SYBD 1: Big Data Driven Materials Science

Time: Tuesday 9:30–12:15 Location: HSZ 02

Invited Talk SYBD 1.1 Tue 9:30 HSZ 02
Materials innovation driven by data and knowledge systems — •Surya Kalidindi — Georgia Institute of Technology, Atlanta, USA

Emerging concepts and toolsets in Data science and Cyberinfrastructure can be strong enablers for systematic mining and capture of Materials Knowledge needed to guide efficient and possibly autonomous explorations of the unimaginably large materials and process design spaces, while synergistically leveraging all available experimental and simulation data. The ongoing efforts in my research group are aimed at accelerating materials innovation through the development of (i) a new mathematical framework that allows a systematic and consistent parametrization of the extremely large spaces in the representations of the material hierarchical structure (spanning multiple length/structure scales) and governing physics across a broad range of materials classes and phenomena, (ii) a new formalism that evaluates all available next steps in a given materials innovation effort (i.e., various multiscale experiments and simulations) and rank-orders them based on their likelihood to produce the desired knowledge (expressed as PSP linkages), and (iii) novel higher-throughput experimental assays that are specifically designed to produce the critically needed fundamental materials data for calibrating the numerous parameters typically present in multiscale materials models. I will present and discuss ongoing research activities in my group.

Invited Talk SYBD 1.2 Tue 10:00 HSZ 02
Network Theory Meets Materials Science — •Chris Wolverton[1], Murat Aykol[2], and Vinay Hegde[3] — [1]Northwestern University, Evanston, IL, USA — [2]Toyota Research Institute, Los Altos, CA, USA — [3]Citrine Informatics, Redwood City, CA, USA

One of the holy grails of materials science, unlocking structure-property relationships, has largely been pursued via bottom-up investigations of how the arrangement of atoms and interatomic bonding in a material determine its macroscopic behavior. Here we consider a complementary approach, a top-down study of the organizational structure of networks of materials, based on the interaction between materials themselves. We demonstrate the utility of applying network theory to materials science in two applications: First, we unravel the complete *phase stability network of all inorganic materials* as a densely-connected complex network of 21,000 thermodynamically stable compounds (nodes) interlinked by 41 million tie-lines (edges) defining their two-phase equilibria, as computed by high-throughput density functional theory. Using the connectivity of nodes in this phase stability network, we derive a rational, data-driven metric for material reactivity, the nobility index, and quantitatively identify the noblest materials in nature. Second, we apply network theory to the problem of synthesizability of inorganic materials, a grand challenge for accelerating their discovery using computations. We use machine-learning of our network to predict the likelihood that hypothetical, computer generated materials will be amenable to successful experimental synthesis.

Invited Talk SYBD 1.3 Tue 10:30 HSZ 02
Verification and error estimates for ab initio data — •Claudia Draxl — Humboldt-Universität zu Berlin, Germany — Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin, Germany

Veracity (uncertainty of data quality), one of the 4V challenges of Big Data, is an issue for the FAIRness of (computational) materials-science results. Creating benchmark data and estimating errors are prerequisites for the interoperability of our research data. The precision of the many different computer codes used in the community has been investigated thoroughly by evaluating the equation of state of 71 monoatomic crystals [1]. More recently, it has been demonstrated how ultimate precision for molecules and solids in DFT calculations can been reached [2] and how different methodology impacts the results [3]. We also address code-specific uncertainties that come from numerical settings commonly used in practice [4]. We do so by systematically investigating total and relative energies as a function of computational parameters, employing four popular DFT codes. Based on this, we propose an analytical model for quantifying errors associated with the basis-set incompleteness and predicting converged results. It will be discussed how our approach enables comparison and interoperability of the heterogeneous data present in computational materials databases [5], for the purpose of data-driven research.

[1] K. Lejaeghere et al., Science 351, aad3000 (2016). [2] A. Gulans, A. Kozhevnikov, and C. Draxl, Phys. Rev. B 97, 161105(R) (2018). [3] A. Gulans and C. Draxl, preprint. [4] C. Carbogno, et al., preprint. [5] https://nomad-repository.eu

15 min. break

Invited Talk SYBD 1.4 Tue 11:15 HSZ 02
Identifying Domains of Applicability of Machine Learning Models for Materials Science — •Mario Boley[1], Christopher Sutton[2], Luca M. Ghiringhelli[2], Matthias Rupp[3], Jilles Vreeken[4], and Matthias Scheffler[2] — [1]Monash University, Melbourne, Australia — [2]Fritz Haber Institute of the Max Planck Society, Berlin, Germany — [3]Citrine Informatics, Redwood City, California — [4]Helmholtz Center for Information Security, Saarbrücken, Germany

Machine learning (ML) promises to accelerate the discovery of novel materials by allowing to rapidly screen compounds at orders of magnitude lower computational cost than first-principles electronic-structure approaches. A critical obstacle for the development of novel ML models is that the complex choices involved in designing them are currently made based on the simplistic metric of the average model test error. Treating models as a black box that produces a single error statistic can render them as insufficient for certain screening tasks while they actually predict the target property accurately in sub-domains of the considered materials. We present an alternative diagnostic tool based on subgroup discovery that detects domains of applicability of ML models. These domains are given as a combination of simple conditions on the unit cell structure (e.g., on the lattice vectors, lattice angles, and bond distances) under which the model error is substantially lower than its global average in the complete materials class. Such descriptions allow to understand and subsequently address systematic shortcomings of the investigated ML model and to focus sampling of candidate materials to regions of low expected error.

Invited Talk SYBD 1.5 Tue 11:45 HSZ 02
Deep learning of low-dimensional latent space molecular simulators — •Andrew Ferguson — Pritzker School of Molecular Engineering, University of Chicago, Chicago, IL 60637

The long-time microscopic evolution of molecular systems is governed by the leading eigenfunctions of the transfer operator that propagates the system dynamics through time. The low-dimensional latent space defined by these eigenfunctions parameterize the slow manifold to which the system dynamics are constrained to evolve. A set of three deep neural networks of different architectures trained over short molecular simulation trajectories provides a means to (i) learn the leading transfer operator eigenfunctions, (ii) propagate the dynamics within the encoded latent space, and (iii) decode the latent space back to the all-atom coordinate space. This technique offers a means to train numerical simulators to conduct molecular simulations and estimate thermodynamic and kinetic observables at orders-of-magnitude lower cost than conventional molecular dynamics calculations.