Downloaded by Stanford Libraries on 7/20/2020 7:06:37 PM.
Published on 15 July 2020 on https://pubs.rsc.org | doi:10.1039/9781839160233-00512

CHAPTER 20

A Few Guiding Principles for Practical Applications of Machine Learning to Chemistry and Materials

S. SHANKAR*[a] AND R. N. ZARE[b]

[a] Harvard University, Applied Physics, Harvard Paulson School of Engineering and Applied Sciences, 29 Oxford Street, Cambridge, MA 02189, USA; [b] Department of Chemistry, Stanford University, 333 Campus Drive, Stanford, CA 94305, USA
*Emails: SShankar@seas.harvard.edu; SShankar@Material-Alchemy.Org

20.1 Introduction

All the structures or devices that affect humans, from a small microprocessor to large airplanes and bridges, have underlying materials and chemicals that enable their forms and functionalities. All chemical and material properties, and their responses to external conditions, are based on chemical composition and structural characteristics. Systematic characterization of materials consists of the following four components: **experiments, theory, data,** and **computations**. Given the complex nature of inorganic and organic materials, all of these specific components are necessary for a fundamental understanding that can help the design of materials for human needs. These components are illustrated in Figure 20.1 as distinct, but interconnected,

Theoretical and Computational Chemistry Series No. 17
Machine Learning in Chemistry: The Impact of Artificial Intelligence
Edited by Hugh M. Cartwright
© The Royal Society of Chemistry 2020
Published by the Royal Society of Chemistry, www.rsc.org

View Online

Downloaded by Stanford Libraries on 7/20/2020 7:06:37 PM.
Published on 15 July 2020 on https://pubs.rsc.org | doi:10.1039/9781839160233-00512

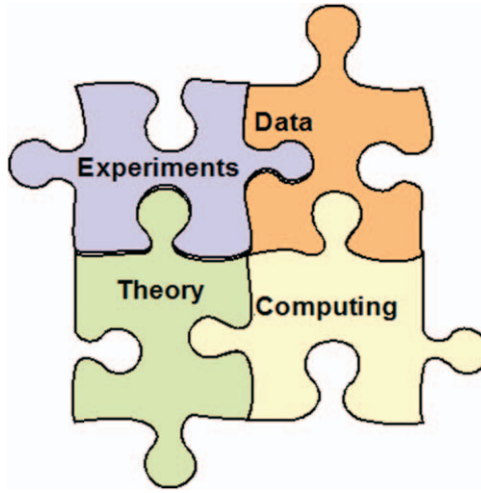


Figure 20.1 Four components for fundamental characterization of chemicals and materials.

which will help in framing this chapter. We will now briefly summarize each of the components with specific examples.

Experiments refer to physical measurements that are used to quantify specific attributes of materials under specific conditions. For example, spectroscopic methods are used to measure chemical bonding and composition of chemicals and materials; X-ray diffraction measurements reveal crystallographic information, secondary-ion mass spectrometry is used for estimating the composition of solid surfaces and thin films, laser-induced fluorescence is used for detecting chemical species and energetics of excited states, calorimetry is used to measure thermal characteristics, *etc.* In an industrial setting, experiments are used to measure responses of chemicals and materials to external fields or conditions, as in four-point probes for electrical resistivities, thermocouples to measure temperatures in processing equipment, tensile measurements to gage mechanical responses, *etc.*

Theory refers to fundamental understanding of physical and chemical characteristics of chemicals and materials. Examples include Schrodinger equation formalism for quantum mechanical description of electronic structure of matter, classical equations of motion for dynamics of particles, constitutive laws for mechanical properties, transition state theory for chemical reaction barriers, thermodynamic laws for ensembles of systems associated with heat, *etc.* Theory provides a basis for linking both computer simulations and experimental measurements, and by its definition is also extrapolative, an important point that we will revisit during the discussion of the guiding principles.

Computations, similar to physical experimental measurements (*in-silico* experiments), refer to simulated experiments done in computers. Examples of computational methods include the use of integration procedures for differential equations, Monte Carlo simulations for probabilistic processes,

View Online

Downloaded by Stanford Libraries on 7/20/2020 7:06:37 PM.
Published on 15 July 2020 on https://pubs.rsc.org | doi:10.1039/9781839160233-00512

thermodynamic methods for phase equilibrium, many-body methods for excited-state properties, Green's function methods for inhomogeneous and non-equilibrium attributes, statistical and empirical methods for correlations, *etc.* Computational simulations may include combining one or more of the techniques with other components. For example, Density Functional Theory (DFT), a simplification of the *theoretical* Schrodinger equation formalism, can be used to *compute* ground-state energies.

Data is the fourth component and is considered critical to transfer and linking of information between the other components. For example, data based on chemical composition from spectroscopic measurements are used to build chemical structures of materials in computers or to establish the rate of chemical reactions. Using appropriate theoretical and numerical formalisms, these chemical structures may be simulated (using computers) to understand dynamics under different conditions than those measured experimentally. Simulation of protein folding is an example in which measured crystallographic structures of proteins are used to estimate long-term dynamics of protein behavior. Data can also be used to develop theory. An example of this is the blackbody radiation law developed by Planck using experimental spectroscopic data of the solar spectrum and building on existing correlations.

For a complete understanding of realistic chemical and material systems, the criticality of combining all these components can be illustrated with an example from plasma processing. Plasma-based processing is one of the most critical steps in semiconductor manufacturing in very large-scale integration of solid-state circuits, the basis of all present-day computers. As plasmas are ''cold, but reactive'' systems with electrons at high energies (10 000 K) compared to the energies of other species like ions, molecules, and chemical specific (greater than 300 K), most of the active chemical processes in plasmas are in non-equilibrium. Understanding these non-equilibrium interactions are necessary for the control of these processes in semiconductors. One of the successful efforts undertaken by a consortium of semiconductor companies in the 1990s quantified different aspects of plasma chemistry for perfluorinated hydrocarbons systematically.[1–4] Figure 20.2 illustrates examples of the above four components as applied to plasma processing. To address the complexity of the non-equilibrium chemistry, the components involved combining measurements (*e.g.* Ion Fourier Transform Spectrometry, film thickness), theory (*e.g.* Binary Encounter Bethe Model, Binary-Encounter-Dipole Model), computations (*e.g.* Variational Green's Function methods, plasma dynamics, gas phase and surface chemistry calculations), and data (*e.g.* electron collision cross sections, ion reaction rates, etching rates).

20.2 Guiding Principles for Applications of Machine Learning

One set of methods, collectively known as Machine Learning (ML) or Artificial Intelligence (AI)-based methods, which are based on prior data, have

View Online

Published on 15 July 2020 on https://pubs.rsc.org | doi:10.1039/9781839160233-00512
Downloaded by Stanford Libraries on 7/20/2020 7:06:37 PM.

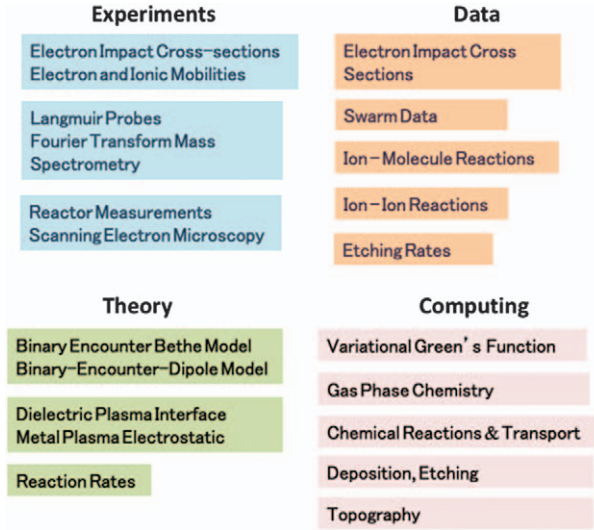


Figure 20.2 Four components for fundamental chemical or material characterization illustrated for plasma processing.

recently become a powerful approach for both the characterization and design of materials in biology, chemistry, and material science.[5–8] ML models are approximations that develop and quantify correlations between input data (''descriptor'') and an output signal (''observation''). These approximations are generally *learned* from a ''training set'', before the accuracy of the model is evaluated by using a separate ''test set''. These methods, based on numerical and statistical methods to analyze experimental data and computer simulations, have accelerated in the past decade due to a variety of factors: 1) Availability of powerful and widely available hardware enabled by Moore's law;[9] 2) Availability of large quantities of training data from high-throughput experimental measurements and computer simulations; 3) Advances in statistical methods and computational algorithms for analysis of data;[10] 4) Wider availability of data, hardware, and software brought about by economical and ubiquitous connectivity.

With the confluence of the above events, ML methods have been used extensively in a wide variety of fields in natural and physical sciences and engineering, including biology, chemistry, physics, chemical engineering, material science, genetics and physiology. ML methods fundamentally consist of formulating an appropriate representation of the data that identifies the features most important for a given application. The data could represent geometrical, structural features, or chemical features, as discrete values or analog functions. They are typically transferred from their native form to digital data, which are then transformed to finite dimensional vectors or some metric of the vectors (*e.g.* Euclidean distances). The key is to extract information with high statistical certainty, which forms a basis representing the underlying relationships. This is done by means of kernels, or

View Online

Downloaded by Stanford Libraries on 7/20/2020 7:06:37 PM.
Published on 15 July 2020 on https://pubs.rsc.org | doi:10.1039/9781839160233-00512

by using other non-linear mapping methods. To satisfy these key requirements, an appropriate feature extractor needs to be developed that maps the original data into optimal representations. Once these representations are identified, a model is used to identify and separate patterns of data. Another set of methods that are being widely used are called Artificial Neural Networks, in which several nodes are interconnected as in a graph emulating the behavior of interconnected neurons in a brain.[11] Each of the nodes is activated through weights of the edges and a non-linear local switching function. The main advantage of these latter techniques is their ability to automatically encode complex input–out relationships. On the other hand, the underlying information on physical or causal relationships is lost in this approach. In addition to the references mentioned above, there are several review papers and special editions of journals related to applications of these techniques to specific problems in chemistry and materials science.[12,13]

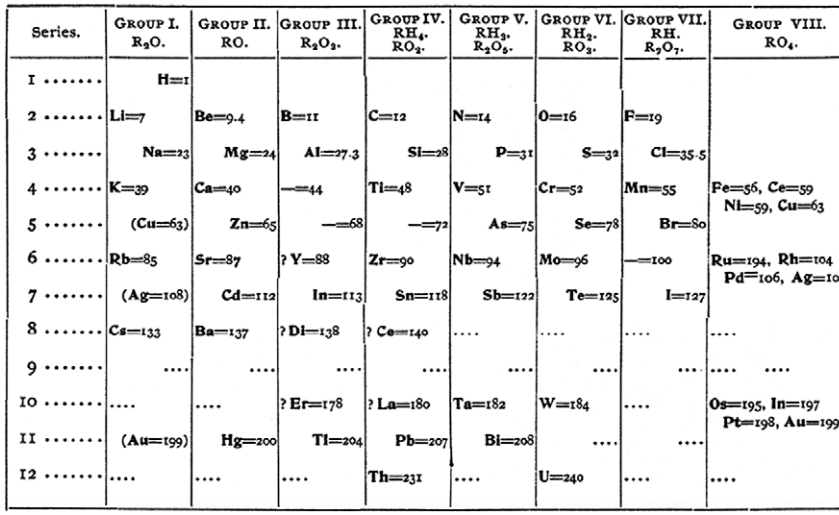
The main purpose of all these ML methods is to systematically reduce dimensions of the original problem by using transformations, without losing the key information that is necessary to reproduce the key chemical or material characteristics consistently. Although these techniques are very powerful, they can be misleading in applications to natural and physical sciences for a variety of reasons: data are relatively sparse in scientific applications compared with pattern recognition applications used in multimedia and consumer applications; the physical and natural laws are complex and need to be self-consistent at any point in time; external surrounding effects and boundary conditions can influence material and chemical behaviors. Given the widespread use of these techniques as mostly black box methods, we have developed a few specific guiding principles that will help other practitioners in cautious application of these techniques. Each of these guiding principles is stated and illustrated with specific examples taken from literature and our own empirical observations from both academic and industrial settings. These examples are neither meant to be comprehensive reviews of all possible ML methods nor to be detailed recommendations of the application of any specific ML method to chemical and materials science problems. In addition, as will be seen below, each of these examples could be associated with more than one guiding principle. Our focus is to use the examples from biology, chemistry, physics, statistics, and engineering to emphasize the need for caution in the broad use of ML to these fields.

Guiding Principle 1: *Use ML for Interpolation but with Care for Extrapolation*

This principle enumerates the point that, although ML methods are very effective in their ability to interpolate in the ranges in which these models are trained, caution is necessary during extrapolation outside the range of the training set. Unlike the theory that describes the physical phenomena in consideration and its applicability, the training data are determined by the limits of available data, including the specific properties or characteristics,

View Online

Downloaded by Stanford Libraries on 7/20/2020 7:06:37 PM.
Published on 15 July 2020 on https://pubs.rsc.org | doi:10.1039/9781839160233-00512



Series.	Group I. R_2O.	Group II. RO.	Group III. R_2O_3.	Group IV. RH_4. RO_2.	Group V. RH_3. R_2O_5.	Group VI. RH_2. RO_3.	Group VII. RH. R_2O_7.	Group VIII. RO_4.
1	H=1							
2	Li=7	Be=9.4	B=11	C=12	N=14	O=16	F=19	
3	Na=23	Mg=24	Al=27.3	Si=28	P=31	S=32	Cl=35.5	
4	K=39	Ca=40	—=44	Ti=48	V=51	Cr=52	Mn=55	Fe=56, Co=59 Ni=59, Cu=63
5	(Cu=63)	Zn=65	—=68	—=72	As=75	Se=78	Br=80	
6	Rb=85	Sr=87	? Y=88	Zr=90	Nb=94	Mo=96	—=100	Ru=104, Rh=104 Pd=106, Ag=108
7	(Ag=108)	Cd=112	In=113	Sn=118	Sb=122	Te=125	I=127	
8	Cs=133	Ba=137	? Di=138	? Ce=140	....	....	....	....
9	....	....	....	....	....	....	....	....
10	....	....	? Er=178	? La=180	Ta=182	W=184	....	Os=195, In=197 Pt=198, Au=199
11	(Au=199)	Hg=200	Tl=204	Pb=207	Bi=208	....	....	
12	....	....	....	Th=231	....	U=240	....	....

Figure 20.3 Mendeleev's Original Periodic Table.

the composition and structural ranges, the conditions of experimental measurement (or simulations), and the spatial/temporal scales. Any ML model, which has been trained and characterized statistically, is applicable within these limits. However, extrapolation outside these limits may provide misleading conclusions, which we will illustrate with our example of the formulation of the periodic table of elements.

One of the most successful applications of classification is the periodic table, celebrated recently by scientists and engineers in 2019 as the "International Year of the Periodic Table" of chemical elements. Formulation of the table is attributed to Dmitri Mendeleev, who demonstrated in 1869 that the chemical elements then known, when arranged by their respective atomic weights, demonstrated periodicity.[14] See Figure 20.3 for Mendeleev's initial published version of his table. Although patterns in the behavior of chemical elements have been observed before, we will use the example of Mendeleev's periodic table to emphasize the difference between interpolation and extrapolation. This is because Mendeleev made many successful predictions of new elements and their chemical natures in addition to correcting known atomic weights of existing materials. The atomic weight could be viewed as a "descriptor" as defined in the ML literature. A few of the key points that Mendeleev used to classify elements are summarized below, and are discussed in more detail in his papers[14,15] and elsewhere:[16]

1. *Elements exhibit periodicity of "properties" when arranged according to the magnitude of their atomic weights*;
2. *The atomic weights of chemically analogous elements either have similar values or increase in a uniform fashion;*

View Online

Downloaded by Stanford Libraries on 7/20/2020 7:06:37 PM.
Published on 15 July 2020 on https://pubs.rsc.org | doi:10.1039/9781839160233-00512

3. *The arrangement according to the atomic weight corresponds to the valencies of the elements and, to a certain degree, to differences in their chemical behavior, e.g. Li, Be, B, C, N, O, F;*
4. *The magnitude of the atomic weight determines the properties of the element;*
5. *It allows one to foresee the discovery of many new elements, e.g. analogs of Si and Al with atomic weights between 65 and 75;*
6. *It is to be expected that some atomic weights will require correction, e.g. Te cannot have an atomic weight of 128, but rather one between 123 and 126;*
7. *The above table suggests new analogies between elements;*

There are many key points that can be seen in Mendeleev's "general deductions" as a forerunner to the current pattern recognition using ML methods. His prediction of classifying elements into groups based on atomic weights, possible existence of newer elements, and correction to the existing atomic weights were all validated by empirical observations. Mendeleev was able to establish the differences between physical properties (*e.g.* atomic weights) and chemical properties that determine how elements combine with other elements (*e.g.* acid or base formation). Even today, the periodic table is a tool for classification, description, analysis, and prediction, as the ML methods purport to do. Although Mendeleev used this formalism to analyze similarities in chemical reactions and his classification seemed to explain a few of the key observations, he was aware of the dangers of overly extending this system to chemical reactions as expressed in his words:[14] "*Because of the diversity of the relations existing between the simple substances, one cannot think of representing their system in the form of a continuous series, since the mutual relations of the substances are extraordinarily* **diverse**. . .". We have highlighted the word "diverse" to make the point that chemistry is complex in the number of possibilities and hence any interpolation built on specific data can be limiting in its application.

Based on this classification, Mendeleev predicted the existence of several unknown elements through the interpolation of atomic weights. An example of the success of this method is evident in his prediction of properties of eka-silicon, which was later discovered and named as germanium[16] (see Table 20.1).

Table 20.1 Comparisons between predicted properties of Eka silicon and Germanium. Adapted from ref. 16 with permission from Oxford Publishing Limited, Copyright 2007.

Property	Ekasilicon	Germanium
Atomic mass	72	72.61
Density	5.5	5.35
Melting point (°C)	High	947
Color	Gray	Gray
Oxide type	Refractory oxide	Refractory dioxide
Oxide density	4.7	4.7

View Online

Table 20.2 Mendeleev's successful and unsuccessful predictions. Adapted from ref. 16 with permission from Oxford Publishing Limited, Copyright 2007.

Mendeleev's element	Predicted atomic weight	Experimental atomic weight	Current element
Coronium	0.4	Not found	Not found
Ether	0.17	Not found	Not found
Eka-boron	44	44.6	Scandium
Eka-cerium	54	Not found	Not found
Eka-aluminum	68	69.2	Gallium
Eka-silicon	72	72	Germanium
Eka-manganese	100	99	Technetium
Eka-molybdenum	140	Not found	Not found
Eka-niobium	146	Not found	Not found
Eka-cadmium	155	Not found	Not found
Eka-iodine	170	Not found	Not found
Eka-cesium	175	Not found	Not found
Tri-manganese	190	186	Rhenium
Dvi-teurium	212	210	Polonium
Dvi-cesium	220	223	Francium
Eka-tantalum	235	231	Protactinium

Downloaded by Stanford Libraries on 7/20/2020 7:06:37 PM.
Published on 15 July 2020 on https://pubs.rsc.org | doi:10.1039/9781839160233-00512

Mendeleev further made many predictions using atomic weight as the classifying parameter. His estimate of atomic weights together with the successful (*e.g.* eka-boron or scandium with atomic weight 44) and unsuccessful predictions (*e.g.* coronium with atomic weight 0.4) are given in Table 20.2. Several interesting observations are evident from these predictions, two of which show the limitations of this approach: 1) Values of some atomic weights are estimated to be less than one (or lighter than hydrogen); 2) The success of the prediction rate is about 50%. As atomic weight was used as the *only* key variable to classify the elements based on the known data in the late nineteenth century, this was a remarkable success rate. However, it also shows the difficulty of trying to extrapolate based on known patterns and behaviors of complex systems, representing the field of chemistry, in a simple graphical form. Replacement of the atomic weight by atomic number and incorporation of the principles of quantum mechanics in the twentieth century led to a more accurate way of classifying elements. New theory, experimental measurements, and computations provided a more detailed and consistent classification. As mentioned before, Mendeleev himself was aware of the limitations during the formulation, but became more confident as a result of the earlier validation, including the discovery of new elements (*e.g.* discovery of germanium in 1886), and tried to extrapolate broadly. Extrapolation of the known classification in itself was limiting, as evident from Mendeleev's predictions (including atomic weights less than 1). This limited its application within chemistry, where the complexity restricts description by simple correlations. Mendeleev's approach stands as a red-flag warning to investigators when applying scientific principles to extrapolate outside the proper context.

View Online

Downloaded by Stanford Libraries on 7/20/2020 7:06:37 PM.
Published on 15 July 2020 on https://pubs.rsc.org | doi:10.1039/9781839160233-00512

Guiding Principle 2: *Ensure Consistency Between Sources of Data Used in the ML Development and the Targeted Application*

It is important that the measurements used in building the model or the theory used in simulating specific properties are consistent with the end applications. Since ML models as they exist today do not have any intelligence beyond the training data on which they are based, this principle is intuitive, and yet it can be misapplied, as we will demonstrate in the example below.

In the microelectronics industry, new materials are integrated in electronic devices for advancing Moore's law. Integration of heterogeneous materials leads to many possible interfaces between these materials. Characterizing these interfaces is critical to ensure that the integrated devices function without failures. One specific attribute of these interfaces is adhesion. Adhesive strength depends on the intrinsic interface chemical bonds and the extrinsic processing conditions under which these interfaces were formed. As a result, detailed characterization of composition, interface morphology, and structure are critical for interfaces but this is difficult and time-consuming due to the buried nature of these surfaces and the multiple surfaces that make up a device. To address this problem, several methods for estimating adhesion between copper and other materials were evaluated by Kong *et al.*[17] A DFT model for estimating interface adhesion energy was compared with the use of an ML method based on experimental measurements of wetting angle for each interface, to compare and contrast the methods in estimating the adhesion. The DFT-based simulations were done at low temperatures (0K), where only the electronic degrees of freedom are taken into account. The wetting experiment is at a higher temperature as the material is heated beyond its melting point, and implicitly includes a phase transition, and the resulting structural and chemical changes. The comparisons are given in Figure 20.4.[17]

Overall, both ways of estimating adhesion are mostly accurate and consistent within the bounds of their applicability. As wetting is measured after increasing the temperature, it is likely that in reality, materials form complex chemical compounds or phases as determined by phase equilibria and

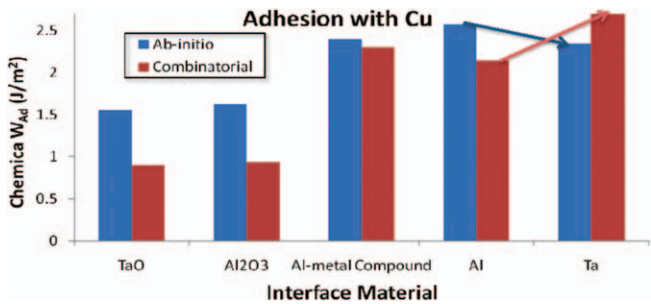


Figure 20.4 Comparisons between quantum-based simulations and ML based on experimental measurements.

View Online

Downloaded by Stanford Libraries on 7/20/2020 7:06:37 PM.
Published on 15 July 2020 on https://pubs.rsc.org | doi:10.1039/9781839160233-00512

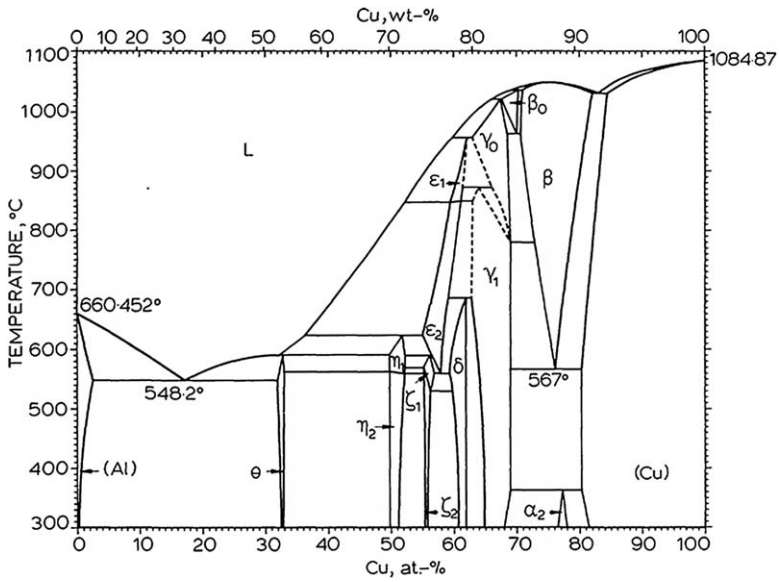


Figure 20.5 Complex phase diagram of Cu–Al.
Reproduced from ref. 38 with permission from Taylor and Francis, Copyright 1985.

kinetics, unlike *ab initio* methods in which the materials are relatively pure and the calculations are done for a temperature of 0 K. Because of the nature of the wetting measurements, the experimental data are expected to capture the realistic phases and chemical nature of the interfaces. Because of the gap between the model assumptions and the collected experimental data, it is likely that oxidation during the process of wetting transforms the interface structures and hence the models make divergent predictions. This can be seen by comparing the phase diagram of Cu–Al (Figure 20.5) with that of Cu–Ta. Copper and tantalum are immiscible under the conditions of measurement and hence are closer to the ideal surface approximations of DFT. Contrast this with Cu–Al, which forms a complex phase equilibrium including chemical compounds, which are possibly captured by wetting experiments. Hence models that are based on single-crystalline interfaces will be inaccurate, as seen from Figure 20.4.

One of the other applications in which one of us applied ML methods in the late 1990s (at that time known as reduced-order models or heuristics) was for advanced process control. As semiconductor technology was advancing towards smaller feature dimensions, the processes used for depositing-processing-growing thin films, and testing devices were also increasing in complexity. One of the biggest hurdles for efficient manufacturing included the need for manual process optimization before and after equipment maintenance. To address these difficulties, a team of engineers across all of the Intel manufacturing fabrication plants developed

View Online

Downloaded by Stanford Libraries on 7/20/2020 7:06:37 PM.
Published on 15 July 2020 on https://pubs.rsc.org | doi:10.1039/9781839160233-00512

advanced process control systems, in which experimental data were used to train physically-based models and integrate them in the station controllers attached to the process equipment. The models based on chemical reactions and diffusion were trained using data from equipment inputs and the resulting process outputs. These led to elimination of misprocessing, and by middle 2000s, over forty applications were demonstrated during manufacturing.[18,19] The success of these models was largely due to the carefully controlled experimentation in manufacturing process equipment, which underscores the need for conducting measurements to train models that are consistent with the desired application conditions.

Guiding Principle 3: *Correlation is not Causation*

This is one of the most fundamental principles in tying empirical observations to underlying causes. Scientific discoveries need to differentiate between observed correlations and the underlying causation.[20] The main difference between the two is determined by time separation of a cause leading to an effect. It is possible for multiple effects to have the same underlying cause or for multiple causes to lead to the same effect. Correlations are statistical associations of data from a given distribution and are relatively valid as long as the external and boundary conditions remain the same. Causation, on the other hand, is predictive and is valid in both static and changing conditions and is not constrained by the statistical distributions.[21] Correlation is statistical and does not indicate the modulation of the distribution function if the conditions in which the data collected were to change due to interventions. As a result, it is difficult to differentiate between causes and effects using only correlational analysis. Although most ML methods traditionally use statistically significant correlations to establish relationships, they are unable to establish causal relationships. To illustrate this we will go to biological systems and try to draw similarities for chemical and material systems.

In evolutionary biology, as the existing conditions are used to trace the causes of these events, understanding the differences between cause and effect is critical. One specific illustration of this is given by de Duve[22] in listing different mechanisms of singularity, which is defined as an occurrence of a singular outcome or event. de Duve specifies six different causes that can lead to the same outcome. These are 1) *Deterministic Necessity* (deterministic laws of nature leading to a singular outcome for every singular cause); 2) *Selective Bottleneck* (externally imposed constraints leading to a singular outcome); 3) *Restrictive Bottleneck* (internally imposed constraints evolving to a singular outcome); 4) *Pseudo-Bottleneck* (time-dependent progressive attrition of all other outcomes except one); 5) *Frozen Accident* (rare statistical events leading to a singular outcome); 6) *Fantastic Luck* (singular or a highly improbable event leading to a singular outcome). Using the same premise, we can list similar mechanisms for complexity in cause-effect relationships in chemistry and materials. In each of these mechanisms, the

View Online

Downloaded by Stanford Libraries on 7/20/2020 7:06:37 PM.
Published on 15 July 2020 on https://pubs.rsc.org | doi:10.1039/9781839160233-00512

outcome is a specific attribute of material that is experimentally measured or simulated, as given below.

1. **Controlled measurements**: The measurement is done under tightly controlled conditions, and the resulting attribute is measured. The key point to note is that the measurement is isolated well enough that the specific property or attribute is controlled. Examples include tensile measurements for elastic moduli, spectroscopic measurements for chemical composition, and low-temperature measurement for electronic properties.
2. **Internal and external interactive effects**: The components of the system interact with one another and also with the environment. Examples include heterogeneous phase equilibrium of materials, and thin film properties modulated by the substrate on which they are measured.
3. **Interactions across length scales**: The different scales interact resulting in the measured property or attribute. These specific interactions are functions of the internal constraints such as composition, sample size, *etc.* Examples include grain growth, which in turn is dependent on the processing or sample size and preparation conditions, which determines the mesoscopic morphology of materials resulting in measured properties of polycrystalline materials.
4. **Interactions across time scales:** Examples include the use of equilibrium properties in lieu of time-dependent kinetics, resulting in the same chemical composition only under narrow conditions.
5. **Rare event effects**: This could result from relatively rare events affecting the property that is being measured. Examples include contamination of the measuring instruments or from errors in software used in calibration or computing. This is especially applicable for highly sensitive measurements or large and complex calculations.
6. **External errors:** This results from equipment set-up or human errors including confirmation biases used in the analysis. Examples include equipment set-up or when there are multiple steps in deconvolving and uncoupling complex properties.

These six mechanisms illustrate the principle that it is difficult to correlate an outcome with its corresponding causation. The focus of this principle is to try to eliminate the alternate mechanisms so that the correlation has a higher probability of being tied to causation. The example of the periodic table formulation illustrated in one of the previous principles also suffered from tying correlations strongly to causation. Although the original periodic table was established based on the correlations of atomic weights with physical and chemical properties, the modern periodic table incorporates the subsequent scientific advances based on the underlying atomic numbers and electronic structures.[23] Thus, using quantum mechanics, the periodic table has been expanded to link with the principles behind the causation.

View Online

Downloaded by Stanford Libraries on 7/20/2020 7:06:37 PM.
Published on 15 July 2020 on https://pubs.rsc.org | doi:10.1039/9781839160233-00512

Guiding Principle 4: *Optimize Information Extraction when Using ML*

This principle is fundamental to the intent of ML in building patterns based on data. As the patterns are expected to reveal the underlying relationships, ML is about quantifying information about the system itself. By information, we refer to the underlying causal relationships, which may include physical and natural laws, conservation laws, symmetries, thermodynamics, or other constitutive relationships. As these are not easy to formulate, compute, or measure for complex chemical and material systems, using data to identify patterns is a powerful alternative, within the range of applicability of the data.

Identifying the appropriate laws for all systems are due to the following: complex interactions, isolation of systems of interest, and quantification of many-body/many-scale effects. The scope of the specific material or chemistry being addressed by physical laws is limited when applying to a real system and hence needs simplifications without trading-off accuracy. This becomes even more difficult when chemical reactions are involved, as the atomic and electronic movements are out of local equilibrium. For example, the accurate many-body models for chemistry scale as $O(n^3\text{-}n^8)$, where n is the number of particles.[24] Most real-world applications consist of more than a single molecule and may contain $O(10^{23})$ molecules. Hence, when applied to a real chemical or material system, the accuracy is traded off with the reduction of complexity of the chemistry or materials either in the number of atoms or electrons or in reducing the time scales. This is qualitatively illustrated in the Figure 20.6 where we have plotted the data requirement

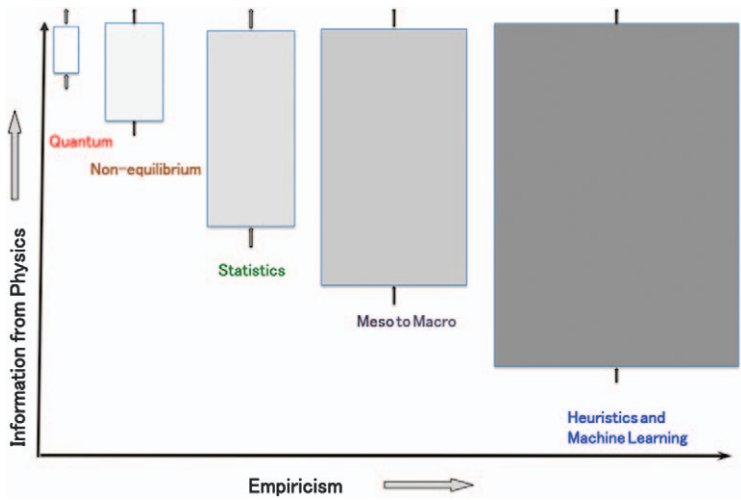


Figure 20.6 Information *vs* for different models compared with ML, indicating how information from physical theories is traded off with empirical data.

View Online

Published on 15 July 2020 on https://pubs.rsc.org | doi:10.1039/9781839160233-00512
Downloaded by Stanford Libraries on 7/20/2020 7:06:37 PM.

from empirical observations against the information (or prediction accuracy) obtained from the fundamental laws. Moving from left to right, the complexity is addressed by the use of more empirical data using less predictive theory, computations, or measurements. As we go from left to right, boxes go from white (representing a fully predictive model) to ML-based models shown as black boxes (representing all empiricism). We will illustrate an example of this guiding principle below, in which ML methods were used to estimate electronic correlation energies using other quantum chemistry models.

Hartree-Fock (HF) and DFT models are typically used to compute ground-state energetics of atoms, molecules, and materials in the solid state. Both these methods find approximations to the Schrodinger equation, which is a many-body, three-dimensional time-dependent formalism for electronic/particle wave functions. The practical difficulties in application of these models are twofold: 1) In addressing the so-called electronic correlation energies for higher prediction accuracies; and 2) In scaling up the computational problem to larger numbers of electrons and atoms, without sacrificing accuracy. The errors result from treating the interaction of one electron with another as a function of a smoothed and averaged electron density. To address the second difficulty, many approximations, including those based on ML, are made to represent correlations. Brockherde *et al.*[25] used ML to develop methods addressing the limitations of DFT, while Miller and co-workers[26,27] used machine learning to build more accurate models of electronic correlation energy based on self-consistent HF methods. We will use the latter as an example of how ML was used to optimize information extraction as related to correlation energy.

In this example, Miller and co-workers[26,27] used ML to build models for increasing the ''chemical'' accuracy of predictability for quantum models. This accuracy amounts to total energies within $\sim$1 kcal mol^{-1} ($\sim$4 kJ mole^{-1}).[24] The error introduced due to the use of HF methods can affect those cases in which chemical bond breaking is important. To address this, the researchers used features based on molecular orbital properties, such as Fock, Coulomb, and exchange matrix elements. These addressed both the challenges in trading off predictability with the size of descriptor set. The models based on ML were compared with the more accurate post-Hartree–Fock methods (Coupled-Cluster, Moller–Plesset perturbation theory). The scalability and transferability of the ML-based models were demonstrated by further refining the techniques by optimizing the feature set and extending it to thermal conditions, and tested on 7211 organic molecules.[27] Their success in optimization of information extraction is evident, as they needed fewer molecular geometries and computations to reach the same accuracy. This work illustrates the guiding principle, where optimal use of ML methods can circumvent the slow scaling of more accurate quantum chemistry models, without trading off accuracy.

View Online

Guiding Principle 5: *Combine Different Methods Consistently, Including Experiments, Theory, and Simulations, to Provide a Larger Window of Applicability*

Downloaded by Stanford Libraries on 7/20/2020 7:06:37 PM.
Published on 15 July 2020 on https://pubs.rsc.org | doi:10.1039/9781839160233-00512

This principle lays out the need for combining all the four components consistently for an integrated solution. Combining experiments, theory, computing, and data ensures self-consistent and systematic analysis. This, in turn, ensures predictability of the analysis and a wider range of applicability. This principle is in alignment with some of the other principles (Principle 1 on limiting use of ML for extrapolation, Principle 3 on causation *vs.* correlation, and Principle 4 on optimizing information extraction). As indicated before, it is important that the theory and measurements used in building the model are consistent with the end applications for which ML is being used.

An illustrative example with the combination of experimental approaches, data, computing, and theory on altered cancer metabolism is given below. This example focused on a clinical application of clear cell renal cell carcinoma (ccRCC).[28] As the most common and lethal subtype of kidney cancer, ccRCC is currently diagnosed using intraoperative frozen section analysis during a partial nephrectomy. If a surgeon can be informed which tissue is to be excised and which is to be retained, this would greatly improve the chances of a successful operation.

Owing to the time consuming and unreliable nature of the analysis, this study determined if desorption electrospray ionization mass spectrometry imaging (DESI-MSI) combined with statistical ML methods can be used as an alternate molecular diagnostic and prognostic tool for delineating surgical margins. In this technique, the tissue sample is mounted on a flat surface that can be translated in the x and y directions. The surface is bombarded with charged microdroplets and the splash of yet smaller microdroplets enters a mass spectrometer, producing a two-dimensional chemical map having very rich content. In this case, the experiment involved DESI-MSI of 23 pairs of fresh-frozen benign and cancer tissue samples. Based on small metabolites, fatty acids, and lipids obtained from DESI-MSI, a binary classifier was successfully trained with 85% accuracy per patient to learn the differences between benign and cancer tissue based on changes in the relative abundances of metabolites. As the training procedure returns a sparse model, the set of chemical compounds was highly interpretable and then used to experimentally identify which particular metabolites are either overexpressed or suppressed in cancer tissues compared with normal tissues. The link between the model features and identifiable metabolites is critical, because it gives confidence in the predictions and has led to novel insights into alterations in cancer metabolism, including the Krebs cycle, fatty acid, and amino acid pathways. For example, this methodology, which found the ratio of glucose to arachidonic acid as highly predictive (>70%) in distinguishing normal and cancer tissue, suggests a dependence on high glucose production and fatty acid breakdown for sustaining cancer growth.

View Online

Downloaded by Stanford Libraries on 7/20/2020 7:06:37 PM.
Published on 15 July 2020 on https://pubs.rsc.org | doi:10.1039/9781839160233-00512

Although the initial aim was to improve frozen section analysis, it can be extended into a 'wider range of applicability', namely, to study metabolic vulnerabilities in kidney cancer to help guide the development of new therapeutic drugs. It is also noteworthy that this example is in accordance with Principle 4 on optimizing information extraction, where the changes in kidney cancer metabolic usage are captured by differences in the relative abundances of metabolites from DESI-MSI.

20.3 Concluding and Cautionary Remarks

ML methods have been found to be very powerful in many consumer applications and effective in applications to solve many aspects of chemistry and materials as evidenced by the many publications mentioned previously. Our intent in this chapter is twofold: to frame the ML applications in the context of the four components of experiments, theory, computing, and data, and to offer cautionary perspectives on the use of ML methods. We have done this by offering five specific guiding principles and how they can be used for a systematic analysis.

We revisit the potential of ML to identify key correlations between different components: experiments, theory, computation, and data. Further, we will illustrate with examples to indicate two different levels in which ML can be applied. This is illustrated in Figure 20.7a and b representing the two levels: micro or intra and macro or inter.

At the first level, using ML within the four components (*micro* or *intra* links) enables acceleration of the analysis. Examples include optimizing transition state analysis,[29] accelerating computations of ground state energetics,[25–27] ML used for learning interatomic potentials for atomistic

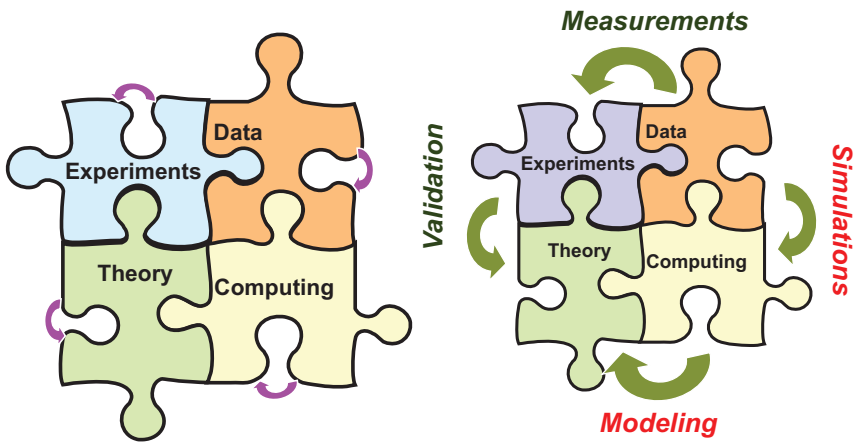


Figure 20.7 Application of ML at two different levels: (a) Micro/Intra: In the first level, ML can be used to accelerate each of the components; (b) Macro/Inter: In the second level, ML can be used to link between the components accelerating the overall analysis.

View Online

Downloaded by Stanford Libraries on 7/20/2020 7:06:37 PM.
Published on 15 July 2020 on https://pubs.rsc.org | doi:10.1039/9781839160233-00512

simulations,[30] and ML trained on reaction data including failed or unsuccessful hydrothermal syntheses to predict reaction outcomes for the crystallization of vanadium selenites,[31] *etc.*

At the second level, linking between the different components enables acceleration of the overall analysis (*macro* or *inter* links). Each of the links between the components provides a specific role: **measurements** provide the link between experiments and data, **validation** acts as the link between experiments and theory, **modeling** is the link between computing and theory, while **simulations** provide a link between computing and data. Examples include linking between ML methods and experiments to drive the analysis for metallic glasses,[32] experiments, computations, and data to control processes or accelerate analysis,[17–19] computations and data linked with ML to develop theoretical understanding of surface reactions,[33] using deep neural networks to gain insight into quantum mechanical-based theory of stability of molecules,[34] and linking of ML with flow reactors for automated optimization of diverse chemical reactions,[35] and by combining experiments, theory, data, and computing, Zare's team[28] demonstrating a clinical application for clear cell renal cell carcinoma, as elaborated in the previous section.

Our five guiding principles are intended to help with applying Machine Learning to physical sciences and are based on years of developing and applying methods to address problems in chemistry and material science. As these techniques become more widely adopted, our hope is that these principles will be used as safeguards during their application in solving problems in chemistry, engineering, and materials science. It is also important to note that these principles are not meant to be prescriptive but are guidelines for both scientists and the students who wish to apply ML to physical sciences. It also emphasized through these principles that ML is not a substitute to any of the four critical components (experiments, theory, computing, data) needed in fundamental understanding.

We illustrate this final overarching principle with the discussions between Dyson and Fermi on meson-proton scattering in 1953.[36] The intent of their collaboration was to compare Dyson's theoretical model and computations with Fermi's measurements. Although the comparisons were close, Fermi was critical of the arbitrary nature of the cut-off parameters. Fermi's opinion represented the prevalent understanding among the physicists that with four arbitrary parameters, one could fit an elephant and with five parameters, the elephant could be made to wiggle its trunk. Using five complex parameters, Mayer *et al.*[37] demonstrated that it was indeed the case. The key message is that validation of experimental data by parameters or methods without physical basis in itself is limiting, and there is a need for caution in the extrapolation of these methods. It is not the closeness of the model fit with data that is important, but rather the fundamental causation behind the physical phenomena for which the data have been collected. This example encapsulates our guiding principles and is consistent with the limitation in application of frequency-based statistics to observations or data for a causal analysis. It is further illustrated by Dyson's own retrospective

View Online

observations, roughly fifty years after the discussion with Fermi. In Dyson's own words:

> *The crucial discovery that made sense of the strong forces was the quark. Mesons and protons are little bags of quarks. Before Murray Gell-Mann discovered quarks, no theory of the strong forces could possibly have been adequate. Fermi knew nothing about quarks, and died before they were discovered. But somehow he knew that something essential was missing in the meson theories of the 1950s. His physical intuition told him that the pseudoscalar meson theory could not be right. And so it was Fermi's intuition, and not any discrepancy between theory and experiment, that saved me and my students from getting stuck in a blind alley.*

This exchange between Fermi and Dyson and the retrospective summary clearly illustrate the central premise of this chapter, namely, that correlations driven only by validation with data (either experimental or computational or both) can be misleading and hence should be augmented by systematic analysis based on physical principles and underlying theory. As we have indicated, ML methods are powerful tools to accelerate learning in each of the four components and in the development of the underlying theory, methods, and algorithms by linking between the components. In the absence of an underlying theory, ML can still be useful for interpolation, but should be applied with caution outside the domain on which ML was trained. It is our hope that the wide use of ML methods will be tempered by the proposed guiding principles in this work.

Downloaded by Stanford Libraries on 7/20/2020 7:06:37 PM.
Published on 15 July 2020 on https://pubs.rsc.org | doi:10.1039/9781839160233-00512

Acknowledgements

The authors would like to acknowledge the Camille and Henry Dreyfus Foundation for the opportunity to present and discuss a different perspective on the use of ML in chemical sciences and materials engineering. The chapter builds on previous joint projects with several researchers and practitioners over the last two decades and discussions in several meetings including the Dreyfus Foundation ML meeting (2019), University of California Los Angeles, Institute of Pure and Applied Mathematics Workshop on Machine Learning and Many Particle Physics, (2016) and Stanford Class on Translation (2012). SS would like to acknowledge the support of the Hearst Foundation and the UCLA IPAM Fellowship. In addition, we acknowledge the information provided by Thomas Miller of Caltech, which was used to illustrate one of the guiding principles.

References

1. V. McKoy, C. Winstead and C.-H. Lee, *J. Vac. Sci. Technol., A*, 1998, **16**, 324.
2. W. L. Morgan, *Adv. At., Mol., Opt. Phys.*, 2000, **43**, 79.

View Online

Downloaded by Stanford Libraries on 7/20/2020 7:06:37 PM.
Published on 15 July 2020 on https://pubs.rsc.org | doi:10.1039/9781839160233-00512

3. S. Shankar, B. V. McKoy and W. L. Morgan, *Sixth U.S. National Congress on Computational Mechanics*, U.S. Association for Computational Mechanics, Dearborn, Michigan, 2001.
4. K. Yoshida, S. Goto, H. Tagashira, C. Winsted, B. V. McKoy and W. L. Morgan, *J. Appl. Phys.*, 2002, **91**, 2637.
5. H. M. Cartwright, *Using Artificial Intelligence in Chemistry and Biology: A Practical Guide*, CRC Press, Boca Raton, FL, 2008.
6. S. Shankar, Machine Learning for Materials Design: Combination of Theoretical methods, Heuristics, and Hybrid Techniques, *Workshop on Synergies between Machine Learning and Physical Models*, University of California-Los Angeles, Dec 5–9, 2016.
7. K. T. Butler, D. W. Davies, H. M. Cartwright, O. Isayev and A. Walsh, *Nature*, 2018, **559**, 547.
8. B. Sanchez-Lengeling and A. Aspuru-Guzik, *Science*, 2018, **361**(6400), 360.
9. https://www.intel.com/content/www/us/en/silicon-innovations/moores-law-technology.html; accessed Jan 31, 2020.
10. T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York City, USA, 2009.
11. Y. LeCun, Y. Bengio and G. Hinton, *Nature*, 2015, **521**, 436.
12. M. Rupp, *Int. J. Quantum Chem.*, 2015, **115**, 1003.
13. I. Tanaka, K. Rajan and C. Wolverton, *MRS Bull.*, 2018, **43**, 659.
14. D. I. Mendeleev, On the relationship of the properties of the elements to their atomic weights, in *Mendeleev on the Periodic Law*, ed. W. B. Jensen, Zeitschrift fur Chemie, Mineola, NY, 2002, pp. 405–406.
15. D. I. Mendeleev, On the periodic regularity of the chemical elements, *Mendeleev on the Periodic Law, Annalen der Chemieund Pharmacie*, 8(Suppl), ed. W. B. Jensen, 2002, pp. 133–229.
16. E. R. Scerri, *The Periodic Table: Its Story and Its Significance*, Oxford University Press, Oxford, UK, 2007.
17. C. S. Kong, M. Haverty, H. Simka, S. Shankar and K. Rajan, *Model. Simul. Mater. Sci. Eng*, 2017, **25**, 065014.
18. J. Luke, T. Albertson, Y. H. Lin, S. Shankar and D. Pantuso, *Intel. Test Assembly J.*, 2003, **6**, 481.
19. S. Shankar, K. Knutson and Y. H. Lin, *DOTS: Advanced Paradigm in Process Control – Yesterday, Today, and Tomorrow*, Intel Advanced Process Control (APC) Summit, Chandler, 2003.
20. J. Woodward, *Philos. Sci.*, 2014, **81**, 691.
21. J. Pearl, *Statist. Surv.*, 2009, **3**, 96.
22. C. de Duve, *Singularities*, Cambridge Univ. Press, New York, 2005.
23. L. N. Ross, *Synthese*, 2018, DOI: 10.1007/s11229-018-01982-0.
24. F. Jensen, *Introduction to Computational Chemistry*, John Wiley & Sons, 2nd edn, 2007.
25. F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke and K.-R. Müller, *Nat. Commun.*, 2017, **8**, 872.

View Online

26. M. Welborn, L. Cheng and T. F. Miller III, *J. Chem. Theory Comput.*, 2018, **14**, 4772.
27. L. Cheng, M. Welborn, A. S. Christensen and T. F. Miller III, *J. Chem. Phys.*, 2019, **150**, 131103.
28. K. Vijayalakshmi, V. Shankar, R. M. Bain, R. Nolley, G. A. Sonn, C.-S. Kao, H. Zhao, R. Tibshirani, R. N. Zare and J. D. Brooks, Identification of Diagnostic Metabolic Signatures in Clear Cell Renal Cell Carcinoma Using Mass Spectrometry Imaging, *Int. J. Cancer*, 2020, **147**, 256.
29. Z. Pozun, K. Hansen, D. Sheppard, M. Rupp, K.-R. Muller and G. Henkelman, *J. Chem. Phys.*, 2012, **136**, 174101.
30. J. Behler, *J. Chem. Phys.*, 2016, **145**, 170901.
31. P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier and A. J. Norquist, *Nature*, 2016, **533**, 73.
32. F. Ren, L. Ward, T. Williams, K. J. Laws, C. Wolverton, J. Hattrick-Simpers and A. Mehta, *Sci. Adv.*, 2018, **4**, eaaq1566.
33. Z. W. Ulissi, A. J. Medford, T. Bligaard and J. K. Nørskov, *Nat. Commun.*, 2017, **8**, 14621.
34. K. T. Schütt, F. Arbabzadah, S. Chmiela, K.-R. Müller and A. Tkatchenko, *Nat. Commun.*, 2017, **8**, 13890.
35. A.-C. Bédard, A. Adamo, K. C. Aroh, M. Grace Russell, A. A. Bedermann, J. Torosian, B. Yue, K. F. Jensen and T. F. Jamison, *Science*, 2018, **361**, 1220–1225.
36. F. Dyson, *Nature*, 2004, **427**, 297.
37. J. Mayer, K. Khairy and J. Howard, *Am. J. Phys.*, 2010, **78**, 648.
38. J. L. Murray, The aluminium-copper system, *Int. Metals Rev.*, 1985, **30**(1), 211.

Downloaded by Stanford Libraries on 7/20/2020 7:06:37 PM.
Published on 15 July 2020 on https://pubs.rsc.org | doi:10.1039/9781839160233-00512