CPP 64: Topical Session: Data Driven Materials Science - Materials Data Management (joint session MM/CPP)

Time: Wednesday 10:15–11:30 Location: BAR 205

Topical Talk CPP 64.1 Wed 10:15 BAR 205

Automated atomistic calculation of thermodynamic and thermophysical data — •Jan Janssen, Tilmann Hickel, and Jörg Neugebauer — Max-Planck-Institut für Eisenforschung, Düsseldorf, Germany

A major challenge in predicting the properties of materials at realistic conditions is the accurate inclusion of finite temperature effects. Doing this on an ab initio level often requires complex simulation protocols. These complex protocols, which often couple several specialized codes, make a quantitative description of error propagation and uncertainty quantification a critical issue.

To handle this high level of complexity we have developed an integrated development environment (IDE) called pyiron[1] - http://pyiron.org. pyiron has been specifically designed to scale simulation protocols from the interactive prototyping level up to the high throughput level, all within the same software framework.

We highlight two recent success stories towards automated calculation of phase diagrams: We first discuss with the automated convergence for all key parameters in DFT codes, followed by the calculation of melting points with a guaranteed precision of better than 1K. These fully automated high-precision tools allow us to study trends over the periodic table in an efficient and systematic way. Examples how such high-throughput screenings allow to develop new strategies in designing materials will be given.

[1]: J. Janssen, et al., Comp. Mat. Sci. 161 (2019)

CPP 64.2 Wed 10:45 BAR 205

Big data in materials science: Status of and needs for metadata and ontologies — •Maja-Olivia Lenz, Luca M. Ghiringelli, Carsten Baldauf, and Matthias Scheffler — Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin

In recent years, the amount of data in materials science has increased exponentially. Consequently, new ways to store and annotate data are necessary to ensure findability, accessibility, interoperability and re-usability, i.e. to fulfil the FAIR principles [1], and to do efficient, good and new science. Data describing and characterizing other data are called metadata. Often, the materials science community has no clear distinction between data and their metadata as it depends on the intended use of the data. In this talk, we present the NOMAD MetaInfo [2], a general descriptive and structured metadata scheme for materials simulations. Ontologies represent the next step on the semantic ladder, as they enrich pure (meta)data structures by relations and thereby enable semantic and syntactic interoperability between different software agents, people, and organizations. In fact, the NOMAD MetaInfo includes a number of relations between concepts and therefore goes beyond the simple metadata picture. It can be interpreted as a light-weight ontology and thus can easily be connected to other ontologies like the European Materials and Modeling Ontology, EMMO. We give an introduction to ontologies, explain why they are useful, and outline their role and current status in materials science.

[1] M. Wilkinson, *et al.*, Sci Data 3, **160018** (2016).
[2] L. M. Ghiringhelli *et al.*, npj Comput. Mater. 3, **46** (2017).

CPP 64.3 Wed 11:00 BAR 205

Benchmarking neural networks on sequence-determined polymer transport through lipid membranes — •Marco Werner[1], Yachong Guo[2], and Vladimir Baulin[3] — [1]Institut Theorie der Polymere, Leibniz-Institut für Polymerforschung Dresden, Germany — [2]National Laboratory of Solid State Microstructure, Department of Physics, Nanjing University, China — [3]Departament d'Enginyeria Química, Universitat Rovira i Virgili, Tarragona, Spain

We consider the transport of amphiphilic polymers through lipid membranes by passive diffusion as a function of the sequence of hydrophilic and hydrophobic building blocks. Massively parallel Rosenbluth sampling of polymer conformations is performed to estimate polymer translocation times through a membrane for all 2^N sequences and chain lengths $N \leq 16$. Our results confirm that smallest translocation times are found for polymers with balanced fraction of hydrophilic and hydrophobic units, and containing short blocks. Sequence-complete databases deliver an important ground truth for benchmarking machine-learning models against training data restrictions and biases. We demonstrate that multi-layer artificial neural networks show remarkable generalization performance when restricting the training data to relatively narrow windows of translocation times. The results indicate that relevant sequence patterns and their physical effect are approximated based on the restricted training set, however, accuracy drops towards unexplored corners in sequence space.

CPP 64.4 Wed 11:15 BAR 205

Analysis of Materials Structural Representations for Machine Learning Interatomic Potentials — •Berk Onat[1], Christoph Ortner[2], and James Kermode[1] — [1]School of Engineering, University of Warwick, Coventry, United Kingdom — [2]Mathematics Institute, University of Warwick, Coventry, United Kingdom

Representations of materials based on atomic structural environments have been used either in machine learning models to predict properties directly or as the core of machine learning interatomic potentials (MLIPs) to enable accurate simulations. Many MLIPs have been developed to translate atomic neighbourhood environments from atom positions to structural representations such as atom-centred symmetry functions, smooth overlap of atomic positions and atomic cluster expansion(ACE) with spherical harmonics. While use of these representations is becoming common practice for applications, the sensitivity of their structural mapping to the materials composition and whether their coverage of the hyper-dimensional space is over-determined or complete have not yet been fully analysed. In this presentation, we provide analysis of the invariance of the model transformation under translations and rotations as well as the sensitivity of descriptors to perturbations. A range of datasets extracted from the NOMAD Archive are used to assess the dimensionality of the representations. The outcomes of our analyses will be presented with discussions on the model sensitivities and their possible limitations. We further provide insights on our continuing affords to utilise structural representations in other models for data-driven materials modelling.