Current Opinion in Solid State and Materials Science xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Current Opinion in Solid State and Materials Science

journal homepage: www.elsevier.com/locate/cossms

Atomistic calculations and materials informatics: A review

Logan Ward, Chris Wolverton *

Department of Materials Science and Engineering, Northwestern University, Evanston, IL, USA

ARTICLE INFO

Article history:
Received 4 April 2016
Revised 22 July 2016
Accepted 29 July 2016
Available online xxxx

Keywords:
Atomistic simulations
Materials informatics
Machine learning

ABSTRACT

In recent years, there has been a large effort in the materials science community to employ materials informatics to accelerate materials discovery or to develop new understanding of materials behavior. Materials informatics methods utilize machine learning techniques to extract new knowledge or predictive models out of existing materials data. In this review, we discuss major advances in the intersection between data science and atom-scale calculations with a particular focus on studies of solid-state, inorganic materials. The examples discussed in this review cover methods for accelerating the calculation of computationally-expensive properties, identifying promising regions for materials discovery based on existing data, and extracting chemical intuition automatically from datasets. We also identify key issues in this field, such as limited distribution of software necessary to utilize these techniques, and opportunities for areas of research that would help lead to the wider adoption of materials informatics in the atomistic calculations community.

1. Introduction

Over the past several decades, atomistic calculations have become increasingly prevalent tools in materials science. These calculations start from a description of the physical interactions between atoms in a solid, and ultimately describe their collective behavior. The fundamental nature of these methods makes them powerful tools for studying the behavior of materials with few assumptions about underlying mechanisms. With the advent of first-principles techniques based on Density Functional Theory (DFT), it is even possible to accurately and efficiently compute the interaction between atoms purely from quantum mechanics [1]. In general, the versatility of atomistic calculation techniques has allowed them to be applied to a broad spectrum of materials problems, including the design of aluminum alloys and predicting the properties of materials formed deep within the core of the Earth [1]. More recently, the ability to efficiently predict the electronic structure of materials from first-principles and advancements of computational power and algorithmic efficiency has enabled "high-throughput" calculations, which involve the predicting the performance of thousands of materials within single studies [2–5].

While powerful tools, atomic-scale calculations are intrinsically limited by their computational cost, which places strong limits on the space and time scales accessible to calculations. While calculations containing trillions of atoms or spanning over a millisecond have been demonstrated [6,7], they are not yet routine and only possible with simplistic models for the interatomic interactions. More accurate techniques (e.g., DFT), can require orders of magnitudes more computational resources. Practically, the large computational cost of atomistic calculations limits the number of materials that can be assessed and, therefore, the size of the materials design space. While several groups have used DFT to evaluate the T = 0 K stability of hundreds of thousands of materials [3–5,8], computing the thermal conductivity, which has a much greater computational cost, has only be accomplished for one hundred [9].

One path to lowering the computational costs and expanding the utility of atomistic calculations is to make better use of the results of previous calculations. Recently, there have been many significant advancements in the application of machine learning to materials science – a field often described as "Materials Informatics" [10]. The goal of materials informatics methods, broadly, is to extract knowledge from large datasets of materials properties. This knowledge can take many forms. For example, the knowledge could be a predictive model for a complex material property based on simple, easier-to-compute properties of the material. Or, it could be a small set of previously-unknown factors that help explain materials behavior. Of course, these activities are often the goals of conventional scientific practice. The power of materials informatics is that creating these models and learning these descriptors can be done quickly and automatically.

In this review, we will discuss the intersection between materials informatics and atomistic calculations with a particular focus

* Corresponding author.
E-mail address: c-wolverton@northwestern.edu (C. Wolverton).

http://dx.doi.org/10.1016/j.cossms.2016.07.002
1359-0286/

on solid, inorganic materials. There has been a great amount of recent work using machine learning in the atomistic simulation of organic materials [11–21] and Metal Oxide Frameworks (MOFs) [22–24], which do not fall within the intended scope of this review. Additionally, we will only focus on a few illustrative examples in detail rather than exhaustively covering the entire breadth of this field. In particular, we will cover several distinct areas that have seen significant advancements in recent years, including the development of techniques to replace *ab initio* calculations with faster surrogate models, using machine learning models to guide the search for new compounds, and extracting intuitive rules from collections of atomistic calculation data. There are topics that we will only allude to such as using machine learning to create density functionals [25–29] and machine-learning-based empirical potentials, which have been reviewed elsewhere [30]. Finally, we will conclude with a discussion of outstanding challenges and opportunities in applying machine learning to enhance atomistic calculations.

2. Principles of materials informatics

All materials informatics approaches are based on three distinct components: (1) a resource of materials data, (2) a representation to quantitatively describe each material, and (3) a machine learning algorithm to discover patterns within the data. The construction of databases of materials properties is a grand challenge in its own right [31–34], and covered in a review by Campbell in this issue [35]. The technology behind algorithms that learn predictive models from data are also described well in many other resources [36,37]. Consequently, for the purposes of this review, much of our discussion will focus on the second ingredient: representations.

The representation of a material is a set of quantitative attributes that describe its relevant characteristics. The attributes included in a representation are what serve as input to the models produced by machine learning. As discussed in Refs. [38,39], a representation should fulfill several requirements:

1. **Complete:** Features of a material relevant to the problem being studied should be captured.

The goal in creating a "complete" representation is to provide enough information to sufficiently differentiate materials. Depending on the target application, it may be necessary to describe different aspects of a material. For example, if all materials in the dataset have the same crystal structure, it may be possible to differentiate the materials in the dataset by their composition. In other cases, it could be necessary to include both the crystal structure and the processing conditions.

2. **Descriptive:** Similar materials should have similar representations.

The approach for creating descriptive representations is to introduce attributes that reflect physical intuition about what factors influence the property of interest. For example, the difference in electronegativity between elements in a compound is correlated to its formation energy. By including the electronegativity difference between constituent as an attribute, it is possible for the machine learning algorithm to recognize the strong relationship between this parameter and the formation energy, and use it to create a predictive rule.

3. **Simple:** Computing the representation should be fast.

The speed of a machine learning model includes both the time to compute the representation and evaluate the machine learning model. The combination of these two calculations should be faster than the method used to create the training set for the model.

4. **Unique:** All materials should have exactly one representation.

If a single material has multiple representations, it is possible to predict different properties for the same material. Unique representations, however, are not a strict requirement. Montavon et al., for example, have accommodated a non-unique representation by including multiple representations of each material in the training set and estimating the average of all representations for a material as the output of the model [12].

Even in the limited context of the application of materials informatics to atomistic calculations, choices in representation can vary drastically. For one, in some problems it may be sufficient to differentiate materials based on the composition and others it is necessary to include the atomic structure. Or, it may even be necessary to introduce the electronic charge density into the representation [29,40]. Also, the maximum acceptable time required to compute the representation may vary. Viewed together, the studies described in this work demonstrate many innovative approaches and general principles in the design of representations.

The final step of creating a machine learning model is the selection of an appropriate learning algorithm. There are a wide variety of available machine learning algorithms, as demonstrated by the Weka [41] and scikit-learn [42] software packages, and no established way to know which is the best for a certain problem *a priori*. Usually, the process of algorithm selection involves using cross-validation techniques to find which algorithm is likely to have the best predictive accuracy (a process that is automatable [43]). However, one must also consider factors such as interpretability, training and evaluation speed, differentiability, and ability to provide robust uncertainty estimates. For example, compressed sensing algorithms can lead to simple, easy-to-interpret linear models that are well-suited for understanding underlying physics [39,44]. Ensembles of decision trees, such as random and rotation forest [45,46], lead to strong predictive accuracy but do not provide a human-understandable model and do not have continuous derivatives [47,48]. Gaussian Process Regression models are differentiable and provide uncertainty estimates for individual predictions, which is useful when using ML for optimization [9,49–52], but have a training time that scales $O(N^3)$ with the dataset size, making it difficult to use with large datasets [53]. As alluded to in the following sections, many of these algorithms have been successfully employed to learn from atomistic calculations.

3. Key applications of materials informatics in atomistic calculations

In this section, we describe several broad areas where materials informatics has been applied to atomistic calculations. First, we discuss examples where machine learning was applied to replace computationally expensive *ab initio* calculations with faster, surrogate models. Then, we describe examples of where machine learning has been used to predict yet-undiscovered crystalline compounds. Additionally, we show how machine learning has been employed to extract knowledge about material behavior out of datasets created using atomistic calculations.

3.1. Replacing ab initio calculations with faster models

Ab Initio calculations offer the ability to compute the properties of materials with minimal experimental input – but at a large computational cost. While single calculations of the electronic structure with DFT can require only 10s of CPU-minutes for

Please cite this article in press as: L. Ward, C. Wolverton, Atomistic calculations and materials informatics: A review, Curr. Opin. Solid State Mater. Sci. (2016), http://dx.doi.org/10.1016/j.cossms.2016.07.002