Materials Advances

ROYAL SOCIETY OF CHEMISTRY

EDITORIAL

View Article Online
View Journal | View Issue

Cite this: *Mater. Adv.*, 2023, 4, 2695

DOI: 10.1039/d3ma90047a

rsc.li/materials-advances

Introduction to Materials Informatics

Krishna Rajan,[a] Jörg Behler[bc] and Chris J. Pickard*[de]

Materials Informatics has emerged from a fusion of the increasing availability of materials data, high throughput experimental and computational methods, first principles and other advanced materials models, and machine learning. It has been fuelled by the dramatic growth in available computational power, and its ubiquity. Materials Informatics is extremely interdisciplinary, drawing from mathematics, computer science and statistics, as well as the physical sciences. The papers brought together in this themed issue capture the nexus of these fields, whether they report new discoveries in materials science or advances in methodologies that enable such discoveries.

Although large amounts of data are often synonymous with the term *informatics*, these papers serve to remind us that the acceleration of knowledge discovery in materials science is the true goal. This could be expressed in a variety of objectives, from finding ways to reduce the amount of data needed to build machine-learning-based models, to accelerating the computational search process, or pushing the computational methods to their limits. With the advancements in high throughput first-principles-based structure analysis and prediction and the development of accessible electronic structure databases, the application of machine-learning methods on such data sets has now become a rich source of scientific exploration. In parallel, there is a growing number of important studies involving machine learning applied to experimentally derived data, and/or simulation or prediction of experimental observables.[1,2] When working with experimental data and/or the simulation of experimental data, one must account for the nature of precision and accuracy in the measurement itself (*e.g.*, specimen geometry, instrumental parameters, and the conditions of observation).[3] Incorporating all these factors can make evaluation and interpretation of spectroscopic profiles based on human-identifiable peaks difficult

[a] *Department of Materials Design and Innovation, University at Buffalo, Buffalo, NY 14260, USA*
[b] *Lehrstuhl für Theoretische Chemie II, Ruhr-Universität Bochum, 44780 Bochum, Germany*
[c] *Research Center Chemical Sciences and Sustainability, Research Alliance Ruhr, 44780 Bochum, Germany*
[d] *Department of Materials Science & Metallurgy, University of Cambridge, Cambridge, UK*
[e] *Advanced Institute for Materials Research, Tohoku University, Sendai, Japan*

Krishna Rajan

Jörg Behler

Chris J. Pickard

and convoluted. This is where machine-learning methods can be a powerful aid to probe structure–property relationships from experimental data.[4] From exploring novel materials discovery under extreme conditions to systems with complex and novel bonding characteristics, these papers demonstrate the continued opportunities for discovery and design in materials science using Materials Informatics.

Atomistic simulations have been a cornerstone of materials science for a long time, as they provide atomic-level information, which is often difficult to directly obtain in purely experimental studies. However, the quality of the results obtained in these simulations depends on the accuracy of the underlying multi-dimensional potential energy surface describing the atomic interactions. Over recent years, machine learning has been increasingly used to develop interatomic potentials, and the resulting machine learning potentials have enabled large-scale simulations of increasingly complex systems with first-principles accuracy.[5–7] Most of these machine learning potentials make use of atomic energies, but also flexible atomic charges can be learned to compute long-range electrostatic interactions. Moreover, other properties that are usually obtained from electronic structure calculations can nowadays be predicted with good accuracy by machine learning, such as magnetic resonance parameters.

In this issue, Burn and Popelier (**https://doi.org/10.1039/D2MA00673A**) contribute a report on ICHOR, a framework for generating Gaussian process regression (GPR) models through active learning. The GPR can be used to express atomic properties, such as energies or electrostatic multipole moments, for constructing machine-learning force fields, *e.g.*, FFLUX, using atomic local frames to describe the atomic environments. It allows for an automatic workflow, from electronic structure calculations to the training process facilitating the construction of interatomic potentials.

While machine-learning potentials usually do not provide direct information about the electronic structure, such properties are also accessible by machine learning, as demonstrated in the contribution by Mazouin, Schöpfer and von Lilienfeld (**https://doi.org/10.1039/D2MA00742H**). They describe a procedure to efficiently predict electronic properties, such as the band gap, by explicitly considering structural features of the investigated molecules. This allows a dramatic reduction in the training data set required as compared to general-purpose approaches trained on highly diverse sets of molecules.

Materials Informatics is founded on models for the microscopic, atomistic, structure of materials. Throughout the 20th Century, our structural models were almost entirely derived from experiment, most notably from X-ray crystallography. From the early recognition that the wavelength of the newly discovered "X" rays was close to the hypothesized distances between atomic planes in crystals, to the large-scale solution of protein and material crystal structures, crystallography has dominated the atomistic sciences. National and international synchrotron facilities are dedicated to solving crystal structures.

This wealth of data has found its way into important, curated, databases such as the Cambridge Structural Database (CSD) and the Inorganic Crystal Structure Database (ICSD). From the start of the 21st Century, these databases fuelled the high throughput computation of materials properties from first principles, using modern, robust, density functional theory (DFT)-based computer codes. Further databases of computed properties, including for species-swapped hypothetical compounds, such as the Materials Project, have transformed materials informatics, and the materials sciences more generally.[8]

In parallel, the advent of first-principles structure prediction has provided a way to generate structural data independently of experiment. These methods are based on the stochastic generation of candidate structures, and either their high-throughput parallel optimisation in methods such at *ab initio* random structure search (AIRSS), or more sophisticated, but not necessarily more performant, evolutionary or nature-inspired algorithms, as implemented in USPEX, and Calypso.[9]

These methods made their biggest initial impact in the field of high-pressure research, where experiments are challenging, and little structure data was available.[10] But with improvements to the methods, and increasing computational resources, more realistic and relevant materials are investigated in state-of-the-art studies, such as those highlighted in this themed issue.

In the manuscript contributed by Liu *et al.* (**https://doi.org/10.1039/D2MA00937D**), we learn about the computational design of a promising family of two-dimensional phosphorus trichalcogenides for photovoltaic applications. The authors construct a large set of candidate structures by considering multiple permutations of a set of elements. Including spin-orbit and many-body effects on the electronic properties, and a consideration of thermodynamic stability, they predict that several of the compositions exhibit promising properties for a variety of applications.

The importance of materials with sustainable applications is clear, and Finkler and Goedecker (**https://doi.org/10.1039/D2MA00958G**) describe a study that investigates another material with photovoltaic applications – methylammonium lead iodide (MaPbI), a hybrid perovskite that has received great attention due to its extraordinary properties. Hybrid perovskites are extremely challenging to describe computationally, due to their structural richness and compositional complexity, and recently competing non-perovskite phases have been identified. Finkler and Goedecker explore this system by combining modern neural-network potentials, to accelerate the evaluation of the energy landscape, with Monte Carlo Funnel Hopping to explore the rich landscape.

Ding *et al.* (**https://doi.org/10.1039/D2MA01012G**) describe an exploration of the relatively unknown chemistry of cerium and nitrogen under pressure. They construct the binary convex hull of the Ce–N system to identify stable and metastable phases at pressures of 100 GPa. This pressure is approximately one third of the pressure at the centre of the Earth, but modern diamond anvil cell experiments can regularly reach such pressures. They use MAGUS, a machine learning accelerated structure prediction method, to tackle this complex system and identify compounds containing nitrogen in a variety of forms, including chains.

Liao *et al.* (**https://doi.org/10.1039/D2MA00920J**) describe an approach to first-principles structure prediction that

exceeds the limits of automated approaches. Carbon is an iconic element, with a rich variety of structures. The SACADA database lists hundreds of established and putative carbon polymorphs, and there are surely more to come. The field of carbon studies contains many mysteries, and many of those involve forms of carbon that have been brought to Earth on meteorites or created on their arrival. One such mystery is the nature of a transparent, large-unit-cell, cubic phase of carbon found in the Popigai crater. The authors construct complex candidate models, currently out of reach of even machine-learning accelerated structure prediction, and compute their properties to claim consistency with many of the measured properties.

References

1 H. Schopmans, P. Reiser and P. Friederich, Neural networks trained on synthetically generated crystals can extract structural information from ICSD powder X-ray diffractograms, *arXiv*, 2023, preprint, arXiv:2303.11699, DOI: 10.48550/arXiv.2303.11699.
2 R. Dong, Y. Zhao, Y. Song, N. Fu, S. S. Omee, S. Dey, Q. Li, L. Wei and J. Hu, DeepXRD, a Deep Learning Model for Predicting XRD spectrum from Material Composition, *ACS Appl. Mater. Interfaces*, 2022, **14**(35), 40102–40115.
3 J. R. Helliwell, Combining X-rays, neutrons and electrons, and NMR, for precision and accuracy in structure–function studies, *Acta Crystallogr., Sect. A: Found. Adv.*, 2021, **77**(3), 173–185.
4 D. Vizoso, G. Subhash, K. Rajan and R. Dingreville, Connecting Vibrational Spectroscopy to Atomic Structure *via* Supervised Manifold Learning: Beyond Peak Analysis, *Chem. Mater.*, 2023, **35**(3), 1186–1200.
5 E. Kocer, T. W. Ko and J. Behler, Neural network potentials: A concise overview of methods, *Annu. Rev. Phys. Chem.*, 2022, **73**, 163–186.
6 P. Friederich, F. Häse, J. Proppe and A. Aspuru-Guzik, Machine-learned potentials for next-generation matter simulations, *Nat. Mater.*, 2021, **20**(6), 750–761.
7 F. Noé, A. Tkatchenko, K.-R. Müller and C. Clementi, Machine learning for molecular simulation, *Annu. Rev. Phys. Chem.*, 2020, **71**, 361–390.
8 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek and S. Cholia, *et al.*, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, *APL Mater.*, 2013, **1**(1), 011002.
9 A. R. Oganov, J. P. Chris, Q. Zhu and R. J. Needs, Structure prediction drives materials discovery, *Nat. Rev. Mater.*, 2019, **4**(5), 331–348.
10 Y. Wang and Y. Ma, Perspective: Crystal structure prediction at high pressures, *J. Chem. Phys.*, 2014, **140**(4), 040901.