MM 70: Topical Session (Symposium MM): Big Data in Materials Science - Managing and exploiting the raw material of the 21st century

Big Data VI

Time: Friday 9:30–11:00 Location: H 0107

Topical Talk MM 70.1 Fri 9:30 H 0107

Transmission Electron Microscopes as a tool generating Big Data: challenges ans opportunities — •Cécile Hébert — LSME, Institut de Physique, Ecole Polytechnique Fédérale de Lausanne, Switzerland

Modern state of the art transmission electron microscopes have become versatile tools fitted with a great variety of detectors. Thanks to tremendous improvement in the stability of components, a single instrument can be operated in various modes. Typically it is possible to operate it in conventional TEM mode or in scanning TEM mode. In TEM mode, cameras are used to reccord images or diffraction patterns. In STEM mode, detectors are used to collect signal as a function of probe position. This signal can be the scattered electrons at various angles, the direct electron beam, eventually analyzed in energy (electron energy loss spectrometry), X-Rays emmited by the specimen or even full diffraction pattern at each probe position. With the ability to scan area of 100 to several 1000 squared pixel, rapid CCD camera (1000 fps) for dynamical experiments, every mid-sized lab can generate data volumes up to petabytes/year. An additional challenge is posed by the fact that several signals are captured by detectors of various brands and delivered in closed undocumented formats. With the current trend towards open science, this poses new challenges but give also opportunities to address the topic on a global level.

MM 70.2 Fri 10:00 H 0107

High-throughput classification and categorization of structures from atomistic simulations — •Lauri Himanen, Patrick Rinke, and Adam Foster — Department of Applied Physics, Aalto University, Espoo, Finland

Our capability of producing, storing and analysing computational materials science data has grown tremendously. As the high-throughput screening of materials is becoming ever more popular, materials databases are being filled with atomic and electronic structure data.

To enable structure-related database queries, specific structural classes need to be defined. Unfortunately the required information is not always provided, and when it is, it is often based on an unspecified definition. To cope with large heterogeneous datasets of atomistic calculations, automated and verifiable methods for analyzing and categorizing atomistic structures are becoming necessary.

We discuss different methods that can be used in extracting structural information from various structural classes. These techniques involve finding a standardized unit cell, finding a translational basis for periodic materials within complex atomic environments and cluster analysis for separating different structural components. We also propose a material map that can be used to categorize the structural space and apply the introduced methods in the automatic classification of pristine crystals, surfaces and 2D materials.

MM 70.3 Fri 10:15 H 0107

Compact representation of crystal structures using three-dimensional diffraction patterns and deep learning — •Angelo Ziletti, Matthias Scheffler, and Luca M. Ghiringhelli — Fritz Haber Institute of the Max Planck Society, Berlin, Germany

Big data is emerging as a new paradigm in materials science. A vast amount of three-dimensional structural data is provided by both computational repositories (e.g. http://nomad-coe.eu) and experiments (e.g. atom probe tomography). Computational methods that automatically and efficiently detect long-range order are of paramount importance for materials characterization and analytics. Current methods are either not stable with respect to defects, or base their representation on local atomic neighbourhoods, which in turn makes it difficult to detect "average" longe-range order. In the proposed approach, for a given crystal structure we first calculate its diffraction pattern, expand it on spherical harmonics, and then use a neural-network model to obtain a compact, low-dimensional representation. We apply this workflow to a subset of materials from the Novel Materials Discovery (NOMAD) Archive, and show that our deep-learning-based approach compactly encodes structural information, is robust to defects (e.g. point defects, and/or strain), and allows to build easily interpretable structural-similarity maps. This work received funding from the NOMAD Laboratory, a European Center of Excellence.

MM 70.4 Fri 10:30 H 0107

Cluster analysis of chemical libraries based on molecular fingerprinting — •Annika Stuke[1], Lei Xie[2], Milica Todorović[1], and Patrick Rinke[1] — [1]Department of Applied Physics, Aalto University, Finland — [2]Department of Computer Science, Hunter College, the City University of New York, USA

Machine learning models promise to greatly accelerate the process of discovering new and better materials. However, it is difficult for learning models to achieve a robust and high prediction performance with imbalanced chemical datasets, in which certain classes of chemical structures are overrepresented. Learning algorithms are easily influenced by the larger classes, leading to biased results. We present an efficient method to generate diverse subsets from large chemical databases with cluster analysis. Databases are split into different clusters with an extended exclusion sphere algorithm based on the pairwise Tanimoto similarity calculated from Morgan fingerprints [1]. A diverse subset is then generated by picking molecules with different substructures from each cluster. The method has been successfully employed to select structurally diverse subsets of a dataset of 64k organic molecules from the Cambridge Crystal Structure Database [2]. We demonstrate the effect of this method on the prediction performance of machine learning models based on kernel ridge regression and neural networks for spectral properties of molecules. [1] D. Butina, J. Chem. Inf. Comput. Sci. 39, 747 (1999), [2] C. Schober et al., J. Phys. Chem. Lett. 7, 3973 (2016)

MM 70.5 Fri 10:45 H 0107

Identifying synthesisable ice structures from first principles — •Edgar Engel[1,2], Andrea Anelli[1], Michele Ceriotti[1], Chris Pickard[2], and Richard Needs[2] — [1]TCM Group, Cavendish Laboratory, UoCambridge, UK — [2]Laboratory of Computational Science and Modeling, IMX, EPFL, Lausanne, Switzerland

We present a comprehensive density-functional-theory study of the crystalline phases of water ice. We construct candidate ice structures on the basis of more than five million tetrahedral networks listed in the Treacy, Deem, and IZA databases, collecting 15,882 locally-stable ice structures. The search for the few synthesisable structures among them is a needle-in-a-haystack kind of problem, which is conventionally tackled using a convex hull construction to identify structures which are stabilised by manipulation of a particular constraint (such as density) chosen on the basis of experimental evidence or intuition. This heavily constrains which stabilisable structures are identified and does not account for the uncertainties inherent to computed structure properties. Hence, we instead employ a recently developed probabilistic generalised convex hull construction to stochastically sample the likelihood of each structure to be stabilised by application of appropriate thermodynamic constraints. We thereby recover (entirely a priori) all known ice phases except the known-to-be metastable ice IV. We further identify several new promising candidates for experimental synthesis, providing a much needed starting point for the determination of accurate structural properties and possible synthetic pathways.