Property Prediction of Crystalline Solids from Composition and Crystal Structure

Author A[*], Author B[†]

Recently Accepted for Publication in the AIChE Journal

Abstract

We propose using kernel regression as a data-driven and rigorous nonparametric statistical technique to predict properties of atomic crystals. A key feature of the proposed approach is the possibility of treating predictors not only as continuous, but also as categorical data. The latter specifically allows the predictive model to capture the discrete nature of crystals with regards to composition (number of atoms in the chemical formula) and spatial configuration (finite number of crystallographic space groups). Another important aspect of using kernel regression is the direct access to its explicit mathematical form, which can be directly embedded in optimal inverse problems to design new crystalline materials with given target properties. The property prediction approach is illustrated by training models to predict electronic properties of 746 binary metal oxides and elastic properties of 1,173 crystals. As a first approach to solving the inverse problem, we describe an exhaustive enumeration algorithm.

Keywords: Crystal Property Prediction, Data Analytics, Kernel Regression, Crystal Composition and Structure, Exhaustive Enumeration Algorithm

1 Introduction

Predicting properties of compounds has received considerable attention across different disciplines and has seen applications in diverse areas. The approach proposed in this paper is influenced by the advances in Group Contribution Methods (GCMs) as will be explained later in this section. GCMs are typically regression-based methods that have been successfully used in chemistry, biochemistry, and biochemical/chemical/process engineering applications to predict thermodynamic and transport properties of chemical compounds based on their molecular structure. The basic idea of GCMs is to use the contribution of *functional groups* (and possibly their interactions) in the molecular structure to the prediction of a given property. Examples of predicted properties are critical properties (temperature, pressure, and

[*]University of Author A
[†]University of Author B

volume), heat capacity, and viscosity. Some notable examples of GCMs include the Lydersen method (Lydersen, 1955), UNIQUAC Functional-group Activity Coefficients (UNIFAC) (Fredenslund, Jones, and Prausnitz, 1975), Klincewicz method (Klincewicz and Reid, 1984), Joback method (Joback and Reid, 1987), Mavrovouniotis method (Mavrovouniotis, 1990), Constantinou-Gani method (Constantinou and Gani, 1994) among others.

A key characteristic of GCMs in general is that the property descriptors (i.e., predictors) are explicitly represented by the molecular structure (and chemical composition) of a compound. For instance, the propanone (acetone) molecule (C_3H_6O) has two methyl groups ($-CH_3$) and one ketone group (>C=O, non-ring). The *forward* problem (i.e., property prediction) uses the contributions of these two functional groups to estimate properties. One can pose an optimal *inverse* (design) problem, which attempts to obtain the original molecule (structure and composition) given a target property value. This is the main goal of the research area called Computer-Aided Molecular/Mixture Design (CAMD) as exemplified by several works in the literature (Odele and Macchietto, 1993; Maranas, 1996; Karunanithi, Achenie, and Gani, 2005; Eljack et al., 2007; Samudra and Sahinidis, 2013).

The main focus of the works on CAMD has been on designing organic molecules used as solvents and extraction agents, refrigerants, and pharmaceutical compounds. In this paper, we focus on *crystalline solids* and propose a regression-based property prediction approach, which uses explicit information about the chemical composition and crystal structure as contributing "groups", thus being amenable to an inverse problem formulation. Next, we present an overview of the literature on property prediction and design of crystals.

The advance of computing, and the development and continuous improvement of *ab initio* quantum chemistry theory and calculations, as well as statistical (or machine learning) methods have greatly influenced research activities related to materials discovery and design (for an excellent recent review, see Kalidindi and De Graef (2015)). This synergy is usually termed high-throughput (HT) computational materials design (Curtarolo et al., 2013). The areas of application of computational HT research include: thermodynamics for the identification of binary and ternary compounds, solar materials, water splitting using sunlight, carbon capture and gas storage, nuclear detection and scintillators, topological insulators, piezoelectrics, thermoelectric materials, materials for catalysis, and battery materials for energy storage.

Several research groups have proposed statistical (machine learning) approaches for property prediction of crystalline solids. Willighagen et al. (2007) used supervised self-organizing maps (a type of artificial neural network) to explore large numbers of crystal structures and to visualize structure-property relationships. Saad et al. (2012) investigated both unsupervised and supervised machine learning techniques to predict structure and properties of crystals with chemical formula AB; the predictors used include the number of valence electrons, electronegativity, boiling point, first ionization potential among others. Ma et al. (2015) developed a machine-learning-augmented model based on artificial neural networks (ANNs) that captures nonlinear adsorbate-substrate interactions; the approach was applied to the electrochemical reduction of carbon dioxide on metal electrodes, and the predictors in the ANN model included characteristics of the d-states distribution, ionization potential, electron affinity among others. Ghiringhelli et al. (2015) used predictors such as nuclear numbers, and other primary descriptors related to energy levels and radii of valence orbitals, as well as algebraic expressions involving them, to predict energy differences in *ab initio*

calculations of crystals with two different crystal structures. Isayev et al. (2015) developed predictive quantitative materials structure-property relationship models, which were used for predicting the critical temperatures of known superconductors. However, while these methods have demonstrated to be efficient at predicting several properties of diverse crystals, it is unclear how and if they can also be directly used in an *inverse problem* setting (materials design), i.e., given a target property, obtain the best or a pool of the best materials whose predicted property matches the specified value within certain tolerance. This represents the main motivation for this work: to use descriptors in predictive models that allow us to directly retrieve the original materials (i.e., their composition and crystal structure) after solving the design problem.

This paper is organized as follows. First, we state the problem under consideration and introduce some notation that is used throughout the paper. We then present the proposed property prediction method, which is based on kernel regression with categorical predictors. Two data sets with crystal properties estimated via *ab initio* calculations are used to illustrate the proposed property prediction method, and an exhaustive enumeration algorithm is presented as a first approach to solving the inverse problem. Finally, conclusions are drawn on the accuracy of the proposed method, and future work on the formulation of an optimal inverse problem is discussed.

2 Problem Statement

We consider a given data set in the form of a table with $i = 1, 2, \ldots, n$ data points (or rows) of the properties to be predicted and predictors. Each row of the table contains data for a crystal. Let $j \in J$ denote the indexed set of properties, and $p \in P$ the indexed set of predictors. In addition, partition the data set table into two tables, one with property data ($Y_{i,j}$) and one with predictor data ($X_{i,p}$). The objective is to obtain a kernel regression model for each property j by regressing its data $Y_{i,j}$ on p predictors whose data are given by the table $X_{i,p}$.

For example, suppose we are interested in predicting the total energy and formation energy per atom of binary metal oxides. We need to choose the predictors for the regression model. In the proposed approach, as will be discussed in the next section, we consider the space group number and the number of atoms in the chemical formula of each crystal in the data set. In this case, the indexed sets can be written as follows: $J = \{\text{energy}, \text{formation\_energy\_atom}\}$ and $P = \{sg, \text{Li}, \text{Na}, \text{K}, \ldots, \text{O}\}$, where sg stands for the space group number. Table 1 illustrates this example with data obtained from the Materials Project database (Jain et al., 2013). Note that the data for the predictors are in fact integer numbers, i.e., they are categorical or discrete. This will be considered in choosing the kernel regression model discussed in the next section.

Table 1: Example of data set and corresponding notation.

(a) Properties ($Y_{i,j}$)

i	Formula	energy (eV)	formation_energy_atom (eV)
1	Li_2O	−14.264	−2.071
⋮	⋮	⋮	⋮
11	Na_2O_2	−8.401	−1.312
⋮	⋮	⋮	⋮
214	TiO_2	−26.902	−3.512
⋮	⋮	⋮	⋮

(b) Predictors ($X_{i,p}$)

i	sg	Li	Na	⋯	Ti	⋯	O
1	225	2	0	⋯	0	⋯	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
11	189	0	2	⋯	0	⋯	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
214	141	0	0	⋯	1	⋯	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

3 Methods

3.1 Property Prediction Model

Before presenting the kernel regression model, we review the basics of regression analysis (Montgomery and Runger, 2003). The regression model is generally written as,

$$Y = f(X) + \epsilon \tag{1}$$

where $f(\cdot)$ is a mathematical formula (called *regression function*) that expresses the relationship between X (input variable, predictor, covariate) and Y (output variable, response), and ϵ is the random error term assumed to have mean zero, homoskedastic (i.e., its variance is constant over the range of X values), and uncorrelated with X. In the context of this work, Y corresponds to a property of interest of a crystal, and X represents a predictor of such property. In this brief overview of regression analysis, and for ease of exposition, we consider that the random variables X and Y each represent 1-D data.

A typical and parametric choice for the regression function is a linear function with parameters β_0 and β_1, i.e., $f(X) = \beta_0 + \beta_1 X$. The goal of a linear regression analysis is to estimate β_0 and β_1 such that the predictions of the model are as close as possible to the observed data of the response variable.

Nonparametric regression models, in contrast, are more general and rely on few assumptions about the underlying populations from which the data are obtained (Li and Racine,

2007; Hollander, Wolfe, and Chicken, 2014). We have chosen kernel regression for its flexibility to handle both continuous and categorical data, and ability to operate on multivariate data. Appendix B contains a brief overview on kernel regression.

In order to predict a desired property of a crystal, we propose using the following predictors:

- 3-D crystallographic space group number (integers between 1 and 230), and
- Number of atoms of each element in the chemical formula.

The space group number indicates the spatial configuration of the atoms in the periodic structure of a crystal. The remaining predictors are the number of atoms contained in the chemical formula. Therefore, the chosen predictors contain information of the crystal structure and its chemical composition. Throughout this paper, we restrict the approach to atomic crystals. The main motivation behind this choice of predictors is to allow posing and solving optimal inverse problems whose output is the best crystal (with structure and composition) for which the predicted properties satisfy given targets. We note that, in the context of an inverse problem (materials design), the design variable space group number would correspond to the *suggested* arrangement of the atoms in the structure with given predicted property. In other words, the approach would not perform geometry or structure optimization.

Since all predictors are in fact categorical, one has to choose an appropriate kernel that is designed for discrete data. We propose using the normalized kernel introduced by Racine and Li (Racine and Li, 2004). The property prediction model proposed in this work is given in equation (2),

$$\tilde{Y}_j = \frac{\sum_{i=1}^{n} Y_{i,j} \prod_{p \in P} \lambda_{p,j}^{z_{i,p}}}{\sum_{i=1}^{n} \prod_{p \in P} \lambda_{p,j}^{z_{i,p}}} \qquad \forall\, j \in J \tag{2}$$

where $\tilde{Y}_j$ is the predicted value of property j, $Y_{i,j}$ is the value of property j of data point i, $\lambda_{p,j} \in [0,1]$ is the (optimal) kernel bandwidth of predictor p obtained for property j, and $z_{i,p}$ is an indicator variable defined as follows,

$$z_{i,p} = \begin{cases} 1, & \text{if } X_{i,p} \neq x_p \\ 0, & \text{otherwise} \end{cases} \qquad \forall\, i = 1, \ldots, n,\ p \in P \tag{3}$$

where $X_{i,p}$ is the value of predictor p of data point i, and x_p represents a given value that the predictor p can take from its *support*. For example, suppose that for all data points in the data set, the predictor $p = \text{Li}$ appears with either 0, 1, or 2 atoms in the chemical formula. Therefore, the support for this predictor is the set $\{0, 1, 2\}$. For the sake of the example, let $x_{\text{Li}} = 2$. Thus, $z_{i,\text{Li}} = 1$ for every data point i such that $X_{i,\text{Li}} \neq x_{\text{Li}} = 2$, and $z_{i,\text{Li}} = 0$, otherwise.

Remark 1. The optimal vector of bandwidths for property j, $\lambda_{p,j}$, $\forall\ p \in P$, can be obtained via least-squares cross-validation (originally proposed in Rudemo (1982)), i.e., via

minimization of the mean squared errors (nonlinear least squares) between observed property data and its predicted values as shown in equation (4),

$$\min_{\lambda_{p,j}\in(0,1]} n^{-1}\sum_{i=1}^{n}\left[Y_{i,j}-\frac{\sum_{i'\neq i}Y_{i',j}\prod_{p\in P}\lambda_{p,j}^{z_{i,i',p}}}{\sum_{i'\neq i}\prod_{p\in P}\lambda_{p,j}^{z_{i,i',p}}}\right]^2 \tag{4}$$

where $z_{i,i',p}=1$ if $X_{i,p}\neq X_{i',p}$, and 0 otherwise. The summation over data points $i'\neq i$ yields the leave-one-out estimator (Li and Racine, 2007) of the regression function for property j.

Remark 2. The optimization problem in equation (4) can be solved with standard local nonlinear optimization solvers, such as CONOPT (Drud, 1994), KNITRO (Byrd, Nocedal, and Waltz, 2006), and IPOPT (Wächter and Biegler, 2006), or global solvers, such as BARON (Tawarmalani and Sahinidis, 2005) and SCIP (Achterberg, 2009). In this paper, we use the routines for implemented in the `np` package Hayfield and Racine (2008) in the R programming language R Core Team (2015). The routine `npreg`, which computes optimal bandwidths for both continuous and discrete data, uses a local solver in a multi-start procedure to increase the likelihood of converging to a better local solution. The `np` package offers many more routines for nonparametric statistics.

Remark 3. Predicting from a kernel regression model requires not only the bandwidths obtained with methods such as the ones discussed in the previous remarks, but also the entire data set used in the training step. Note that the evaluation of the right-hand side of equation (2) uses the values of the response data for property j ($Y_{i,j}$) and of $z_{i,p}$, which in turn depends on the data set of predictors ($X_{i,p}$) and the new value(s) of the predictor(s) (x_p). This is in contrast with parametric regression models, in which only the estimated parameters and the new value(s) of the predictor(s) are needed.

3.2 Materials Project Database and Test Cases

The kernel regression property prediction model was tested on data obtained from the Materials Project (MP) database (Jain et al., 2013), which is a program of the Materials Genome Initiative that uses high-throughput computing to uncover the properties of all known inorganic materials. The materials properties are calculated using quantum chemical approximations such as Density Functional Theory (DFT) (Hohenberg and Kohn, 1964; Kohn and Sham, 1965). Data acquisition step was performed using the Materials Application Programming Interface (MAPI), which is based on the REpresentational State Transfer (REST) (Ong et al., 2015).

The proposed property prediction model was evaluated on the following two test cases.

- *Electronic Properties of Metal Oxides*: band gap and Fermi level of 746 binary metal oxides. Briefly, the band gap is the difference in energy between the valence and conduction bands in a solid; the Fermi level is the total chemical potential for electrons, and can be interpreted as the energy level with 50% chance of being occupied at finite temperature.

- *Elastic Properties*: mechanical properties, such as the bulk and shear moduli according to the Voigt and Reuss averages of 1,173 crystals (for details, see de Jong et al. (2015)). Other elastic properties, such as isotropic Poisson ratio and universal elastic anisotropy can be calculated from those moduli.

3.3 Prediction Performance Metrics

The following measures of fit are used in this paper, where $Y_{i,j}$ and $\tilde{Y}_{i,j}$ are the observed and predicted values of data point i and property j, respectively.

- Root Mean Squared Error (RMSE)

$$\mathrm{RMSE}_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_{i,j} - \tilde{Y}_{i,j})^2} \qquad \forall\, j \in J \tag{5}$$

- Mean Absolute Error (MAE)

$$\mathrm{MAE}_j = \frac{1}{n}\sum_{i=1}^{n}|Y_{i,j} - \tilde{Y}_{i,j}| \qquad \forall\, j \in J \tag{6}$$

- Mean Absolute Percentage Error (MAPE)

$$\mathrm{MAPE}_j = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_{i,j} - \tilde{Y}_{i,j}}{Y_{i,j}}\right| \qquad \forall\, j \in J \tag{7}$$

If $Y_{i,j} = 0$ for any i and j, then MAPE_j is not calculated. Note that both RMSE_j and MAE_j have the same units of property j, whereas MAPE_j is a percentage quantity.

3.4 Software and Hardware Specifications

All kernel regression computations were carried out using the `npreg` function of the `npRmpi` package, which is a parallel implementation of the R (R Core Team, 2015) package `np` (Hayfield and Racine, 2008). It can exploit the presence of multiple processors and the Message Passing Interface (MPI) approach for parallel computing to reduce the computational run time associated with kernel methods. Scripts and data sets are provided as Supplementary Information material.

Computational experiments were run on a Ubuntu Server 15.04 virtual machine (VM) enabled by Google Compute Engine (GCE), which is a component of the Google Cloud Platform (`https://cloud.google.com/compute/`). The VM contains 32 Intel® Xeon® CPUs at 2.30 GHz and 120 GB of RAM.

4 Results and Discussions

In addition to the target properties for each test case (conductivity and elasticity), we also estimated basic properties (total energy, formation energy per atom, and density) obtained from Density Functional Theory (DFT) calculations in the Materials Project database. The optimal bandwidths used in the predictions were the best obtained with 5 multi-start optimization runs (default option in the `npreg` function).

4.1 Electronic Properties of Metal Oxides

Table 2 shows the prediction performance metrics of basic and electronic properties of 746 binary metal oxides. The last column in the table contains the total execution time for obtaining optimal bandwidths with 5 multi-start runs. The optimal bandwidths of the best multi-start run are provided as Supplementary Information. The accuracy of the kernel regression model is in general very high as implied by the low values for the performance metrics.

Table 2: Prediction performance of basic and electronic properties for various metal oxides. An entry with "–" indicates that at least one response data point for a property is zero, thus the MAPE was not calculated.

Property	RMSE	MAE	MAPE (%)	Wall Time (s)
Total Energy (eV)	0.6531	0.1392	0.32	1,862.11
Formation Energy per Atom (eV)	0.0024	0.0011	0.11	2,464.41
Density (g cm^{-3})	0.2546	0.1397	3.20	2,522.03
Band Gap (eV)	0.0528	0.0126	–	1,294.47
Fermi Level (eV)	0.2432	0.1183	10.09	1,466.84

The performance of the proposed property prediction model is also illustrated in Figure 1, which shows actual response data vs. predicted property values. Overall, the prediction of band gaps is in very good agreement with the DFT-calculated data. The three data points with the largest prediction residuals (positive or negative) were ThO_2 with space group #225 (residual = actual − predicted = 0.9962 eV), Rb_2O with space group #225 (residual = 0.8419 eV), and Al_2O_3 with space group #2 (residual = −0.2167 eV). The relatively lower accuracy of Fermi level predictions, as evidenced by the higher MAPE value, is indicated by the larger number of circles that do not lie on the 45° line. The kernel regression model tends to overestimate negative Fermi levels. The maximum and minimum residuals were for the same metal oxide, TiO_2, with space group numbers 14 (residual = 1.2735 eV) and 35 (residual = −1.7507 eV).

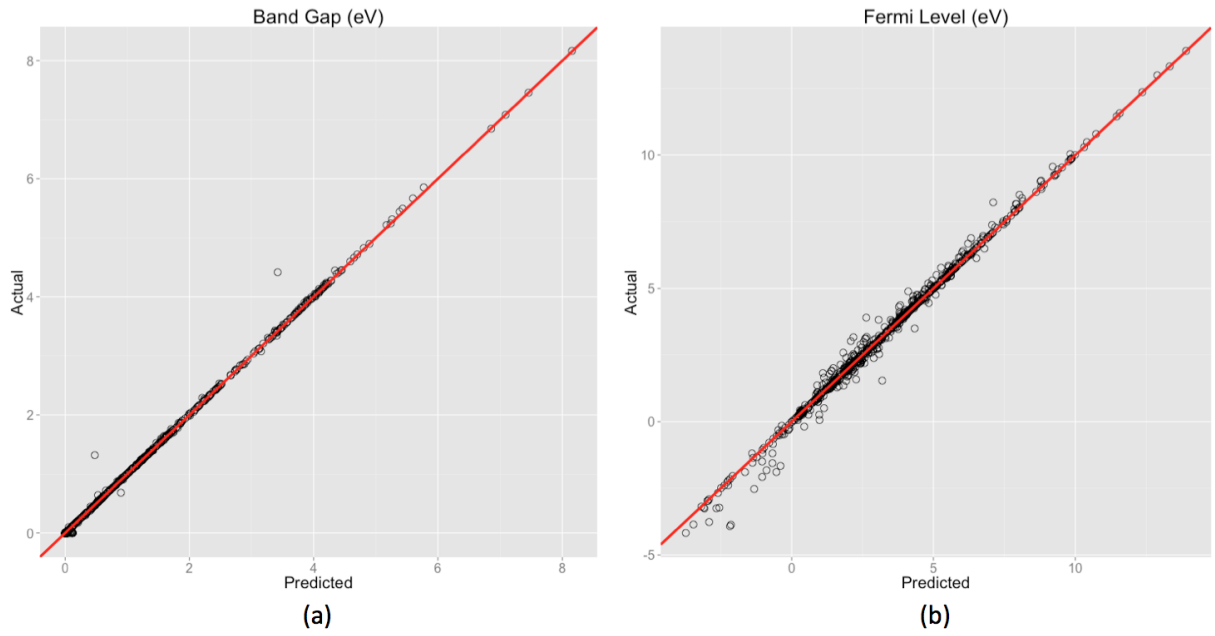

Figure 1: Actual response data vs. predicted values of (a) band gap and (b) Fermi level.

4.2 Elastic Properties

Table 3 shows the prediction performance metrics of basic and elastic properties of 1,173 crystals. The last column in the table contains the total execution time for obtaining optimal bandwidths with 5 multi-start runs. Note that the times are longer than for the previous set of crystals, which is due to the larger number of data points and predictors. The optimal bandwidths of the best multi-start run are provided as Supplementary Information. Note that this data set contains very diverse crystals, with one through four elements in the composition, and a total of 63 chemical elements used as predictors belonging to every group in the period table (except noble gases). This diversity in the data set is likely to have yielded relatively lower accuracy of the kernel regression model in predicting the total energy, even though the prediction of the remaining properties exhibited high accuracy.

Table 3: Prediction performance of basic and elastic properties for various crystals, where K_V and G_V are the bulk and shear moduli Voigt average, respectively, and K_R and G_R are the respective moduli according to the Reuss average. An entry with "–" indicates that at least one response data point for a property is zero, thus the MAPE was not calculated.

Property	RMSE	MAE	MAPE (%)	Wall Time (s)
Total Energy (eV)	3.7988	0.9049	7.34	4,158.17
Formation Energy per Atom (eV)	0.0132	0.0053	–	5,004.54
Density (g cm^{-3})	0.2938	0.1285	2.41	3,921.21
K_V (GPa)	2.1823	0.7046	0.80	4,457.92
G_V (GPa)	1.0373	0.3373	1.01	2,726.10
K_R (GPa)	1.2713	0.3811	0.57	4,920.23
G_R (GPa)	0.7118	0.2316	1.00	3,876.37

Figure 2 shows actual response data vs. predicted values of the elastic properties. In general, the prediction of band gaps is in very good agreement with the DFT-calculated data.

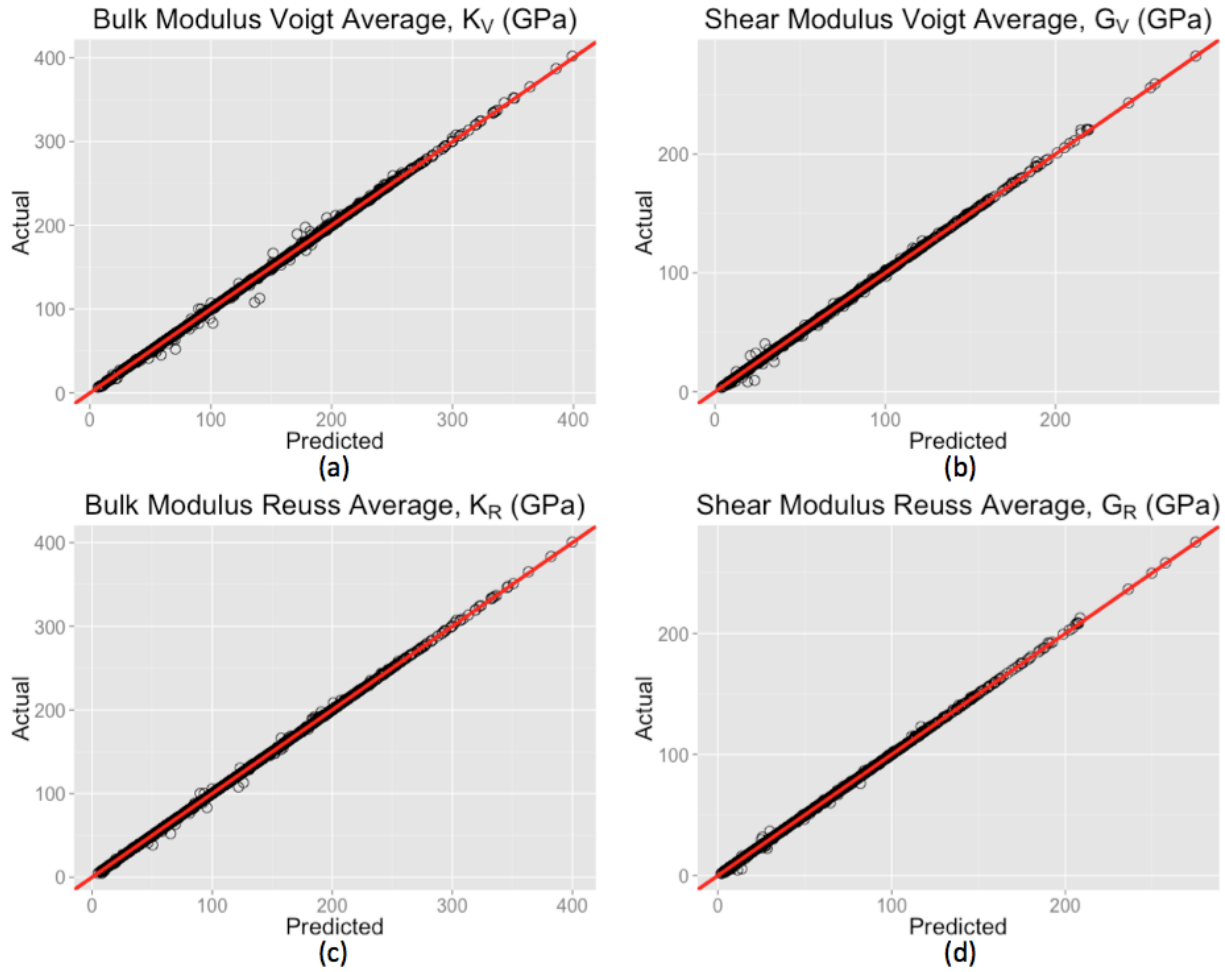

Figure 2: Actual response data vs. predicted values of (a) bulk and (b) shear moduli Voigt average, and (c) bulk and (d) shear moduli Reuss average.

4.3 Inverse Problem via Exhaustive Enumeration

A special feature of the proposed kernel regression method to model the relationship between composition and crystal structure (inputs) and crystal properties (output) is the treatment of the predictors as categorical or discrete data. This fact gives a *combinatorial* aspect to the inverse problem, which can be cast as a mathematical optimization problem similarly to the approach taken in the computer-aided molecular/mixture design literature (Achenie, Gani, and Venkatasubramanian, 2003).

Perhaps the simplest approach to solve the combinatorial inverse problem to obtain the crystal (or a pool of crystals) whose properties match pre-specified values is the Exhaustive Enumeration (EE) method. The principle of this method is to evaluate all combinations

of the discrete variables, which in this work correspond to $z_{i,p}$ (see equations (2) and (3)). Therefore, a general EE method would entail predicting properties for every combination of the predictors, i.e., space group numbers and number of atoms of each chemical element.

A shortcoming of EE is the large number of evaluations needed to cover the entire combinatorial search space of discrete variables. In this particular problem, however, there are some restrictions on the combinations of predictors. For instance, the predictor of space group number must always be present in an evaluation. In addition, at least one chemical element with more than 0 atoms (i.e., $X_{i,p} > 0$) must be present in any evaluation. Even though these inherent *constraints* restrict the search space, the computational cost may be considerable or even prohibitive depending on the number of data points and predictors.

Before presenting the EE algorithm and the results, we will introduce some notation and an illustrative example. First, note that the table of predictor data, $X_{i,p}$, may contain several repeated entries for the same predictor. For example, there may be a few different crystals with the same space group number. Also, each row may contain several zeros, because, in general, only a small fraction of the chemical elements are present in the composition of a crystal. In the case of the set of metal oxides data, each row of $X_{i,p}$ contains zeros for all metals except one and the oxygen atom. This motivates the definition of the following binary matrix in equation (8) that has the same dimensions as $X_{i,p}$.

$$\bar{X}_{i,p} = \begin{cases} 1, & \text{if the pair } (i,p) \text{ is } \textit{unique} \\ 0, & \text{otherwise} \end{cases} \qquad \forall\, i = 1, \ldots, n,\ p \in P \tag{8}$$

The meaning of "unique" in the definition of $\bar{X}_{i,p}$ is as follows. For a given predictor p, if two data points i and i' are such that $X_{i,p} = X_{i',p}$, then either the pair (i,p) or the pair (i',p) is considered unique. For example, if $X_{1,sg} = X_{2,sg} = 225$, where sg represents the space group number predictor, then either $\bar{X}_{1,sg} = 1$ and $\bar{X}_{2,sg} = 0$, or vice versa.

In order to clarify the definition of $\bar{X}_{i,p}$, consider the following illustrative example. Consider a data set with 4 crystals and respective space group numbers: Li_2O (#225), Li_2O_2 (#194), Na_2O (#225), NaO_3 (#44). The tables/matrices $X_{i,p}$ and $\bar{X}_{i,p}$ in this example are represented in Figure 3. Note that we adopted a convention in the construction of $\bar{X}_{i,p}$: without loss of generality, the first row of this matrix contains all entries equal to 1, i.e., all pairs (i,p) for $i = 1$ are considered unique by default. We reemphasize that only valid combinations of crystals are evaluated in the EE algorithm, i.e., each combination contains space group numbers and numbers of atoms that are present in the original data set, thus respecting the support of each predictor. In this illustrative example, the support of sg is $\{44, 194, 225\}$, Li is $\{0, 2\}$, of Na is $\{0, 1, 2\}$, and of O is $\{1, 2, 3\}$.

	sg	Li	Na	O
Li_2O (#225)	225	2	0	1
Li_2O_2 (#194)	194	2	0	2
Na_2O (#225)	225	0	2	1
NaO_3 (#44)	44	0	1	3

$X_{i,p}$

sg	Li	Na	O
1	1	1	1
1	0	0	1
0	1	1	0
1	0	1	1

$\bar{X}_{i,p}$

Figure 3: Table of predictor data, $X_{i,p}$, and corresponding matrix of unique entries, $\bar{X}_{i,p}$, for the illustrative example of the exhaustive enumeration algorithm.

The use of $\bar{X}_{i,p}$ significantly reduces the number of combinations in the EE algorithm, since only the combinations of unique pairs (i, p) must be evaluated. We will focus on the case of metal oxides, which puts additional restrictions on the search space by enforcing that an evaluation must include a space group number, a metal (one or more atoms), and oxygen (again, one or more atoms). Note that the definition of $\bar{X}_{i,p}$ is general, and it can be used for data of crystal in general.

The pseudo-code of the EE algorithm is given in Algorithm 1. Note the use of conditional (`if`) statements (lines 3, 7, and 10) to restrict the prediction of properties for only the unique cases (i.e., $\bar{X}_{i,p} = 1$). To illustrate the EE algorithm, consider the example from Figure 5.3. In the first pass of the algorithm, as one reaches line 11, the running indices are $(i, i', i'') = (1, 1, 1)$ and $p = \{\text{Li}\}$; thus, the crystal whose properties are evaluated in the first pass is Li_2O (#225), where we ensure that all other metals (in this example, only Na) have number of atoms equal to 0. Following the same rationale, the next set of running indices are $(i, i', i'') = (1, 1, 2)$ and $p = \{\text{Li}\}$, which corresponds to the crystal Li_2O_2 (#225), and so on. The algorithm is amenable to parallelization, since it contains independent `for` loops. We implemented the algorithm in the R programming language (see Supplementary Information) and parallelized the first `for` loop using the R `foreach` package (Revolution Analytics and Weston, 2014).

The EE algorithm was applied to the data set of electronic properties of metal oxides. The total execution time was 71,097.17 seconds (approximately 19.7 hours), and the total number of unique combinations (i.e., the size of the output list of predicted properties) was 1,153,504. This serves as a practical example of the computational expense of the EE algorithm, which motivates future work on the development of effective optimization-based strategies to handle the combinatorial aspect of the inverse (design) problem.

After obtaining the list of the property predictions for all possible unique combinations, we only have to iterate over the list and select the entry(ies) that match given specifications within certain tolerance. Some results (crystal composition and space group number) are shown in Table 4. Depending on the specifications bounds and the tightness of the tolerances, several crystals or none may satisfy the targets. Here, we only show a few of the crystals that exhibited the smallest absolute deviations from the tolerances when applicable. The results show that we can not only retrieve crystals whose properties have been previously calculated or measured, but also discover new crystals whose predicted properties match the given specifications. It is important to keep in mind that the kernel regression model has

Input : List with names of properties and matrices of predictor data $X_{i,p}$ and unique combinations $\bar{X}_{i,p}$
Output: List of predictions for all unique combinations for each property

```
for i from 1 to n                                   /* Space group number loop */
do
    if X̄_{i,sg} = 1 then
        for i′ from 1 to n                                        /* Metal loop */
        do
            foreach p ∈ P \ {sg, O} do
                if X̄_{i′,sg} = 1 and X_{i′,p} ≠ 0 then
                    for i″ from 1 to n                           /* Oxygen loop */
                    do
                        if X̄_{i″,O} = 1 then
                            /* Current data point indices:                    */
                            /* Space group number i, Metal i′, Oxygen i″      */
                            Predict the properties j ∈ J of the current metal oxide;
                            Store predictions in the output list;
                        end
                    end
                end
            end
        end
    end
end
```

Algorithm 1: Pseudo-code of the Exhaustive Enumeration (EE) algorithm to evaluate the properties of all *unique* combinations of metal oxides.

some inherent prediction error, which may prevent retrieving a crystal originally present in the data set given its property values.

5 Conclusions

In this paper, we proposed the use of a nonparametric statistical technique, namely kernel regression with categorical or discrete data, to model the relationship between properties (response) and chemical composition and crystal structure (predictors) of crystalline solids. The discrete nature of the predictors is explicitly accounted for in building kernel regression models. The proposed method was tested on two data sets with diverse crystals and different predicted properties. In the data set of electronic properties of metal oxides, the kernel regression models obtained resulted in overall low (at most 10.1%) mean absolute percentage error (MAPE) values. In the data set of elastic properties of several diverse crystals, the MAPE values were less than 7.5%.

Table 4: Results (crystal composition and space group number) of the Exhaustive Enumeration algorithm for some given specifications. Crystals that are not present in the original data set (i.e., potential new crystals) are marked with an asterisk.

Specifications	**Crystals**
$\lvert\tilde{Y}_{\text{band\_gap}} - 1.23\rvert \leq 10^{-2}$	Bi_2O_3 (#197), Mn_7O_{25} (#164)*
$\min\ \tilde{Y}_{\text{band\_gap}}$ $0.5\,\text{eV} \leq \tilde{Y}_{\text{efermi}} \leq 1.5\,\text{eV}$	Rb_9O_{13} (#167)*
$\tilde{Y}_{\text{band\_gap}} \leq 2\,\text{eV}$ $\lvert\tilde{Y}_{\text{efermi}} - 1.5\,\text{eV}\rvert \leq 10^{-2}$	$Co_{21}O_{22}$ (#141)*, Tb_8O_{21} (#130)*, CsO_{48} (#185)*

An important advantage of the proposed property prediction method to be used in an inverse (design) problem is the choice of predictors, which allow for the direct retrieval of a material (i.e., its chemical composition and crystal structure). We described an exhaustive enumeration algorithm that solves the inverse problem by evaluating all possible combinations of crystals from the data set. The computational expense serves a motivation to the development of an optimization-based inverse problem that can efficiently handle the combinatorial aspect of selecting the best material that meets given target property values. This is the subject of on going work and future publication.

Acknowledgments

The first author is thankful to Dr. Marc De Graef (Department of Materials Science and Engineering at Carnegie Mellon University) for the discussions about crystallography.

Appendix A Overview of Crystal Structure

There are standard references on the crystalline structure of materials (Rohrer, 2004; Hahn, 2005; Tilley, 2006; De Graef and McHenry, 2012) as well as their properties (Newnham, 2005; White, 2011). In this section, we provide a brief overview of the main concepts that are relevant to the nonparametric regression model for property prediction proposed in this work.

In simple terms, a crystal structure is a regular and periodic arrangement of atoms or molecules. The arrangement or *pattern* is composed of a *motif* that is repeatedly placed at each point of a net or *lattice* (see Figure A.1 for a 2-D example). For an atomic crystal, the motif is represented by an atom, and its average position in the structure does not change with time. Therefore, more formally, we say that a crystal structure is a time-invariant, three-dimensional arrangement of atoms or molecules on a lattice.

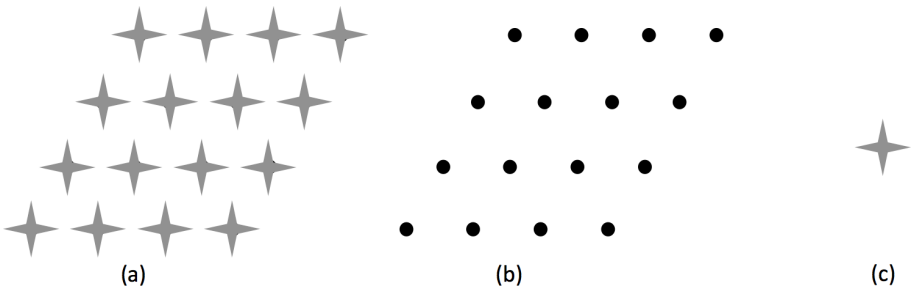

Figure A.1: Schematic of a 2-D crystal structure: (a) periodic pattern or structure, (b) lattice or net, and (c) motif.

A lattice is a mathematical abstraction whose points reflect the *translational symmetry* of the complete crystal. Using algebraic and symmetry arguments, it can be shown that there are only five 2-D lattices and fourteen 3-D lattices. They are called the Bravais lattices of which two are illustrated in Figure A.2. It is important to notice the *discrete* nature of lattices not only in terms of the position of points in space, but also in the number of possible arrangements they can yield in a crystal structure.

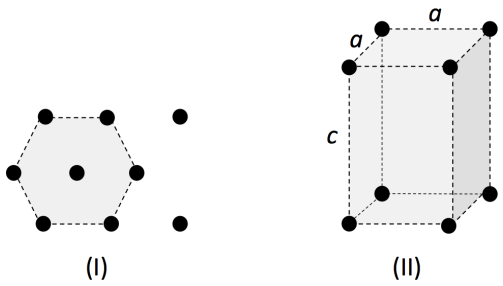

Figure A.2: Two Bravais lattices in 2-D and 3-D: (I) hexagonal or rhombic, and (II) tetragonal (primitive) where $a \neq c$.

The Bravais lattices in 2-D and 3-D are combined with the crystallographic point groups to form a finite number of plane groups (2-D) and space groups (3-D). Briefly, a point group is a set of symmetry operations (e.g., rotations or reflections) that leaves one or more points in space unmoved. It can be shown that there are 14 plane groups and 230 space groups. Figure A.3 illustrates the crystal structure of TiO_2 anatase.

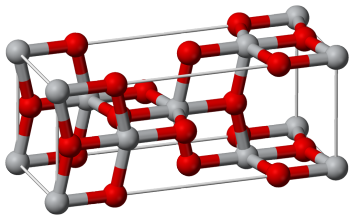

Figure A.3: Crystal structure of TiO_2 anatase. Space group: #141. Crystal system: tetragonal.

Appendix B Overview of Kernel Regression

Consider the following general regression model,

$$Y = f(X) + \epsilon \tag{B.1}$$

where Y and X are output and input *continuous* random variables, respectively, $f(\cdot)$ is the regression function, and ϵ is a random error term. Kernel regression is a nonparametric statistical technique to estimate $f(\cdot)$.

A typical choice for the mathematical expression of the *kernel estimator* of the regression function is the Nadaraya-Watson kernel (Nadaraya, 1964; Watson, 1964),

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)} \tag{B.2}$$

where Y_i and X_i are data points of the output and input variables, respectively, and $K(\cdot)$ is the kernel function with bandwidth $h > 0$, which is a smoothing parameter. Figure B.1 illustrates the output of a kernel regression analysis.

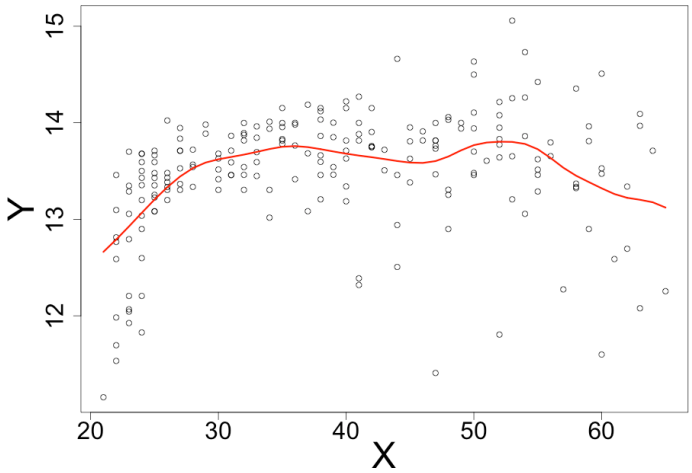

Figure B.1: Illustration of kernel regression, where the estimated regression function is represented by the smooth line.

For multivariate data (multiple predictors), a common approach is to use the *product kernel*, which is simply the product of individual (univariate) kernels for each predictor. The multivariate version of equation (B.2) is as follows,

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} Y_i \prod_{p \in P} K\left(\frac{x_p - X_{i,p}}{h_p}\right)}{\sum_{i=1}^{n} \prod_{p \in P} K\left(\frac{x_p - X_{i,p}}{h_p}\right)} \tag{B.3}$$

where P is the indexed set of predictors.

Different kernel functions may be used (e.g., Gaussian, Epanechnikov (Epanechnikov, 1969)), but choosing the bandwidth is more important for the accuracy of the kernel estimator. A general and rigorous approach is via least-squares cross-validation (Li and Racine,

2007), where the following optimization problem is solved,

$$\min_{h_p>0} n^{-1} \sum_{i=1}^{n} \left[Y_i - \frac{\sum_{i' \neq i} Y_{i'} \prod_{p \in P} K\left(\frac{x_p - X_{i,p}}{h_p}\right)}{\sum_{i' \neq i} \prod_{p \in P} K\left(\frac{x_p - X_{i,p}}{h_p}\right)} \right]^2 \tag{B.4}$$

where the summation over data points $i' \neq i$ yields the leave-one-out estimator of the regression function. This cross-validation approach has the advantage that it can automatically handle the possibility of the existence of *irrelevant* input variables. This can happen when the output variable and the *relevant* input variables are independent of the other (irrelevant) predictors.

Lastly, kernel regression can also be used with discrete and mixed data (continuous and discrete). In this case, a kernel function for discrete or categorical data is also used (e.g., Aitchison-Aitken (Aitchison and Aitken, 1976), Racine-Li (Racine and Li, 2004)). Let the superscripts c and d denote continuous and discrete predictors, respectively. The multivariate, mixed kernel estimator can then be written as follows,

$$\hat{f}(x^c, x^d) = \frac{\sum_{i=1}^{n} Y_i \prod_{p \in P^c} K\left(\frac{x_p^c - X_{i,p}^c}{h_p}\right) \prod_{p \in P^d} L\left(X_{i,p}^d, x_p^d, \lambda_p\right)}{\sum_{i=1}^{n} \prod_{p \in P^c} K\left(\frac{x_p^c - X_{i,p}^c}{h_p}\right) \prod_{p \in P^d} L\left(X_{i,p}^d, x_p^d, \lambda_p\right)} \tag{B.5}$$

where λ is the vector of bandwidths for the discrete predictors, and $L(\cdot, \cdot, \cdot)$ is the kernel function for discrete data.

References

Achenie, L. E. K.; Gani, R.; and Venkatasubramanian, V., eds. 2003. *Computer Aided Molecular Design: Theory and Practice*, volume 12 of *Computer-Aided Chemical Engineering*. Elsevier Science B.V. Amsterdam, The Netherlands.

Achterberg, T. 2009. SCIP: Solving Constraint Integer Programs. *Mathematical Programming Computation.* 1(1):1–41.

Aitchison, J., and Aitken, C. G. G. 1976. Multivariate Binary Discrimination by the Kernel Method. *Biometrika.* 63(3):413–420.

Byrd, R. H.; Nocedal, J.; and Waltz, R. A. 2006. KNITRO: An Integrated Package for Nonlinear Optimization. In Di Pillo, G., and Roma, M., eds., *Large-Scale Nonlinear Optimization*, volume 83 of *Nonconvex Optimization and Its Applications*. Springer Science+Business Media, Inc. 35–59. New York, NY. USA.

Constantinou, L., and Gani, R. 1994. New Group Contribution Method for Estimating Properties of Pure Compounds. *AIChE Journal.* 40(10):1697–1710.

Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; and Levy, O. 2013. The High-Throughput Highway to Computational Materials Design. *Nature Materials.* 12(1):191–201.

De Graef, M., and McHenry, M. 2012. *Structure of Materials: An Introduction to Crystallography, Diffraction, and Symmetry.* Cambridge University Press, second edition. Cambridge, UK.

de Jong, M.; Chen, W.; Angsten, T.; Jain, A.; Notestine, R.; Gamst, A.; Sluiter, M.; Ande, C. K.; van der Zwaag, S.; Plata, J. J.; Toher, C.; Curtarolo, S.; Ceder, G.; Persson, K. A.; and Asta, M. 2015. Charting the Complete Elastic Properties of Inorganic Crystalline Compounds. *Scientific Data.* 2(150009):1–13.

Drud, A. S. 1994. CONOPT – A Large-Scale GRG Code. *ORSA Journal on Computing.* 6(2):207–216.

Eljack, F. T.; Eden1, M. R.; Kazantzi, V.; Qin, X.; and El-Halwagi, M. M. 2007. Simultaneous Process and Molecular Design—A Property Based Approach. *AIChE Journal.* 53(5):1232–1239.

Epanechnikov, V. A. 1969. Non-Parametric Estimation of a Multivariate Probability Density. *Theory of Probability and Its Applications.* 14(1):153–158.

Fredenslund, A.; Jones, R. L.; and Prausnitz, J. M. 1975. Group-Contribution Estimation of Activity Coefficients in Nonideal Liquid Mixtures. *AIChE Journal.* 21(6):1086–1099.

Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; and Scheffler, M. 2015. Big Data of Materials Science: Critical Role of the Descriptor. *Physical Review Letters.* 114(10):105503.

Hahn, T., ed. 2005. *International Tables for Crystallography*, volume A: Space-Group Symmetry. Springer, fifth edition. Dordrecht, The Netherlands.

Hayfield, T., and Racine, J. S. 2008. Nonparametric Econometrics: The np Package. *Journal of Statistical Software.* 27(5).

Hohenberg, P., and Kohn, W. 1964. Inhomogeneous Electron Gas. *Physical Review.* 136(3B):B864–B871.

Hollander, M.; Wolfe, D. A.; and Chicken, E. 2014. *Nonparametric Statistical Methods.* Wiley Series on Probability and Statistics. John Wiley & Sons, Inc., third edition. Hoboken, NJ. USA.

Isayev, O.; Fourches, D.; Muratov, E. N.; Oses, C.; Rasch, K.; Tropsha, A.; and Curtarolo, S. 2015. Materials Cartography: Representing and Mining Materials Space Using Structural and Electronic Fingerprints. *Chemistry of Materials.* 27(3):735–743.

Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; and Persson, K. A. 2013. The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Materials.* 1(1):011002.

Joback, K. G., and Reid, R. C. 1987. Estimation of Pure-Component Properties from Group-Contributions. *Chemical Engineering Communications.* 57(1-6):233–243.

Kalidindi, S. R., and De Graef, M. 2015. Materials Data Science: Current Status and Future Outlook. *Annual Review of Materials Research.* 45(1):171–193.

Karunanithi, A. T.; Achenie, L. E. K.; and Gani, R. 2005. A New Decomposition-Based Computer-Aided Molecular/Mixture Design Methodology for the Design of Optimal Solvents and Solvent Mixtures. *Industrial & Engineering Chemistry Research.* 44(13):4785–4797.

Klincewicz, K. M., and Reid, R. C. 1984. Estimation of Critical Properties with Group Contribution Methods. *AIChE Journal.* 30(1):137–142.

Kohn, W., and Sham, L. J. 1965. Self-Consistent Equations Including Exchange and Correlation Effects. *Physical Review.* 140(4A):A1133–A1138.

Li, Q., and Racine, J. S. 2007. *Nonparametric Econometrics: Theory and Practice.* Themes in Modern Econometrics. Princeton University Press. New Jersey, NJ. USA.

Lydersen, A. L. 1955. Estimation of Critical Properties of Organic Compounds by the Method of Group Contributions. Technical report, University of Wisconsin-Madison. College of Engineering.

Ma, X.; Li, Z.; Achenie, L. E. K.; and Xin, H. 2015. Machine-Learning-Augmented Chemisorption Model for CO_2 Electroreduction Catalyst Screening. *Journal of Physical Chemistry Letters.* 6(18):3528–3533.

Maranas, C. D. 1996. Optimal Computer-Aided Molecular Design: A Polymer Design Case Study. *Industrial & Engineering Chemistry Research.* 35(10):3403–3414.

Mavrovouniotis, M. L. 1990. Group Contributions for Estimating Standard Gibbs Energies of Formation of Biochemical Compounds in Aqueous Solution. *Biotechnology and Bioengineering.* 36(10):1070–1082.

Montgomery, D. C., and Runger, G. C. 2003. *Applied Statistics and Probability for Engineers.* John Wiley & Sons, Inc., third edition. New York, NY. USA.

Nadaraya, E. A. 1964. On Estimating Regression. *Theory of Probability and its Applications.* 9(1):141–142.

Newnham, R. E. 2005. *Properties of Materials: Anisotropy, Symmetry, Structure.* Oxford University Press. Oxford, UK.

Odele, O., and Macchietto, S. 1993. Computer Aided Molecular Design: A Novel Method for Optimal Solvent Selection. *Fluid Phase Equilibria.* 82(1):47–54.

Ong, S. P.; Cholia, S.; Jain, A.; Brafman, M.; Gunter, D.; Ceder, G.; and Persson, K. A. 2015. The Materials Application Programming Interface (API): A Simple, Flexible and Efficient API for Materials Data Based on REpresentational State Transfer (REST) Principles. *Computational Materials Science.* 97(1):209–25.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Racine, J., and Li, Q. 2004. Nonparametric Estimation of Regression Functions with Both Categorical and Continuous Data. *Journal of Econometrics*. 119(1):99–130.

Revolution Analytics, and Weston, S. 2014. *foreach: Foreach Looping Construct for R*. R package version 1.4.2. URL: http://CRAN.R-project.org/package=foreach.

Rohrer, G. S. 2004. *Structure and Bonding in Crystalline Materials*. Cambridge University Press. Cambridge, UK.

Rudemo, M. 1982. Empirical Choice of Histograms and Kernel Density Estimators. *Scandinavian Journal of Statistics*. 9(2):65–78.

Saad, Y.; Gao, D.; Ngo, T.; Bobbitt, S.; Chelikowsky, J. R.; and Andreoni, W. 2012. Data Mining for Materials: Computational Experiments with *AB* Compounds. *Physical Review B*. 85(10):104104.

Samudra, A. P., and Sahinidis, N. V. 2013. Optimization-Based Framework for Computer-Aided Molecular Design. *AIChE Journal*. 59(10):3686–3701.

Tawarmalani, M., and Sahinidis, N. V. 2005. A Polyhedral Branch-and-Cut Approach to Global Optimization. *Mathematical Programming*. 103(2):225–249.

Tilley, R. J. D. 2006. *Crystals and Crystal Structures*. John Wiley & Sons, Inc. Hoboken, NJ. USA.

Wächter, A., and Biegler, L. T. 2006. On the Implementation of a Primal-Dual Interior Point Filter Line Search Algorithm for Large-Scale Nonlinear Programming. *Mathematical Programming*. 106(1):25–57.

Watson, G. S. 1964. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*. 26(4):359–372.

White, M. A. 2011. *Physical Properties of Materials*. CRC Press, second edition. Boca Raton, FL. USA.

Willighagen, E. L.; Wehrens, R.; Melssen, W.; de Gelder, R.; and Buydens, L. M. C. 2007. Supervised Self-Organizing Maps in Crystal Property and Structure Prediction. *Crystal Growth & Design*. 7(9):1738–1745.