020). In particular, they first construct descriptors by applying of erators iteratively M times on the primary feature space $\mathbf{X}^{<0} = \boldsymbol{X} = (xi, \ldots, Xp) \in R.nxp$ to construct an ultra-high dimensional descriptor space x<M) = $_{Q(M)}$ (X) of O(p^{2}M) descriptors, assuming binary operators are used in each iteration. Then variants of generic statistical methods areadopted to select variables fromx<M). Alongthis line, the method SISSO (Sure Independence Screening and Sparsifying Operator) proposed by Ouyang et al. (2018) builds on Sure Independence Screening, or SIS (Fan and Lv 2008), O'Connor et al. (2018) uses LASSO (Tibshirani 1996), and Liu et al. (2020) adopts Bayesian variable selection methods.

SISSO iswidely perceived as one of the mostpopular methods for materials genome. It uses SIS to screen out P descriptors, from which the single best descriptor is selected using an *lo*-penalized regression. If a total of k descriptors are desired, this process will be iterated for k times yielding k sets of SIS-selected descriptors, followed by an lo-penalized regression to select the best k descriptors from all the SIS-selected descriptors. Note that in each iteration, SIS is employed to screen out P descriptors from the remaining descriptor set with an updated response vector given by its least squares residuals projected onto the space spanned by previously SIS-selected descriptors. Users must define the composition complexity of the descriptors through $M,$ that is, the order of compositions of operators. In a typical application of SISSO, the composition complexity M is no greater than 3, the number of candidate descriptors in each SIS iteration is less than 100, and the number of descriptors k

is no larger than 5 (Ouyang et al. 2018). Note that selecting five descriptors with 100 SIS-selected descriptors in each iteration amounts to fitting at most $\binom{5}{0} \approx 2.6 \times 10^{11}$ different regressions, which is computationally intensive.

A major drawback of these one-shot descriptor construction procedures is the introduction of a highly correlated and ultra-high dimensional descriptor space *x<M)*. High correlation often hampers the performance of a variable selection method, and the ultra-high dimensional descriptor space with large *p,* as common in modern applications, can make it computationally prohibitive for such methods. In practice, these methods often resort to ad hoc adaptation or sacrifice the complexity level of candidate descriptors.

Another thread of work is the well-developed *automatic feature engineering* in machine learning, aiming to generate complex features from given constructor functions adaptively (Markovitch and Rosenstein 2002; Feurer et al. 2015; Khurana, Samulowitz, and Turaga 2018). However, the overwhelming focus of this literature is on increasing the predictive power of the primary features *X.* We instead focus on discovering the underlying functional relationship between the response and the predictors and revealing data-driven insights into the underlying physics of materials design. In addition, the sample size in materials genome is typically limited, hampering the use of machine learning methods that rely on large training data.

In statistics, transformations have been commonly used to expand the predictor space, including polynomials, logarithmic, power transformations, and the previously noted interactions. The induced feature spaces from these elementary transformations are often overly simple to capture the intricate dynamics of the response in materials genome, particularly compared to high-order compositions of a larger operator set. There has been a rich literature on nonparametric variable selection, but descriptor selection relies on a linear model with feature engineering that favorably points to interpretable insights for domain experts as the functional forms of selected variables are explicitly given and the feature space could be composed usingdomain-relatedknowledge. The nonparametric module in PAN only serves as a dimension reduction tool, and the desired performance calls for new investigation under the context of OIS. Overall, materials GWAS may play an analogous role that GWAS have played in motivating new statistical methods and concepts, andto the best of our knowledge, the present article is the first statistical work on this topic.

2. The 01S Framework

2.1. Operator-Induced Structural Model

We begin with a standard ultra-high dimensional linear regression model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p^*} x_{p^*} + \varepsilon, \tag{2}$$

wheres $\sim Nn(0,c;^2\mathrm{I})$ isa Gaussian noise vector,and the regression coefficients *f3* aresparse. The dimension of predictors p^* is ultra-high, at the materials GWAS scale that typically exceeds the maximum size of matrices allowed by a modern personal computer, while the sample size n is on the order of tens. In this article, we assume that this seemingly ultra-dimensional descriptor space obeys an operator-induced structure (OIS). In particular, we assume that the predictors *x1, ... , Xp•* in (2), or *descriptors,* are generated by applying operators in O iteratively M times on a primary feature space $\boldsymbol{X} = (\mathrm{xi}, \ldots, Xp) \in \mathrm{]Rnxp,}$

$$(\mathrm{x1},\ldots, Xp\bullet) = \mathrm{X(M)} = o(M)(\mathbf{X}) = \underbrace{O \circ \cdots \circ O}_{\text{Mtimes}}(\mathrm{X}) = O \circ \mathrm{o(M\text{-}I)(X)}, \tag{3}$$

where *x<M) = o<M)(***X***)* denotes M-composition of O on X and can be defined iteratively as above. The operator set O is user-defined; for concreteness, we focus on the common example given in (1), unless stated otherwise.

We adopt the following convention. Evaluation of operators on vectors is defined to be entry-wise, for example, $\mathrm{Xi} = (\mathrm{Xi\_p} \ldots, x\,.1)T$ and $X1 + X2 = (\mathrm{X1,1} + X1,2, \bullet\ldots, Xn,I + Xn,2)T$. Throughout this article, we assume that all descriptors in *x<M)* are uniquely defined in terms of their numerical values. For instance, onlyone of the descriptors in $\{\mathrm{xi}, x_I \times xii\}$ will be kept in *x<M).* This can be easily achieved in practice by identifying and removing perfectly correlated descriptors.

We hereafter refer to the linear regression model in (2) along with the operator-induced structure (OIS) in (3) as the OIS model To facilitate a precise OIS model definition using predictors with nonzero coefficients, we define *M-composition descriptor* as follows.

Definition 2.1 (M-composition descriptor). We defineJ<M)(X) to be an M-composition descriptor if it is constructed via M compositions of operators on some primary features **X:** *J<M)*(X) = o<M)(X) = *OMO*t<M-l)(X) = *OMO* OM-I $\circ \cdots \circ$ *o,* (X), where $Om \in O$ is the mth composition operator(s) for 1 ::=; m ::=; *M,* andf(l)(X),... ,J<M-l)(X) are the necessary intermediate descriptors for constructing the descriptor*J<M)*(X).

Note that if the mth composition operator is a binary operator, there mayexist two (m- l)th composition operators but we suppress the notation in the definition above for simplicity. Furthermore, ifan M-composition descriptor*J<M)*(X) only depends onasubsetofprimaryfeaturesXs,whereS s; [p] = {1,... ,p}, we also write it asJ<M)(Xs) and call it an (M,S)-descriptor.

Definition 2.2 ((M,S)-composition OIS model). An (M,S)-OIS model assumes

$$Y = f3o + \sum_{k=l}^{K} L,fiMkl(Xsk)fh + s, \tag{4}$$

where $M = \max_{k=1,\ldots,K} Mk$ denotes the highest order of operator compositions, K denotes the number of additive descriptors, *Xsk* is thesetof primaryfeatures used in the kthdescriptors, and $S = \mathrm{Uf=i}\, Sk$ is the set of all active primary feature indices.

Throughout the article, we assume the data follows an (M,S)-OIS model in (4). We use Mo for the oracle highest composition complexity and S_0 for the oracle set of active primary feature indices. Descriptors in the (M,S)-OIS model and their intermediate descriptors are called *active.* We next use

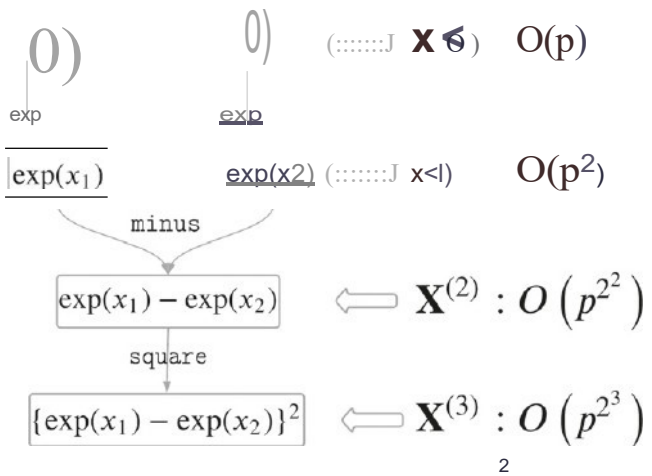

Figure 1. A tree diagram for generating {exp(x1) - exp(x2)} 2. The dimension of descriptor space increases double exponentially with the composition complexity.

a toy example to illustrate the introduced concepts in OIS and the challenges posed by descriptor selection. Suppose that the data-generating model is

$$y = \beta_0 + \beta_1 f_1^{(M_1)}(X_{\mathcal{S}_1}) + \beta_2 f_2^{(M_2)}(X_{\mathcal{S}_2}) + \varepsilon,$$

where M1 = 3,M2 = 2,S1 = {1,2},S2 = {3,4},j_1(Mt\X) = {exp(x1) - exp(x2)}2, and $J\}M^2\backslash X)$ = sin(rrx3x4). Here {exp(x1) - exp(x2)}2 is a 3-composition descriptor or a (3,{1, 2})-descriptor, andsin(rr x_{3x4}) isa 2-composition descriptor or a (2,{3, 4})-descriptor. Both descriptors arise from applying O iteratively three times on the primary features: $\mathbf{X^{<3}}) = O \circ O \circ O(X)$. The composition ofoperators resembles a tree-like structure for generating descriptors; Figure 1 describes the tree-like workflow for generating{exp(x1) - exp(x2)}^{2}.

Tosee how the descriptor space dimension increases with the number of iterations, let Cu and Cb denote the number of unary and binary operators, respectively, and let p_j denote the dimension of the jth descriptor space $\mathbf{X}m$. Note that the dimension of $\mathbf{X^{<1}}$) is $p_1 = cup + CbP(P - 1)/2$, which is on the order of O(p^2); the dimension of $\mathbf{X^{<2}}$) is P2 = $CuPI + CbP1(p1 - 1)/2$;:::: O(p^2•^{2}) = O(p^2\ the dimension of x(3) is p_3 = $CuP2 + CbP2<P2 - 1)/2$;:::: O(p$^{2^2}$•^{2}) = O(p^2\ The dimension of descriptor space will increase double exponentially with the number of binary operator compositions, for example, with M compositions of binary operators on X the resulting descriptor space has a dimension of order O(p^{2}M). Similarly, the double exponential expansion applies to the number of binary operators Cb. Excluding redundant descriptors does not prevent this doubleexponential expansion.Asshown in Section 3.4, building $\mathbf{X^{<2>}}$from $p = 59$ primary features results in an astronomical descriptor space containing over 5.5 x 10^7 = 0(59^{22}) descriptors evenafter removing redundant and nonphysical descriptors. In addition, the number of active (intermediate) descriptors will be nonincreasing with Mas shown in Figure 1: there are four active descriptors exp(x$_1$), exp(x2), x_3, x_4 in x(I), but only two active descriptors exp(x1) - exp(x2) and x_3x_4 in$\mathbf{X^{<2>}}$,and two active descriptors {exp(x$_1$) - exp(Xi)}^{2} and sin(rrx3x4) inX(3)_

2.2. PAN Descriptor Selection for 01S Model

We propose an iterative descriptor construction and selection procedure PAN for the OIS model, which generates descriptors by iteratively applying operators and selecting the potentially useful intermediate descriptors between each iteration of descriptors synthesis. The iterative descriptor selection procedure excludes irrelevant intermediate descriptors from the descriptor generating step, achieving a *progressive* variable selection and enabling variable selection based on ab initio primary features. This reduces the dimension of the subsequent descriptor space x<m) and mitigates collinearity among the descriptors in comparison to the one-shot descriptor construction approaches.

To describe the method in its most general form, we allow *different sets of operators Om* S O for each iteration m = 1,...,M, leading to the descriptor space

$$\mathcal{O}_M \circ \mathcal{O}_{M-1} \circ \cdots \circ \mathcal{O}_2 \circ \mathcal{O}_1(\boldsymbol{X}). \tag{5}$$

The framework ofour iterative descriptor selection procedure is as follows.

PAN descriptor selection procedure:

1. **Repeat** the following until at least one descriptor exhibits a stronglinear association with the response variable y (i = 0):
   (a) Use a nonparametric variable selection procedure to perform descriptor selection on x<Oand obtain the selected descriptors x(i)';
   (b) Apply the ith operator set O_i on all of the previously selected descriptors, Umx<m)', yielding a newdescriptor space, x(i+l) =O;(LJm x<m)'), where $O_i$can bedifferent for each iteration i;
2. Once there exist descriptor(s) that exhibit a strong linear association with response variable y, use a linear parametric variable selection procedure to perform descriptor selection on x<O, and obtain the selected descriptors, $\boldsymbol{X}^*$ S **x<i).**

We keep all the selected descriptors in the main loop to facilitate the creation of high-order complexity descriptors with the help oflow-order complexity descriptors. For instance, constructingxf using$O$ defined in (1) requires us to keep x_1 E **X<^{0}>'** selected at the base iteration and XI E x<l)' selected at the first iteration.

To see how this iterative procedure helps reduce the dimension of descriptor space significantly, let s_i = IX(i)'I be the number ofdescriptorsselected in the ithiteration andp_i = IX(i)I be the dimension of the ith descriptor space. Suppose that the number of selected descriptors is sparse ineach iteration, that is, s_i « p_i. Assuming binary operators were used, the dimension of the (i + l)th descriptor space in PAN is on the order of O(sf) « $0(pf)$, and this holds for all iteration i :::: 0. If we further assumes; ;:::: $0(p)$ for all i :::: 0, where p = IXI is the number of primaryfeatures, then the dimension of the (i+ l)th descriptor space for PAN is on the order of O(p^2), compared to O(p$^{2^{i+1}}$) for the one-shot methods. Note these assumptions are reasonable according to the discussion in Section 2.1. In the simulation studyin Section 3.4with p = 10 primary features, we observed thats; ;::::0(10^1) and p_i ;::::0(10^2) for all iterations of

the PAN procedure. On the contrary, a one-shot method, such as SISSO, generates a descriptor space containing 9.26×10^9 descriptors in thesame setting.

The use of a nonparametric variable selection procedure in Step l(a) is necessary because the intermediate descriptors may not have a strong linear association with the response variable. Thus, a method that accounts for model misspecification, such as a nonparametric method, is more suitable for preliminary screening of the intermediate descriptors. In addition, a suitable nonparametric module for PAN needs to account for the geometry embedded in OIS and the unconventional goal of selecting operators along with variables; the next section discusses performance requirements for this step and introduces an implementation of PAN, iBART, that is particularly suitable for OIS. In Step 2, the oracle descriptors are linearly associated with the response. Hence, we employ linear parametric variable selection methods, such as LASSO (Tibshirani 1996), to reduce false positives and select the final descriptors.

2.3. Choosing Nonparametric Module in PAN and iBART

In OIS, the ultimate goal is to recover or approximate the true functional relationship between the response and the primary features. This goal together with PAN entails new performance requirements for the nonparametric module in PAN. To illustrate such a need, let us consider a simple OIS model

$$y = f^{(M)}(X) + \varepsilon = \sqrt{x_1 + x_2} + \varepsilon. \tag{6}$$

In traditional nonparametric variable selection problems, the regression model only sees the primary features, X, and the desired performance is successful identification of the active index set, $S_0 = \{1, 2\}$. The base iteration of PAN has a similar goal as we would want theselected index set S to be a superset of S_0. However, the intermediate descriptor space $x<m)$ in the mth iteration consists of nonlinear transformations of the selected primary features X_8, and the active index set is no longer well-defined.

Due to the iterative structure of PAN, a good nonparametric selection method for PAN must be able to generate and identify the m-composition descriptors $ft)(X)$ at the mth iteration (i.e., all active intermediate descriptors). To this end, a suitable nonparametric module should satisfy the following *PAN criterion:*

> It selects *all* of them-composition descriptors that are necessary for constructing the true M-composition descriptor, for all O ::: m .:::: M iterations.

Taking model (6) as an example, failingto select either $f_{1(o)} = x_1$ or $fio) = x_2$ in the base iteration will preclude the generation of $f(I) = (x_1 + x2)$ and $j<^2) = x1 - x2$ in subsequent iterations. The PAN criterion thus favors a "conservative" nonparametric method, in which false positives during selection are allowed but false negatives *must not* occur in any iteration. Here false positives are defined within each intermediate descriptor space $x<m)$_

The literature has provided a rich menu of nonparametric selection methods; however, the asymmetric effect of false positives and false negatives on descriptor selection illustrated by the PAN criterion motivates our choice of tree-based nonparametric approaches. To see this, suppose the true descriptor is $f(x_1) = \log(x_1)$ and the design matrix $O(X)$ consists of I-composition transformations of the primary features $X = (x_1, \ldots, xp)$- A typical nonparametric method aims to identify the oracle primary predictors (x_1 in this case) that associate with y through an unknown function $f(\bullet)$, without considering transformations of x_1 as possible candidate predictors. However, the goal in the presence of OIS is to identify the true primary predictors (xi) and the correct operator composition (log(-)). This goal, in the presence of many non-signal but highly correlated unarytransformationsof x_1, namely, Jxi, $Ix_1 1$, etc., is shown to be difficult for many nonparametric methods; see supplementary material Section A.2.1 for detailed analyses. Tree-based methods areinvariant to monotonic transformations and thus tend to be robust to related transformations that are often piecewise monotonic. Consequently, theymayselect unary transformationsof x_1 inaddition to $f(x_1) = \log(x_1)$-such false positives, although increasing the candidate search space, are favorably compatible with the PAN criterion. We next provide a review of BART (Chipman, George, and McCulloch 2010) and describe BART-G.SE (Bleich et al. 2014)-the default nonparametric module in the PAN framework.

BART isa Bayesian nonparametric ensemble tree method for modeling $y = f(X) + e$, the unknown relationship between a response vector y and a set of predictors $xi, \ldots, xp$- More specifically, BART models the regression function f bya sum of regression trees

$$Y = \sum_{i=1}^{m} g;(X1, \bullet\ldots, Xp; Ti, /1,;) + 8, \qquad \varepsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

Each binary regression tree $g;$ consists of a tree structure Ti partitioning observations into $B;$ terminal nodes, anda setofterminal parameters $/1,; = \{\mu ii, \ldots, \mu;B;\}$ attached to these nodes. Observations within a given terminal node b are constrained to have the same terminal parameter $\mu;b$- The prior distributions for (Ti, /1,;) constrain each tree to besmall so that each tree contributes to approximatefin a small and distinctfashion. Readers are referred to Chipman, George, and McCulloch (2010) for the full details of BART and posterior sampling.

The primary usage of BART is prediction, and the predicted values j, for y serve no purpose in variable selection. However, a variable inclusion rule can be defined based on the variable inclusion proportion $q;$ of $x;,$ which can be easilyestimated from the posterior samples. To this end, we adopt the permutation-based selection threshold based on the permutation null distribution of $q = (qi, \ldots, qp)$ proposed by Bleich et al. (2014).

Specifically, B permutations of the response vector $Yi, \ldots,)1$ are generated, and a BART model is fitted for each of the permuted response vectors with the same predictors $xi, \ldots, xp$- The variable inclusion proportions from the permuted BART models $qt, \ldots, qi$ are then used to create a permutation null distribution for the non-permuted variable inclusion proportion q. The predictor $x;$ is selected if $q; > m; + C'' \bullet s;,$ where $m;$ and $s;$ are the mean and standard deviation of the permuted variable inclusion proportion $qf = (qt_1, \ldots, qf,B)$, and $C^* = \inf\{cE!R+ \{Vi, f1I=f=l H(qtb .:::: m; + C \bullet s;) > I - a\}$ is the smallest global standard error multiplier (G.SE) that gives

a simultaneous 1 - *a* coverage across the permutation null distributions of *q;* for all predictors. We refer to BART with the permutation-based selection procedure described above as BART-G.SE.

Various BART-related methods have been recently developed (Linero 2018; Horiguchi, Pratola, and Santner 2021; Liu, Rockova, and Wang 2021). They often incorporate sparsity-inducing priors into BART and prove to be highly effective in various tasks. However, it is unclear whether the excellent performance of them developed in traditional settings carries over to being compatible with the PAN criterion. Indeed, we have found that methods aiming at optimally choosing nonparametric variables in traditional settings may incur fewer false positives but have a higher chance to miss the true descriptors in intermediate iterations-such false negatives are devastating in the context of PAN and OIS variable selection. The PAN criterion provides a useful guide in choosing not only the regression method but also the selection rule, for which we recommend BART-G.SE. In our numerical experiments, we vary the nonparametric module in PAN bycomparing several recent BART-related methods and other nonparametric selection methods, and find that the proposed iBART, PAN with BART-G.SE as the nonparametric module, is particularly well suited for OIS variableselection andtends to give the best overall performance; see the supplementary material for numerical results and a comprehensive discussion.

2.4. Practical Consideration and the Algorithm

The operators in O in (1) can be classified into the unary operators Ou = {J,exp,log, j • J, ,-l ,2 ,sin(rr-),cos(rr-)} and the binary operators Ob = {+,-,x,/,1 - j}, each posing different challenges to descriptor selection. The unary operators Ou inducestrongcollinearityamong the engineered descriptors; for instance, cor(xf, jx$_1$j) > 0.9 when x_1 ~ Nn(0,I) with n = 200. The binary operators Ob increase the descriptor space dimension double exponentially and generate complex nonlinear descriptors. These two issues are compounded when the two operator sets are used together. Therefore, we propose to decouple the two operator sets and alternate them, leading to two special cases of (5):

$$\mathcal{O}_{A_u}^{(M)}(\boldsymbol{X}) = \underbrace{\cdots \mathcal{O}_b \circ \mathcal{O}_u \circ \mathcal{O}_b \circ \mathcal{O}_u}_{\text{Mtimes}}(\boldsymbol{X}), \tag{7}$$

$$O;[; (\mathrm{X}) = \underbrace{\cdots Ou \circ Ob \circ Ou \circ \mathrm{Ob}(\mathrm{X})}_{\text{Mtimes}} = \mathrm{ot}\text{-}'\!>\!\mathrm{o}Ob(X). \tag{8}$$

In addition to the binary operators identified earlier, we include an additional binary operator, rr$_1$: ^{2} – - defined by rr$_1$ (*a, b)* = *a,*which allows intermediate descriptors to bepassed down unchanged. Note the two alternating descriptor spaces $O;.,$ (X) and $\mathcal{O}_{A_u}^{(M)}(\mathbf{X})$ are not equivalent to the full descriptor space $o<M(>X)$ with thesame composition complexity M. However, one can show that O;., ")(X) and $O;., b)$ (X) can recover o<M>(X) with some $Mu > M$ and $Mb > M$, respectively. This is formally described in the theorem below and its proof isavailable in supplementary material Section D.

Theorem2.1. LetX= *(xi,... ,xp)* E nxpbeaprimaryfeature space and *O* be a set of operators such that it can be partitioned into a unary operator set *Ou* and a binary operator set *Ob*. Suppose that *I* E *Ou* and rr$_1$ E *Ob*. Then for any *M* E N, there exists *Mu* *M* and *Mb* *M* such that $otu>(\mathrm{X})$ 2 $o<M(>\mathrm{X})$ and otb)(X) 2 o<M>(X), respectively.

For instance, the 2-composition descriptor space0$<^{2}>$(X) generated using (3) contains descriptors such as $/_1$ = *(xf* + *xJ)* and *Ji* = (x; + *Xj)*^{2}. These two descriptors can be generated using O;.,;(X) in (8): f_1 = add(square o J!"J *(x;,Xj),* square Orr, (Xj,X;)) E or;<X) andfi = square0 add(x;,Xj) E o;.,2;(X). Using(7),fi andfi can also begenerated as *f1* = add(square(x;)s, quare(xj)) E O;i,^{2}!(X) and h = square0 add(I(x;),I(Xj)) E or!(X). As such, under the M-composition OIS model in (4), one may consider using $o<M>,$ ot·>**(X),** or $otb)$**(X)** as long as they contain the true descriptors. In what follows we will focus on $O;.,$ (X) and $O;.,;$ **(X)** instead of$_{O(M)}$ for their aforementioned advantages.

We adopt the descriptor generating strategies in (7) and (8) for different scenarios. In particular, ifthe primaryfeatures X are believed to generate the model through their intricate interactionsthat will becaptured bybinary operators and compositions of such binary operators, we recommend (8). This is often the case in real-world applications, such as in Section 4, where the domain scientists have chosen a large set of potentially useful primary features, and unary transformations of these primary features are less interpretable and thus less desired. If such prior knowledge is not available and the relevant functional form of primary features is unknown, as in Section 3, we would recommend (7) to first identify such functional forms byselecting between unary descriptors.

The stopping criterion in Step 1 can be easily modified depending on practical needs. For example, we can specify the maximum composition complexity Mmax, like SISSO, or implement a data-driven criterion that terminates Step 1 when there exists a descriptor such that its absolute correlation with response variable y exceeds a pre-specified threshold *Pmax* that isclose to 1. We note that iBART allows larger Mmax than SISSO as iBART does not rely on one-short feature engineering, and the use of *Pmax* allows early termination.

It is also common in practice that one may only want to select k descriptors for easy interpretation, such as k ::: 5. If the cardinality of the selected descriptors X^* is greater than k, then an £$_0$-penalized regression may be used to choose the best k descriptors.

We allow user-specified options for these considerations when implementing iBART, summarized in Algorithm 1. The following default settings will be used in our experiments. We use the BART-G.SE thresholding variable selection procedure implemented in the R package bartMachine to perform nonparametric variable selection for Step l(a) in PAN and use LASSO implemented in the R package glmnet to perform parametric variable selection for Step 2 in PAN. For BART-G.SE, we set the number of trees to 20 that is recommended by Bleich et al. (2014), the number of burn-in samples to

10,000, the number of posterior samples to 5000, the number of permutations of the response to SO, and the rest of the parameters to the default values. With these values, our Markov chain Monte Carlo (MCMC) runs appear to have reached a sufficient number of iterations in our experiments. We choose the penalty term >.. in LASSO by minimizing the mean squared error loss through a 10-fold cross-validation procedure and set the other parameters to the default values. Unreported results using real data indicate that other tuning methods to debias LASSO yield similar performance. The algorithm terminates if the composition complexity M reaches Mmax = 4 or the maximum absolute correlation IPI reaches *Pmax* 0.95, whichever occurs first. Setting Mmax to 4 appears sufficient for the considered materials GWAS application as descriptors with more complexity become challenging to interpret.

Algorithm 1: iBART

Input: *Pmax* E (0, 1) = maximum absolute correlation with the response variable;
Mmax = maximum composition complexity;
Lzero = whether to perform fa-penalized regression;
k = number of selected descriptors byfa-penalized regression. Only required when Lzero == TRUE.
Output:.XZ: selected descriptors
Data: X: primary features; *Ou:* set of unary operators; *Ob:* set of binary operators; *y:* response vector
M=0
p =maxxEX(M>*cor(x,y)*
while *M* Mmax *or* IPI *Pmax* **do**
  x<M)' +- BART-G.SE selected descriptors on x<M)
  x<M+1> +- *OM+1* (u;x<;>')
  M+-M+1
  p +- maxxEx<M> *cor(x,y)*
end
X* +- LASSO selected descriptors on x<M)
if *Lzero* == *TRUE and* IX*I > k then
  XZ +- best k descriptors from fa-penalized regression
end
else
  I *XZ* +-X*
end

3. Simulation

3.1. Outline

In Sections3.2 and 3.3wedemonstrate that the employed BART-G.SE tendsto satisfythe PAN criterion whenselecting unaryand binary operators, respectively, and evaluate its false positives. Section 3.4 assesses iBART relative to several existing descriptor selection methods in view of the PAN criterion and OIS variable selection accuracy using a complex simulation setting. Section 3.5 showcases the robustness of iBART in the initial input dimension *p*. Section 3.6 examines the performance of iBART when the operator set O can not generate the ground truth model. We use Algorithm 1 and the default settings described in Section 2.4 when implementing iBART, unless stated otherwise. Each simulation is replicated 100 times. We also compare variants of iBART by varyingthe nonparametric module;see the supplementary material for details.

3.2. Unary Operators

We consider all unary transformations of five primary features ***X*** = *(xi, ... ,x5),* that is, the descriptor space is *Ou(X)* = Ui=i{x;, *x;-1, xf,.jxi,* log(x;),exp(x;), lxd, sin(Jtx;), cos(:,r*x;)* }. For each unary operator *Uj* E *Ou,* we generate the response vector by

$$y = 10u_j(x_1) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}_n(0, I), \tag{9}$$

with sample size n = 200, yielding nine independent models in total. We draw the primary features X independently from the standard normal distribution if the domain of *Uj* is IR, and the Lognormal(2, 0.5) distribution if the domain is IR a-

The left plot in Figure 2 shows the number of true positives (TP) of BART-G.SE. For each of the nine models in (9), the true descriptor is selected 100/100 times. This suggests that BART-G.SE is capable of identifying the true descriptor with high probability when the descriptor space is populated with unary operators *Ou.* Other nonparametric methods did not select all TP 100/100 times for all ninescenarios, failing the PAN criterion; see the supplementary material for further results and discussion.

The left plot in Figure 2 shows the number of FP of BART-G.SE for each simulation setting. Although BART-G.SE is capable of capturing the true descriptor with high probability, it also selects some non-signal descriptors in some cases. The number of FP is especially high when the true descriptor is *xi,* lx1I, exp(x1),log(x1), *x1,* or *,jxi.* This isdue to high collinearity among thesesixdescriptors. In particular, the empirical Pearson correlation between log(x_1) and *Fi* is over 0.99 under simulation setting (9) and thus some false positives are expected. Selecting inactive descriptors in one iteration isless of a concern in variable selection with OIS as theydo not constitute misspecified models, and can be further screened out during Step 2 of PAN.

3.3. Binary Operators

In this simulation study, we apply binary operators *Ob* {+,-,x,/,1 - 1,7ti} on the five primary features *X* *(xi, ... ,X5).* Thedescriptor space *Ob(X)* E IR^{2}ooxss contains all binary transformations of all possible pairs of the five primary features. For each binary operator *bj* E *Ob,* we generate the response vector by

$$y = 10b_j(x_1, x_2) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}_n(0, I),$$

with sample size n = 200, yielding a total of six independent models. We generate the primary features X following *X1, ... ,X5* $\overset{\text{iid}}{\sim}$ *Nn(O,l).*

Theright plot in Figure 2 shows the number of TP for each of the six models in (lo). BART-G.SE is able to identify the true binary descriptor 100/100 times in all sixsettings, similar to earlier observations in Section 3.2. The right plot in Figure 2 shows that BART-G.SE mayalso select some irrelevant descriptors but

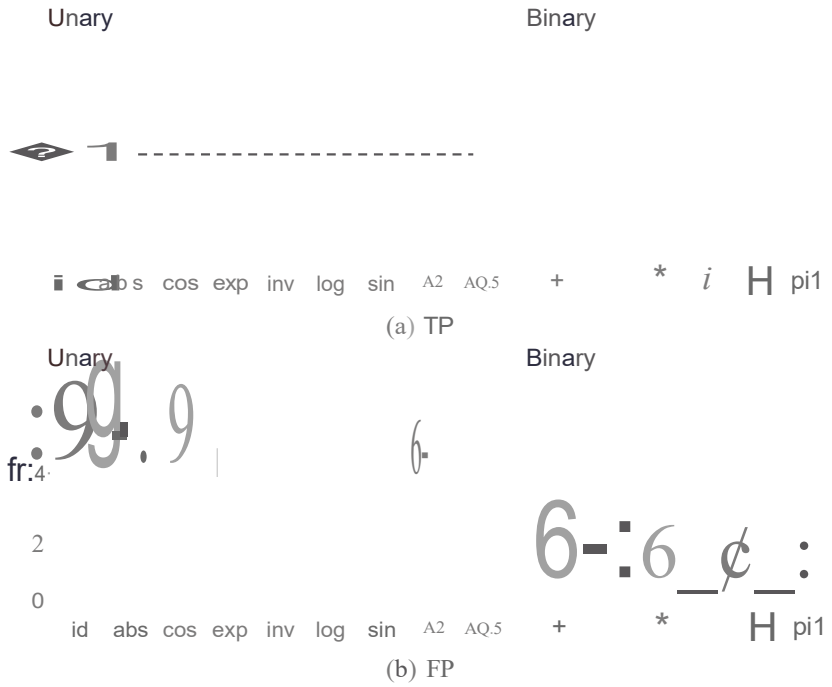

Figure 2. Boxplots ofTPandFPover100simulation replicatesunder (9) and(10).

considerably fewer than in Section 3.2. In particular, when the truedescriptor is *(x1 -xi),* BART-G.SEcorrectly selects $(x_1\text{-}x2)$ butalso selects ***Ix,-xii***100/100 times. Unlike in Section 3.2, the inclusion of Ix, *-xi* I is not due to highcorrelation between Ix, -x21and $(x_1 - x2)$. In fact, their empirical Pearson correlation is merely 0.07. We suspect that ***Ix, -*** $x2_1$ was selected due to its piecewise monotonicity in $(x_1 - x2)$ that tendsto be invariant to tree-based methods.Similarly, BART-G.SE selects both $(x_1\text{-}x2)$ and***Ix, -*** *Xi* I with high probability when the true descriptor is Ix, - *xi* I- When the true descriptor is (x, + *xi), x,* x *xi, xif*x2, or $rr_1(x_1, x2)$, the FP of BART-G.SE is nearly zero. This shows that BART-G.SE is not too conservative when high collinearity does not exist.

Our investigation using unary and binary descriptors suggests that the permutation threshold ofBART-G.SE tends to satisfy the PAN criterion without being too conservative, making it a good candidate for PAN and OIS variable selection. We next use a more complex simulation to assess iBART in view of the PAN criterion and OISselection accuracy.

3.4. Complex Descriptors with High-Order Compositions

In this section, we compare iBART with existing approaches under a complex model and demonstrate superior performance of iBART. We use the following model

$$y = 15\{\exp(x,) - exp(xi))i + 20\sin(rrX3X4) + e, \quad e \sim Nn(0,a^2I), \qquad (11)$$

with $n = 250$, $p = 10$, and $a = 0.5$. Here the number of primary features p is set to a relatively small number since competing one-shot methods cannot complete the simulation when p::::: 20 due to the ultra-high dimension of their descriptor spaces. In Section 3.5, we demonstrate that iBART scales well in p and gives a robust performance. We use the operator set CJ defined in (1) with rr1, and the primary feature vectors x_i aredrawn independently from a uniform distribution, namely, $x_1, \ldots, x_p \overset{iid}{\sim}$ Un(-1, 1). Section B of supplementary material demonstrates the effect ofdependent primaryfeatureson iBART and other methods using the same functional relationship in (11).

iBART is implemented in R based on Algorithm 1 in Section 2.4 using the default settings. We chose the descriptor generating process (7) in Section 3 since the iid primary features do not show strong collinearity. Two versions of iBART are considered: iBART without and with lo-penalized regression, labeled as "iBART" and "iBART+*lo*", respectively. The l_0-penalization finds the best subset of k variables from the set of selected variables using the Akaike irlformation criterion (AIC) with $k \in \{1, 2, 3, 4\}$. Wecompare the performance of iBART and iBART+*lo* with SISSO and LASSO.SISSO is implemented using the Fortran 90 program published by Ouyang et al. (2018) on GitHub with the following settings: the descriptor magnitude allowed in the descriptor space is set to $[1 \times 10^{-6}, 1 \times 10^{5}]$; the size of the SIS-selected subspace is set to 20; the operator composition complexity M is set to 3; the maximum number of operators in a descriptor is set to 6; and the number of selected descriptors $k \in \{1, 2, 3, 4\}$ is tuned by AIC. The R package glmnet isused to implement LASSO and the penalty parameter .11. is tuned via 10-fold cross-validation to minimize the mean squared error loss. Since the size of $CJ(^3)(X)$ exceeds the limit of an R matrix, we give LASSO an advantage by reducing the descriptor space to $CJ_u \circ O \circ O$ (X) that aligns with the true data-generating process. We also use the lo-penalized regression step as in SISSO and iBART+*lo* for LASSO, leading to LASSO+l_0.

For each method, we calculate the number of TP selections, FP selections, and false negative (FN) selections. We use the

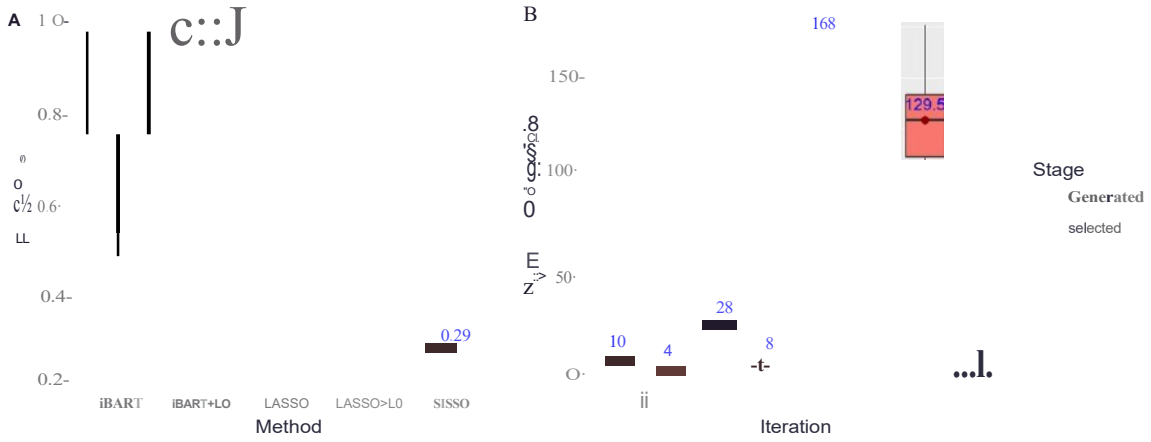

Figure 3. Left: boxplots of F1 scores over 100 simulations for different methods under Model (11). Right Boxplots of iBART generated and selected descriptors in each iteration.

F_1 score as an overall metric to quantify the performance of each method: $F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$, where Precision $=$ 'T'P1J'FP and Recall $=$ IPZFN. The value $F_1 = 1$ means correct identification of the true model without having any FP and FN. In thissimulation, the two TPsare/$_1(X) = \{\exp(x_1)\text{-}\exp(x_2)\}^2$ andfi(X) $=$ sin(rrx3x$_4$).

As shown in Figure 3A, both iBART-based methods achieve very high F_1 scores, with a median F_1 of 0.8 and **1** for iBART and iBART+£$_0$, respectively. In particular, both iBART and iBART+fo have 2 TPs in all simulations while incurring an average FP of 0.93 and 0.3, respectively. This demonstrates that iBART satisfies the PAN criterion in all iterations as it is able to identify the 2 TPs 100/100 times in simulations. Furthermore, iBART+£$_0$ gives a very low FP,scoring a perfect F_1 score 75/100 times. LASSO-based methods and SISSO have lower F_1 scores than iBART and iBART+fo butfor different reasons. LASSO selects 37.65 descriptors on average, and the 2 TPs are always selected. This means that LASSO also enjoys the PAN criterion but its F_1 score is hindered by the large number of FP. With the help of the £$_0$-penalized regression, LASSO+fo reduces the average number of FP from 35.65 to 2 while maintaining 2 TPs 100/100 times, substantially increasing the F_1 score to 0.67, the third-highest score but still considerably lower than iBART and iBART+£$_0$. SISSO, on the other hand, has only 1 TP but 3 FPs in all simulations, which unfortunately does not satisfy the PAN criterion. In particular, SISSO can identify sin(rrx3x$_4$) 100/100 times but always selects a false signal, $|\exp(x_1) - \exp(x_2)|$, a descriptor that has an absolute correlation over 0.9 with the TP, $\{\exp(x_1) - \exp(x_2)\}2$. In summary, both iBART-based and LASSO-based methods satisfy the PAN criterion but LASSO-based methods incur more FPs. SISSO, however, fails to identify one of the two descriptors 100/100 times but selects a highly correlated counterpart, indicating its relatively weakened ability to distinguish the true descriptor when there are descriptors highly correlated with the TP. Results not reported here show that replacing BART-G.SE with LASSO in the PAN framework missed TPs with high probability even at the base iteration, indicating the importance of nonparametric variable selection in PAN.

To gain insight into the scalability of iBART, Figure 3(B) shows the boxplots of the number of iBART generated and selected descriptors in each iteration. Throughout the 100 simulation replicates, iBART generates no more than 168 ::: $2p^2$ descriptors, which is significantly less than the number of descriptors generated by SISSO (9.26×10^9) and LASSO (1.2×10^6). Such dimension reduction not only reduces runtime and memory usage but also enables iBART to tackle data with much larger p, as we show in Sections 3.5 and 4.

3.5. Largep

In thissection, wedemonstrate that the performance of iBART is robust to increase in input dimension p while one-shot descriptor selection methods are not. Under Model (11) with $p = 20$, SISSO with the same parameter settings in the precedingsection generates more than3.8×10^{11} descriptors and failed to complete one simulation within 24 hr on a server with 1.3TB of memory and 40 CPUcoresavailable to SISSO. LASSO, on the other hand, failed at $p = 20$ since the glmnet function in the R package glmnet cannot handle a matrix of size 250 x (1.57×10^7). The proposed method iBART instead scales well in the input dimension p partly because of the efficient dimension reduction via the PAN strategy. In particular, when $p = 20$, iBART tookan average of 497 sec to finish, and its memory usage peaks out at 10.85 GB in 100 simulation replicates.

We implement iBART for p E {10,20,50, 100,200} using the same simulation settings in Section 3.4. Figure 4 shows that iBART'sF_1 score is robust to the increase in p. In fact, benefiting from the *ab initio* mechanism through the PAN strategy, the performance of iBART would be identical for varying $p$as long as it selects x_i, x_2, x_3, x_4 in the first iteration, which is often the case based on Figure 4 at least under the current simulation setting. We acknowledge that the stability of the F_1 scores across different values of p may be attributed, in part, to the relatively small noise standard deviation. In contrast, one-shot descriptor selection methods suffer significantly from just a small increase in the input dimension p since the descriptor space increases double exponentially in p. One way for one-shot description selection methods to circumvent the ultra-high dimension issue is to reduce the maximum composition complexity Mmax· However, this undesirably rules out descriptors with the correct composition complexity. For example, the descriptor in (11)/$_1(x) =$

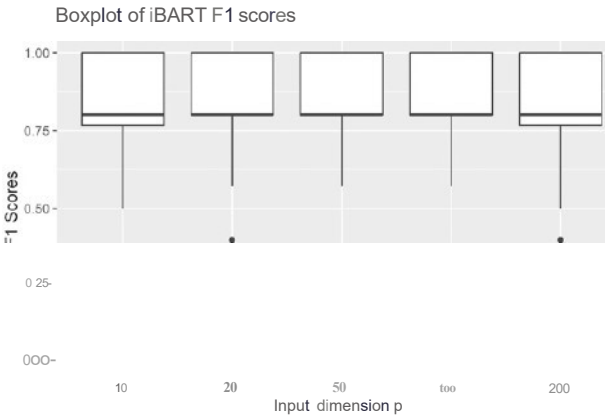

Figure 4. Boxplot of F1 scores for iBART under Model (11) with $p \in \{10,20,50,100,200\}$.

$\{\exp(x_1) - \exp(x_2)\}^2$ requires at least three compositions of operators and reducing Mmax to2 means that $Ji(x)$ would never be generated and hence would never be selected. Thus, in a complexscenario, one-shot descriptor selection methodscannot generate and select the complex descriptors unless p is small. We considered a small $p = 10$ in the preceding section to accommodate such limitation of one-shot methods.

3.6. Model Misspecification

In this section, we examine the behavior of iBART when the operator set O is not sufficient to generate the data-generating function. We generate data from the model $y = xJ^{\cdot 7} + e$ with $e \sim$ .Nn(O,I), $n = 250$, $p = 10$, use the same operator set(') asinSection 3.4, and draw the primary features from a uniform distribution, that is, $x_1, \ldots, x_p \; Vn(0, 3)$. In this simulation, iBART iterates once after screening the primary features. Since iBART onlyselects x_1 in the first iteration,100/100times, it does not apply the binary operators in 0.

In contrast to the preceding sections, the computation of true and false positives, as well as the F_1 scores, is not feasible here due to the model specification by design. To evaluate the performance of our method, Figure 5(A) shows the frequency of unique descriptors selected by iBART and the average Pearson correlation between the selected descriptor and $x^{\cdot 7}$ in parentheses. Note that multiple descriptors might be selected in one replication, and therefore the sum of frequencies exceeds 100. We can see that, although the true descriptor $xj-7$is not in the candidate descriptor space, iBART selects highly correlated descriptors from the candidate space, including xy for 98% of the time and x1 for 87% of the time. Figure 5(B) shows that the RMSEof the iBARTselected modelsdoes not deviate much from the RMSE of the true model, indicating a similar explanatory power. These observations suggest that when the employed operator space is not sufficient to generate the ground truth, iBART iscapable of selecting a model that closely resembles the true model, at least in the considered simulation setting. This is reassuring as **in** manyapplications ofOIS,such as the oneinSection 4, the goal is precisely to find an accurate but interpretable model that well approximates complex real-world physical systems.

4. Application to Single-Atom Catalysis

Weapply theproposed method to analyze a single-atom catalysis dataset (O'Connor et al. 2018) in which the goal is to identify physical descriptors that are associated with the binding energy of metal-support pairs calculated by density functional theory (DFf). Single-atom catalysts are popular in modern materials science and chemistry as they offer high reactivity and selectivity while maximizing utilization of the expensive active metal component (Yang et al. 2013; O'Connor et al. 2018; Wang, Li, and Zhang 2018). However, single-atom catalysts suffer from a lack ofstability caused by the tendency for single metal atoms to agglomerate in a process called sintering. To prevent sintering in single-atom catalysts, onecan tune thebindingstrength between single metal atoms and oxide supports. While first principle simulations can calculate the binding energy for given metal-support pairs, modeling their association requires explicit statistical modeling and is key to aid the design of single-atom catalysts that are robust against sintering. Feature engineering leads to physical descriptors constructed using mathematical operators and physical properties of the supported metal and the support, and gains popularity in materials informatics as the obtained descriptors are interpretable and provide insights into the underlying physical relationship. The key challenge is to select the most relevant physical properties among large-scale candidate predictors that haveexplanatory and predictive power to the binding energy; often the sample size is small as first principle simulations are computationallyintensive.

The data comprise bindingenergy of $n = 91$ metal-support pairs and $p = 59$ physical properties of these metal-support pairs, in which the bindingenergy serves as the response y while the 59 physical properties constitute the primary features $X = (x_1,\ldots, x_{59})$. We use the operator set given in (1) but exclude sin(n •) and cos(n ·) because they are not physically meaningful in this data application.

In this analysis, we compare the best k descriptors $(k = 1, \ldots, 5)$ of iBART+fo (referred to as iBART for brevity), SISSO, and the method proposed in O'Connor et al. (2018). The method proposed in O'Connor et al. (2018) combines LASSO with extensive domain knowledge; in particular, they rule out a substantial proportion of less promising descriptors using expert knowledge of single-atom catalysis. Such expert-guided, tailored dimension reduction is not needed for SISSO or iBART, but is necessary for LASSO-type methods because the astronomical size of$0^{<2)}$(X) in this application far exceeds the maximum matrix size allowed in many modern programming languages. We refer to O'Connor et al. (2018) as LASSO" to distinguish it from LASSO+fo implemented in Section 3.4. To enhance interpretability, we eliminate nonphysical descriptors, such as *volumn + speed,* by automatically comparing the units of the constructed descriptors. Results without this constraint are reported in the supplementary material, showing that this constraint does not alter our conclusions when comparing methods but substantially improves the interpretability of the identified model.

For iBART, wefollow thesettings described in Section 2.4 and use the descriptor generating process (8). The SISSO parameters are set as followed. The maximum number of descriptors

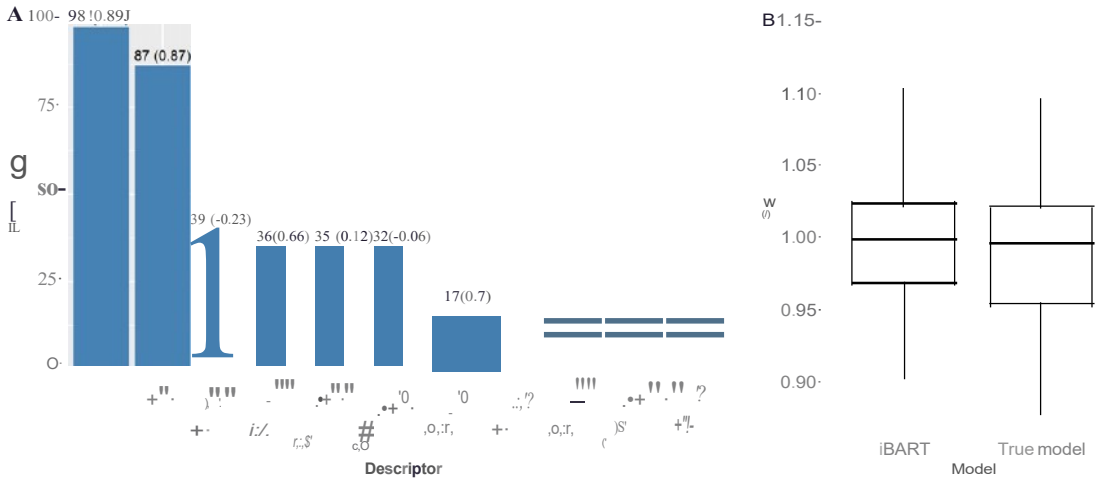

Figure5. Left:frequency (averagePearson correlation withxJ-7) of iBARTselected descriptors. Right:root meansquares error (RMSE) of iBARTmodelsand the truemodels.

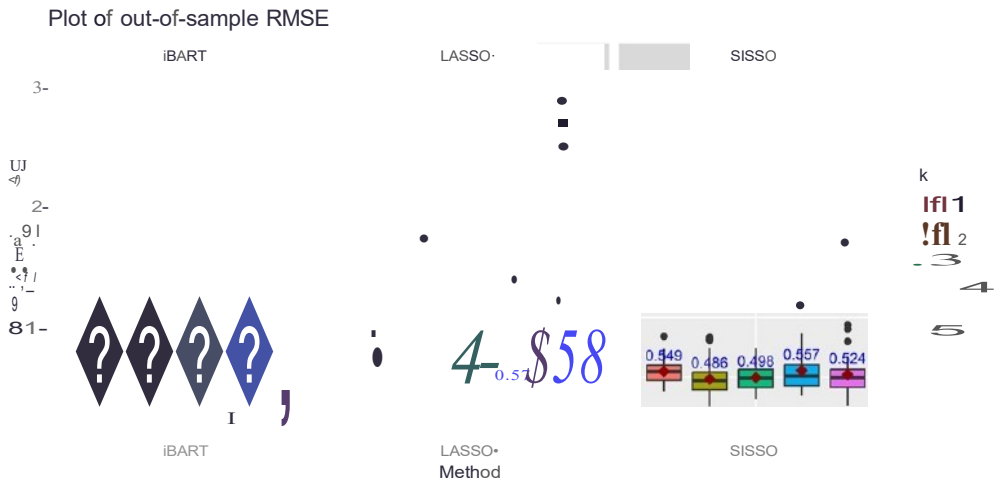

Figure 6. Boxplot of the out-of-sample RMSE for each method across SO random partitions with 1 k 5 (left to right in each plot). The blue numbers and the red rhombusesindicate the average out-of-sample RMSE.

allowed in a model is set to 5; the descriptor magnitude allowed in the descriptor space isset to [1 x 10-^{8}, I x 10^{8}]; the size of the SIS-selected subspace is set to 40; the composition complexity M is cap at 2 because0<^{3}) (X) and higher complexity space exceed the maximum number of elements allowed in a Fortran 90array. The LASSO* procedure is implemented using the MATLAB code published by O'Connor et al. (2018) with all parameters left as default.

Each method's performance is assessed using out-of-sample RMSE, runtime, and the number of generated descriptors. To calculate out-of-sample RMSE, we randomly partition the n = 91 observations into a 90% training set (82 samples) and a 10% testing set (9 samples) and repeat this process SO times. For brevity, we herein refer to out-of-sample RMSE as RMSE.

From Figure 6, the smallest average RMSE is attained at k = 3 for iBART (0.41), k = 1 for LASSO* (0.471), and k = 2 for SISSO (0.486), that is, iBART reduces the RMSE by 13% relative to LASSO* and 16% relative to SISSO; also see the first row of Table 1. iBART outperforms LASSO* and SISSO with smaller average RMSE and reduced variability for k 2. The larger average RMSE of iBART at k = 1 may be partially due to its smaller descriptor space as a result of its alternating descriptor generation process, and our implementation details that give advantages to LASSO*. Multiple descriptors are often needed to approximate complex physical systems, and the predictive performance ofiBART reassuringly improves with moredescriptors in the model. On the contrary, the RMSE of LASSO* increases with more descriptors in the model, indicating overfitting. For all k, we observe iBART does not report large deviations from the average performance, suggesting robustness to training sets compared to LASSO* and SISSO.

Table 1. Performance comparison of threemethods:out-of-sample RMSE, runtime, and the number of generated descriptors, averaged over50 cross-validations.

	iBART	LASSO*	SISSO
RMSE	0.41	0.471	0.486
Runtime	225 sec	5511 sec x 20	6943 sec
Number of generated descriptors	627	3.3 x 10^{5}	5.5 x 10^{7}

In addition to the performance gain in RMSE, iBART also leads to a substantial reduction in computing time and memory usage. Table 1 shows that iBART leads to over 30-fold (6943/225 = 30.86) speedup compared to SISSO, and over 480-fold (5511 x 20/225 = 489) speedup compared

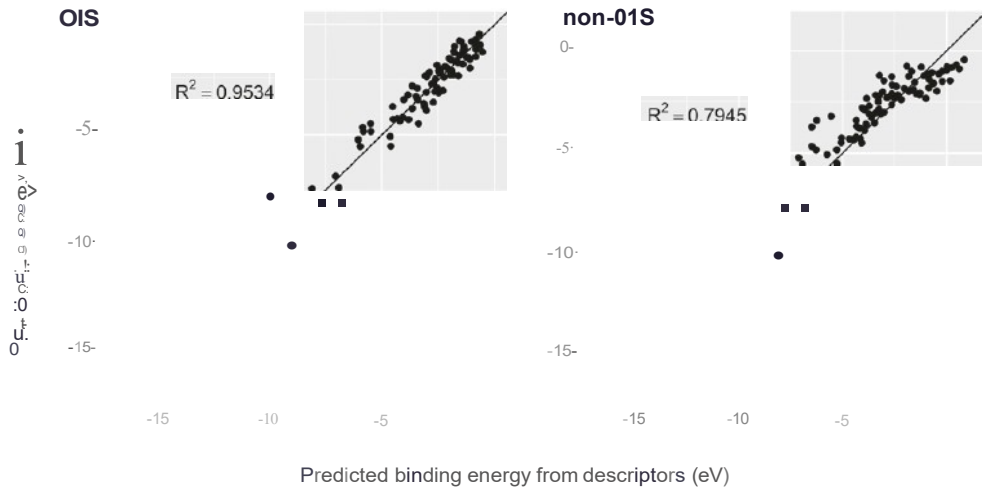

Figure 7. DFT binding energies versus predicted valuesusing linear modelswith015and without015.The black lineis$y = x$.Each model has$k = 3$ descriptors.

Table 2. Selected linear modelsby iBART fork E {1,2,3,4, 5).

k	Selected descriptors
2	$\left\lvert\frac{\Delta H_{sub}-\Delta H_{f,ox,bulk}}{\Delta E_{vac}}\right\rvert$, $\frac{EA^s\cdot(\Delta H_{sub}-\Delta H_{f,ox,bulk})}{N^m_{val}/CN^m_{bulk}}$, $\left\lvert\frac{\Delta H_{sub}-\Delta H_{f,ox,bulk}}{\Delta E_{vac}}\right\rvert$
3	$EA^s\cdot\Delta H_{f,ox,bulk}$, $\left\lvert\frac{\Delta H_{sub}-\Delta H_{f,ox,bulk}}{\Delta E_{vac}}\right\rvert$, $\left\lvert\frac{(\eta^{1/3})^m}{N^m_{val}\cdot\Delta E_{vac}}\right\rvert$
4	$EA^s\cdot\Delta H_{f,ox,bulk}$, $\frac{(\eta^{1/3})^m}{\phi^s}$, $\frac{(\eta^{1/3})^m\cdot IE^s_4}{(N^m_{val})^2}$, $\left\lvert\frac{EA^s\cdot(\Delta H_{sub}-\Delta H_{f,ox,bulk})}{\Delta E_{vac}}\right\rvert$
5	$\frac{(\eta^{1/3})^m}{\phi^s}$, $\frac{\Delta H_{f,ox,bulk}}{\Delta E_{vac}}$, $\frac{EA^s\cdot(\Delta H_{sub}-\Delta H_{f,ox,bulk})}{N^m_{val}/CN^m_{bulk}}$, $\frac{(\eta^{1/3})^m\cdot IE^s_4}{(N^m_{val})^2}$, $\left\lvert\frac{\Delta H_{sub}-\Delta H_{f,ox,bulk}}{\Delta E_{vac}}\right\rvert$

Table 3. Descriptions of the selected primary featuresby iBART.

Primary feature	Physical meaning
ti.H,ub	Heat of sublimation
ti.Htox bulk	Oxidation energy of the bulk metal
ti.fv c•	Oxygen vacancy energy
EA'	Electron affinity of support
all..	Number of valence electrons in metaladatom
(N'b lk	Coordination number of the surface metal atomin the bulk phase
	Discontinuity in electron density of metal adatom
	Chemical potential of the electrons in support
	Fourth ionizationenergy of support withthebulk metal in the 4+oxidation state

to LASSO*, tested on an Intel Xeon Gold 6230 CPU @ 2.10 GHz using either **1** or 20 CPU cores. We use the published code by authors for competing methods, and the runtime reported here does not isolate the effect of various programming languages (R for iBART, MATLAB for LASSO*, and Fortran 90 for SISSO). We did not obtain an accurate single-core runtime of LASSO* because of its poor scalability, and instead multiplied its 20-core runtime by 20 for comparison (i.e., 5511 x 20 x 50/3600 = 1530 hr); the exact speedup of iBART compared to the single-core runtime of LASSO* mayvary due to the reduced communication cost among multiple cores and other factors. We remark that the LASSO* method implemented here is given an advantage with an additional dimension reduction step, and an exploration of higher complexity descriptor space as in iBART is computationally prohibitive for LASSO*. The excellent scalability of iBART transforms into memory efficiency as the descriptor space in iBART is orders of magnitude smaller than that of competing methods. Table 1 shows that iBART generates a descriptor space of size 627 < $O(p^2)$ in the last iteration on average. Thanks to this significantly smaller descriptor space, we were able to run iBART on a laptop with only 16GB of memory; in contrast, LASSO* and SISSO failed at the descriptor generation stepowing to the enormous descriptor space theytry to generate, and require server-grade computing facilities.

Table 2 reports the selected descriptors by iBART with various k using the full dataset, and Table 3 reports the physical meanings of the selected primary features. We can see that some descriptors are recurrent in various models, such as $\left\lvert\frac{\text{ti.H,ub-t,.Hr,ox,bulk}}{\text{,'.Evac}}\right\rvert$. This reassuringly suggests the stability of the proposed descriptor selection method across k. Readers are referred to Liu et al. (2022) for in depth analysis and physical interpretation of these selected descriptors from the perspective of catalyst design.

Figure 7 demonstrates a clear advantage of the OIS model over the non-OIS model. The OIS model is the iBART model with $k = 3$, the optimum model suggested by RMSE. The non-OIS model is a simple least squares model with X as the design matrixand no feature engineeringstep. For ease of comparison, the non-OIS model also has $k = 3$ predictors determined by best subset selection.The OIS modelyieldsan R^2 of0.9534, indicating high explanatory power. In contrast, the non-OIS model gives an $R^2 = 0.7945$ and shows poor fitting performance as the scatterplot deviates from the diagonal line $y = x$ considerably. Both the OIS and the non-OIS models have analytical forms, but the OIS model gains more explanatory power and gives an insightful description of the response through nonlinearity of its predictors. In particular, the non-OIS model is .Ynon-OIS = -4.7- 0.4 x t:.Hf,ox + 0.3 x N!a1 + 1.0 x t:.Evac, while the OIS model is

$$.Yrns = -0.01 + 0.4 \times \left(EA^5 \cdot \frac{\text{t:.Hf,ox,bulk}}{\text{f:.Hsub-}} - \frac{\text{f:.Hf,ox,bulk}}{\text{l:.Evac}}\right) - 0.6 \times \left\lvert\frac{(771/3r}{\text{f:.Evac}}\right\rvert - 19.6 \times$$

This O1S model pinpoints targeted descriptors to guide further investigation into the underlying physical discovery.

5. Discussion

In this article, we study variable selection in the presence of feature engineering that is widely applicable in many scientific fields to provide interpretable models. Unlike in classical variable selection, candidate predictors are engineered from primary features and composite operators. While this problem has become increasingly important in science, such as the emerging field of materials informatics, the induced new geometry has not been studied in the statistical literature. We propose a new strategy "parametrics assisted by nonparametrics': or PAN, to efficiently explore the descriptor space and achieve nonparametric dimension reduction for linear models. Using BART-G.SE as the nonparametric module, the proposed method iBART iteratively constructs and selects complex descriptors. Compared to one-shot descriptor construction approaches, iBART does not operate on an ultra-high dimensional descriptor space and thus substantially mitigates the "curse of dimensionality" and high correlations among descriptors. We introduce the O1S framework, define a PAN criterion, and assess iBART through the lens of this criterion. This sets a foundation for future research in interpretable model selection with feature engineering.

Other than the methodological contributions, we use extensive experiments to demonstrate appealing empirical features of iBART that may be crucial for practitioners. iBART automates the feature generating step; one-shot methods such as SISSO often require user intervention as otherwise they are not scalable to handle the overwhelminglylarge descriptor space-this descriptor space increases double exponentially in the number of primary features and binary operators as a result of compositions. iBART has excellent scalability and leads to robust performance when the dimension of primary features increases to a level that renders existing methods computationally infeasible. Compared to SISSO, which is widely perceived as state-of-the-art in the field of materials genomes, our data application shows that iBART reduces the out-of-sample RMSE by 16% with an over 30-fold speedup and a fraction of memory demand. Overall, the proposed method accomplishes traditionally "server-required" tasks using a regular laptop or desktop with improved accuracy. Beyond materials genomes illustrated in Section 4, iBART can be applied to a vast domain where interpretable modeling is of interest.

There are interesting next directions building on our iBART approach. First, O1S provides a useful perspective for structured variable selection by introducing operators and compositions. Certain operator sets may be particularly useful depending on the application; for example, the composite multiply operator leads to high-order interactions. It is interesting to examine the performance of iBART with a diverse choice of operators. Also, we have focused on finite sample performance when assessing selection accuracy through simulations partly in view of the limited sample size typically available in the motivating example. It is nevertheless interesting to theoretically investigate the necessary conditions for the PAN criterion to hold for selected nonparametric variable selection methods such as BART when the sample size diverges.

Supplementary Materials

Code and data for replicating the study resultsare available at *https://github.com/xylimeng/OIS*. An R package that implements iBART is available at *https:/ Igithub.com/mattshengliBART*. The supplementary materials include additional simulations and proof. Supplementary A contains a detailed comparison of variants of iBART by varying the nonparametric module. Supplementary B contains additional simulation results using correlated primaryfeatures.Supplementary C presents additional real data application results without enforcing unit consistency. Supplementary D contains the proof of Theorem 2.1.

Acknowledgments

We thank the editor, associate editor, and reviewers for constructive comments that helped to improve the article.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

This research was partially supported by the Big-Data Private-Cloud Research Cyberinfrastructure MRI-award funded by NSF under grant CNS-1338099 and by Rice University's Center for Research Computing. Shengbin Ye's research was supported by NIH grant T32CA096520. Meng Li's research was partially supported by NSF grants DMS-2015569 and DMS/NIGMS-2153704. Thomas P.Senftle's research was supported by NSF grant CBET-2143941.

ORCID

Shengbin Ye http:l/orcid.org/0000-0001-8767-2595
Thomas P.Senftle http:l/orcid.org/0000-0002-5889-5009
Meng Li$ http://orcid.org/0000-0003-2123-2444

References

Bleich, J., Kapelner, A., George, E. I., and Jensen, S. T. (2014), "Variable Selection for BART: An Application to Gene Regulation:• *Annals of Applied Statistics,* 8, 1750-1781. [82,85,86]

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010), "BART: Bayesian Additive Regression Trees," *Annals of Applied Statistics,*4, 266-298. [85]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional FeatureSpace," *Journal of theRoyal Statistical Society,*Series B,70,849-911. [82]

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., and Hutter, F. (2015), "Efficient and Robust Automated Machine Learning," in *Advances in Neural Information Processing Systems* (Vol. 28), Curran Associates, Inc. [83]

Foppa, L., Ghiringhelli, L. M., Girgsdies, F., Hashagen, M., Kube, P., Havecker, M., Carey, S. J., Tarasov, A., Kraus, P., Rosowslct, F., Schlogl, R., Trunschke, A., and Scheffler, M. (2021), "Materials Genes of Heterogeneous Catalysis from Clean Experiments and Artificial Intelligence," *MRS Bulletin,* 46, 1016-1026. [81]

Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V.,Drax!, C., and Scheffler, M. (2015), "Big Data of Materials Science: Critical Role of the Descriptor:• *Physical Review Letters,* 114,105503. [81]

Hart, G. L., Mueller, T., Toher, C., and Curtarolo, S. (2021), "Machine Learning for Alloys," *Nature Reviews Materials,* 6, 730-755. [81]

Horiguchi, A., Pratola, M. T., and Santner, T. J. (2021), "Assessing Variable Activity for Bayesian Regression Trees," *Reliability Engineering & System Safety,* 207, 107391. [86]

Keith, J. A., Vassilev-Galindo, V., Cheng, B., Chrniela, S., Gastegger, M., Miiller, K.-R., and Tkatchenko, A. (2021), "Combining Machine Learningand Computational Chemistry for Predictive Insights into Chemical Systems:' *Chemical Reviews,*121, 9816-9872. [81]

Khurana, U., Samulowitz, H., and Turaga, D. (2018), "Feature Engineering for Predictive Modeling Using Reinforcement Learning;' in *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18).* [83]

Lin, Y., Saboo, A., Frey, R., Sorkin, S., Gong, J., Olson, G. B., Li, M., and Niu,C. (2021), "CALPHAD Uncertainty Quantificationand TDBX;' *The Journal of The Minerals, Metals & Materials Society,*73, 116-125. [81]

Linero, A. R. (2018), "Bayesian Regression Trees for High-Dimensional Prediction and Variable Selection," *Journal of the American Statistical Association,* 113, 626-636. [86]

Liu,C.-Y., Ye,S., Li, M., and Senftle, T. P. (2022), "A Rapid FeatureSelection Method for Catalyst Design: Iterative Bayesian Additive Regression Trees (iBART)," *The Journal of Chemical Physics,*156,164105. (82,92]

Liu, C.-Y., Zhang, S., Martinez, D., Li, M., and Senftle, T. P. (2020),"Using Statistical Learning to Predict Interactions between Single Metal Atoms and Modified MgO(100)Supports; *npj Computational Materials,*6, 102. (82]

Liu, X., Wu, Y., Irving, D. L., and Li, M. (2021), "Gaussian Graphical Models with Graph Constraints for Magnetic Moment Interaction in High Entropy Alloys," ArXiv e-prints. (81]

Liu, Y., Rockova, V., and Wang, Y. (2021), "Variable Selection with ABC Bayesian Forests:' *Journal of the Royal Statistical Society,* Series B, 83, 453-48i. I86]

Markovitch, S., and Rosenstein, D. (2002), "Feature Generation Using General Constructor Functions," *Machine Learning,* 49, 59-98. (83]

O'Connor, N. J., Jonayat, A. S. M., Janik, M. J., and Senftle, T. P. (2018), "Interaction Trends between Single Metal Atoms and Oxide Supports Identified with Density Functional Theory and Statistical Learning," *Nature Catalysis,* 1, 531-539. (82,90,91]

Ouyang, R., Curtarolo, S., Ahrnetcik, E., Scheffler, M., and Ghiringhelli, L. M. (2018), "SISSO: A Compressed-Sensing Method for Identifying the Best Low-Dimensional Descriptor in an Immensity of Offered Candidates:' *Physical Review Materials,* 2, 083802. (82,83,88]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso:' *Journal of the Royal Statistical Society,* Series B, 58, 267-288. (82,85]

Wang, A., Li, J.,and Zhang, T. (2018), "Heterogeneous Single-Atom Catalysis:' *Nature Reviews Chemistry,* 2, 65-81. (90]

Yang, X.-F., Wang, A., Qiao, B., Li, J., Liu, J., and Zhang, T. (2013), "Single-Atom Catalysts: A New Frontier in Heterogeneous Catalysis; *Accounts of Chemical Research,* 46, 1740-1748. (90]

Zhong, M., Tran, K., Min, Y., Wang, C., Wang, Z., Dinh, C.-T., De Luna, P., Yu, Z., Rasouli, A. S., Brodersen, P., Sun, S., Voznyy, 0., Tan, C.-S., Askerka, M., Che, F., Liu, M., Seifitokaldani, A., Pang, Y., Lo, S.-C., Ip, A., Ulissi, Z., and Sargent, E. H. (2020),"Accelerated Discovery of CO2 ElectrocatalystsUsing Active Machine Learning," *Nature,*581, 178-183. (81]