IPAM Program on Machine Learning & Many-Particle Systems - *Recent Progress and Open Problems*

A. Tkatchenko, M. Afzal, C, Anderson, T. Baker, R. Banisch, S. Chiama, C. Draxl, M. Haghighatlari, F. Heidar-Zadeh, M. Hirn, J. Hoja, O. Isayev, R. Kondor, L. Li, Y. Li, G. Martyna, M. Meila, K.S. Ruiz, M. Rupp, H. Sauceda, A. Shapeev, M. Stöhr, K.-R. Müller, S. Shankar

March 1, 2017

1. Overall summary

Interactions between many constituent particles, i.e. quarks, electrons, atoms, molecules, or materials, generally give rise to collective or emergent phenomena in matter. Even when the interactions between the particles are well defined and the governing equations of the system are understood, the collective behavior of the system as a whole does not trivially emerge from these equations. Despite many decades of prominent work on interacting many-particle (MP) systems, the problem of N interacting particles is not exactly soluble. In fact, computational complexity typically increases exponentially with N. Although many attempts have been made to define collective (emergent) variables in numerous fields of natural sciences, the progress is often painfully slow. This situation stems from the lack of easily identifiable symmetries in complex dynamical systems such as materials, chemicals, and proteins. In fact, the identification and understanding of descriptive collective variables is among the most time-consuming and rewarding processes in a multitude of sciences. In this context, the main goal of the proposed program was to develop and apply novel machine learning (ML) methods to significantly accelerate the discovery of descriptive variables in complex MP systems at the microscopic scale. Examples of collective behavior are abundant in nature, manifesting themselves at all scales of matter, ranging from atoms to galaxies. Examples of collective behavior include spontaneous assembly of organic and inorganic crystalline structures on surfaces and in the bulk, self-assembly of proteins and DNA in cells, the behavior of human and animal crowds, the dynamics of sand dunes, formation of clouds, and formation of galaxies.

Machine learning methods have been used extensively in a wide variety of fields ranging from e.g. the neurosciences, genetics, multimedia search to drug discovery. ML models can be thought of as universal approximators that learn a (possibly very complex) non- linear mapping between input data (descriptor) and an output signal (observation). ML approaches are frequently applied as "black box" approximators but have been rarely used to learn new physical models for MP systems. Therefore, the aim of this IPAM long program was to develop the "black box" scientifically by bringing together experts in MP problems in condensed-matter physics, materials, chemistry, and protein folding, together with experts in mathematics and computer science. This helped address the problem of tackling emergent behavior and understanding the underlying collective variables in MP systems. Only collaborations during and after the program, between researchers in these areas could lead to breakthroughs in our understanding of complex emergent behavior in MP systems.

The combination of ML with atomistic simulations is an emerging and quickly growing field that brings many challenges as well as potential opportunities. As such, it is clearly impossible to cover all topics of interest discussed during the long program in this document. We have organized the document around important areas of research that were prominently discussed during the long program.

2. Topics

In the next few sections, we will summarize many of the exciting discussions and developments that was part of the 2016 program. The sections are divided into the following six sections: Machine Learning Models and Representations; Many-Body Interactions; Collective Variables and Nonlinear Dimension Reduction; Potential-Energy Landscapes; Benchmarks and Data Repositories; and Nano-Engineering. In each of the sections the report summarizes the background, applications, and future problems. In the concluding section, we summarize the overall program summary and potential directions for future research.

2.1: Machine learning models and representations: Kernels, networks and graphs

R. Banisch, M. Haghighatlari, M. Hirn, O. Isayev, R. Kondor, M. Meila, M. Rupp

2.1.1: Background

A fundamental problem in every machine learning task is finding an adequate representation of the data that exposes the features most important for the scientific problem. In the case when data is discrete, geometric, or takes other forms than finite dimensional vectors, it is necessary also to transform the data into vectors or distances, the natural inputs of most machine learning algorithms, while preserving the information we ultimately want to extract. Kernels do this by mapping into an expanded, and continuous, feature space. Non-linear dimension reduction methods map the data from high-dimensions to a reduced set of descriptors, often called collective coordinates, that represent the slow modes. Deep learning architectures pass the data through a cascade of multiple linear transformations and nonlinear pooling operations, extracting complex multiscale information from the data.

The development of machine learning methods has exploded this century, with novel algorithms being applied to a multitude of domains including computer vision, natural language processing, and music. This program focused on the synergy between machine learning and physics, and in particular many particle systems such as molecules, crystals, and materials. Quantum-mechanical many-body methods are crucial for the computational study of such systems, but are limited by their prohibitive computational cost. Machine learning can significantly reduce those costs by on the one hand accurately interpolating between reference calculations, and on the other hand learning collective variables that succinctly characterize the landscape of such data. These methods potentially enable orders of magnitude improvement in the size of screened databases, the length of dynamics simulations, and the scale of investigated systems. Successful invention and application of these methods will significantly extend the reach of computational methods in nanoscale physics, chemistry and materials science.

Essential to the success of this program is the development of new machine learning approaches that obey the underlying physical laws of the system. For example, the total ground state energy of the system is invariant to translations, rotations and reflections,

and is stable to deformations of the relative positions of the atoms. In addition, force field models for closed systems must obey the law of energy conservation. New machine learning approaches that incorporate these physical constraints have been developed across kernel based algorithms, deep learning architectures, and unsupervised learning.

2.1.2 Recent Developments

While kernel based machine learning encompasses a wide area ranging from Bayesian methods such as Gaussian processes (closely related to kriging and kernel ridge regression) to support vector machines and structured output learning, a common feature of all of these algorithms is that the choice of kernel is critical to their success. In general, the kernel $K(x, x')$ needs to satisfy a combination of mathematical and domain specific requirements: (a) it must be a positive semi-definite function; (b) it must incorporate all the internal symmetries of x; (c) it must capture the right notion of similarity between any two datapoints x and x'; (d) it must be efficient to compute.

It was quickly realized that modeling many particle systems requires developing fundamentally new kernels, because in this field the nature of criteria (b) and (c) above are fundamentally different than in other branches of ML. Some of the participants of the program have made fundamental contributions to many particle kernels, and kernel design remained one of the recurring themes of discussion. To name just a few of the approaches:

1. Rupp (core participant) and coworkers pioneered the use of kernels based on the so-called Coulomb matrix , incorporating information about the charge and position of atoms making up a given atomic system or environment.

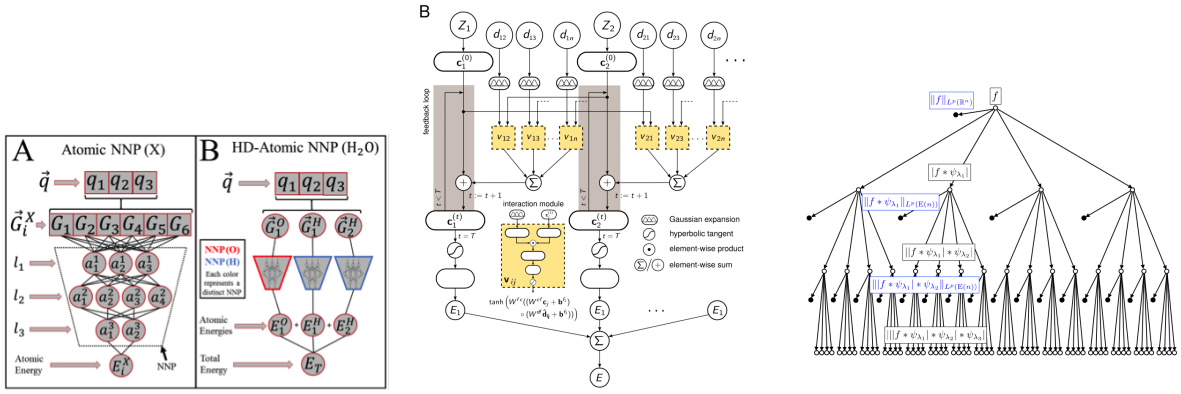
2. Kondor (core participant) and others directly use concepts from the representation theory of the underlying symmetry groups to build invariant kernels (bispectral kernels)

3. Bartok (Workshop III) et al introduced the so-called SOAP kernels , standing for smooth overlap of atomic potentials, which compute $K(x, x')$ by forming invariant functionals of an effective density induced by the atomic configuration.

4. Ferre (core participant) et al introduced a new framework for lifting graph kernels (which have a substantial literature in machine learning) to a kernel between atomic environments.

In the Coulomb matrix approach the kernel is constructed by defining a representation $\Phi(x)$ of the system x (here the Coulomb matrix), which satisfies the correct physical properties, and then defining the kernel on this representation. Ensuing synergistic work during the program resulted both in new insights into relationships between various representations and new ideas, including moment tensor potentials (A. Shapeev, core participant) and many-body tensor representations (M. Rupp, core participant).

Neural networks for potential energy surface fitting of one molecule or material predate this program by more than two decades. However, the recent re-emergence of deep learning architectures in machine learning has spurred the development of several new



(a) ANI-1 Network (b) Deep Tensor Network (c) Scattering Network

Figure 1: Deep learning architectures for many particle systems.

deep networks for the prediction of energies across chemical compound space. Aside from incorporating the aforementioned physical constraints, these networks capture the intrinsic multiscale nature of the problem through their multilayer architecture. The variety of approaches is illustrated in Figure 1, which shows three such networks:

1. ANI-1 neural network potentials (O. Isayev, core participant), which takes the inputted molecular coordinates and generates through the multilayer architecture an atomic environment vector for each atom.

2. Deep tensor neural networks (K. T. Schütt, core participant), which builds through the layers of the network a series of representations spanning local atomistic descriptions to global interaction coefficients.

3. Scattering transform networks (M. Hirn, core participant), which cascade multiscale wavelet transforms and nonlinear pooling operations applied to a density based representation of the molecule.

2.1.3 Future Directions

Within both kernel and deep learning based approaches, work over the next few years will focus on methodological development and application to more challenging systems such as multi-component alloys and transition metal oxides. Potential applications include hard problems such as crystal structure prediction. Approximation bounds, efficient learning rate guarantees, uncertainty quantifications, and characterizations of the space of physical functionals that the machine learning algorithms can learn are needed to better understand the synergistic relationship between physical models and machine learning. A challenge/necessity for the future is to incorporate more subtle physics into these methods such as appropriate conservation laws.

Given the emerging state of this community, these machine learning algorithms are often developed using different programming architectures, and can require a certain level of familiarity and expertise in order to be tweaked and used properly. In order to

facilitate broader dissemination of these tools, and to continue to grow the community of researchers working in machine learning for many particle systems, M. Haghighatlari (core participant) has compiled many of these methods and libraries in a modular, multi-purpose toolkit that promises to make them more accessible to the chemical and materials community. This software suite, *ChemML*, will be open source under BSD 3-clause license and an initial release is planned for the end of 2016.

2.2 Many-body interactions: Methods, algorithms, and applications

G. Martyna, J. Hoja, M. Stöhr, H. Sauceda, A. Tkatchenko

2.2.1 Background

A reliable description of interactions in molecules and materials must be based on accurate quantum-mechanical many-body calculations for solving the Schrödinger equation. Such many-body calculations constitute the foundation of multiscale modeling and thus can be used for obtaining accurate molecular potential-energy landscapes, reliably modeling nanoscale materials, and creating benchmarks for materials properties, which can be deposited in big data repositories. The grand challenge stems from the fact that explicitly correlated approaches to solving the Schrödinger equation are extremely expensive (they grow exponentially with the number of particles), hence robust approximation methods must be developed, along with efficient algorithms to implement these methods to exploit massive parallelism on high performance computers.

Several useful databases of quantum-mechanical calculations of molecules and materials already exist. However, most of them contain ordered equilibrium structures and employ density-functional theory (DFT) as a reference, for example the Perdew-Burke-Ernzerhof (PBE) functional. While these calculations are very useful, they are not sufficiently accurate for most purposes discussed above. The understanding of properties and functions of materials requires long time-scale molecular dynamics calculations (millions/billions of time steps) for non-equilibrium structures with thousands of atoms. In this situation, quantum mechanical many-body calculations beyond standard DFT approximations become essential.

2.2.2 Recent Developments

The IPAM program covered a wide variety of methods and algorithms for modeling many-body interactions. The methods included modeling non-local electronic correlations (including van der Waals interactions) in molecular systems with Drude and quantum harmonic oscillators (MBD method), local chemical effects with embedded atom potentials (COMB method), and constructing interatomic many-body potentials based on multipoles and polarizable atomic fragments. Workshop 3 also touched upon highly correlated methods for molecules and materials, such as coupled cluster, density matrix renormalization group (DMRG), GW, and Bethe-Salpeter equation. In addition, several

participants discussed the general problem of describing quantum states (determinants, tensor networks, density matrices, etc.).

It is critical to develop new algorithms that avoid the exponentially growing complexity of the many-body problem for increasing system size. Novel developments in quantum chemical methods, including coupled cluster and DMRG [R. Schneider] allow quantum mechanical all-electron analysis [M. Reiher] and description of few-atom systems, with excellent accuracy. Large molecules and periodic crystals, on the other hand, are still mostly described adopting approximate and increasingly efficient [K. Burke] Density-Functional Theory (DFT) approaches. While providing a successful quantum mechanical description of chemical bonds, and an appropriate basis for perturbation theory techniques [L. Lin], DFT functionals employ semi-local approximations to the electron correlation. One of important shortcomings of semi-local DFT is the lack of ubiquitous van der Waals (vdW) interactions. Coarse grained methodologies, based on the accurate parametrization of coupled atomic polarizabilities [T. Bereau] have recently emerged as a very promising tool, capable of capturing the complex collective nature of long-range correlation interactions, and will likely benefit from novel model reduction techniques [Y. Maday]. The Drude Hamiltonian [G. Martyna] and the related many-body dispersion (MBD) model [A.Tkatchenko] were shown to accurately capture the complex collective electronic fluctuations that are responsible for dispersion and induction forces in large molecular systems. Coherent wave-like electron density fluctuations were also recently predicted in nanostructured systems [A. Ambrosetti], causing enhancements in both range and magnitude of vdW forces. These findings highlight the necessity for a correct quantum mechanical many-body description of vdW interactions up to the largest nanoscale systems, nowadays out of reach for semi-local DFT. Methods that treat vdW interactions in an explicit many-body fashion open the way towards reliable simulations of large scale bio-molecules, well beyond ad hoc empirical force fields. Even larger scale molecular dynamics simulations, ideally reaching the 100,000 atom scale will likely benefit in the near future from a quantum-mechanical MBD description of vdW forces, thanks to a Car-Parrinello approach [A. Ambrosetti] developed at IPAM, and based on a computationally-efficient evolution of the system response function that avoids explicit diagonalization operations. Latest machine learning developments of accurate force fields and molecular dynamics [G. Csanyi, A. Shapeev, O. Isayev, S. Chmiela, K. Schuett] are extremely promising, and will most likely benefit from efficient analysis and predictions of collective electron charge dynamics, currently under development.

2.2.3 Future Directions

There are many challenges that the IPAM program has identified in the development and application of many-body methods:

1. how to balance accuracy and efficiency of existing DFT+vdW methods,
2. how to develop accurate interatomic potentials that treat electrostatics, polarization, and dispersion to all orders,

3. how to extend the applicability of explicitly correlated methods to increasingly larger systems.

2.3 Collective variables and nonlinear dimension reduction

T. Baker, R. Banisch, M. Haghighatlari, M. Hirn, O. Isayev, R. Kondor, L, Li, M. Meila, M. Rupp

2.3.1. Background

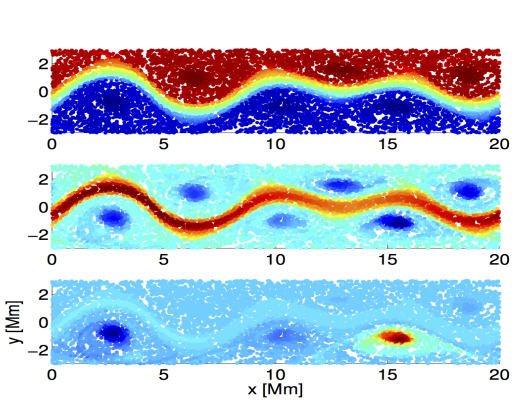
Unsupervised learning algorithms, rather than using training data to learn a model for prediction, analyze the intrinsic structure of (unlabeled) data. A recurring theme in this program was the appearance of new algorithms capable of learning, unsupervised from data, the collective variables of a physical system via nonlinear dimension reduction tools. While these methods take many forms, novel extensions of manifold learning and compressed sensing played a prominent role.

In Molecular Dynamics (MD), one typically deals with orders of magnitude difference in multiscale dynamical structure from very fast vibrational modes on the order of femtoseconds, to slow modes representing conformational changes which happen on the order of μs or ms. Often it is those slow modes, which are difficult to access through numerical simulation, one is most interested in. Additionally, the dynamics happens in the extremely high-dimensional space of molecular conformations ($\mathbb{R}^{3N}$, where N can be 1000 or more). Finding a parameterization of the (typically low-dimensional) subspace spanned by the slow modes is a major challenge, and the first step for many methods for coarse-graining, computing reaction rates, and more. Usually, one thinks of this parameterization in terms of a family of coordinate functions $\xi_i : \mathbb{R}^{3N} \to \mathbb{R}$, which are called collective variables.

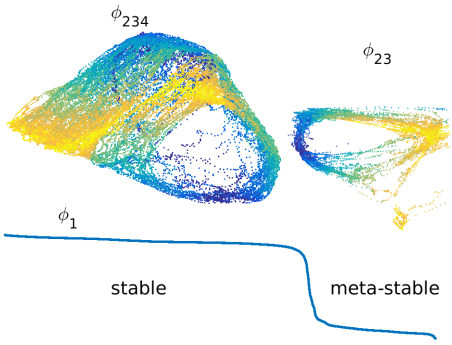
2.3.2 Recent Developments

Unsupervised learning can play a crucial role in solving this challenge. Given samples from a MD simulation trajectory, if we can design a metric which incorporates the slow/fast dynamical features, then the task of finding collective variables can be performed with established manifold learning methods, e.g. diffusion maps. There are several possible choices for designing such a metric. One choice, which was explored in (F. Noé, Workshop II), mimics the diffusion distance construction, but with the diffusion heat kernel replaced by the propagator of the dynamics. This was recently followed up by (C. Clementi, Workshop II), the commute distance constructed therein eliminates the τ parameter and is interpretable in terms of the commute time, i.e. the average time it takes to go back and forth between two states. Embeddings produced by the commute distance can be used to construct kinetic models which are accurate on long time scales. Constructing a distance which captures dynamical coherence is also possible (R. Banisch, core participant); see also Figure 2.

M. Meila (core participant) introduced methods that endow manifold learning algorithms with the ability to estimate and preserve intrinsic geometric information in the



(a) Diffusion maps analysis of a jet stream



(b) Low dimension embedding of aspirin

Figure 2: (a) The first three dominant eigenvectors of a diffusion maps analysis show the three slowest modes of a jet stream model. The jet stream is the red band singled out by the 2nd eigenvector (middle), which is flanked by six rotating vortices, two of which are singled out by the 3rd eigenvector (bottom). The 1st eigenvector divides the state space into a region north and south of the jet stream (top). (b) Non-linear embedding of 52,000 states of the aspirin molecule from a simulation at $T = 500K$. The first eigenvector, ϕ_1 reveals that the states form two distinct clusters. In the meta-stable state, the next two eigenvectors ϕ_2, ϕ_3 describe a circle which traces the torsion of the O=C–C–H bond. In the stable state, the next three eigenvectors ϕ_2, ϕ_3, ϕ_4 describe a two dimensional surface in which the same torsion is represented smoothly.

original data. Figure2 illustrates these methods on multiple states of aspirin. These algorithms are implemented in the python package `megaman` , capable to perform non-linear dimension reduction on millions of points in hundreds of dimensions. Johannes Garcke (Workshop I) contributed bindings for `megaman` to the fast eigensolver SAMG, speeding up embedding by a factor of 5 (from 30 min to 5 min for 1.5M data points).

Complementing the manifold learning approach, Nathan Kutz (Workshop II) presented work that combines compressive sensing techniques and machine learning with nonlinear dynamical systems to discover governing equations from noisy measurement data.

In development of new methods, Baker, Li, Burke & co-workers used a one-dimensional (1d) test laboratory developed in an article last year, allowing to quickly search through a myriad of possibilities faster than in three dimensions while developing tools immediately applicable to three-dimensional (3d) systems, to generate exact results from the highly accurate DMRG. A local density approximation (LDA) was also found in this model. Together with Kieron Burke (core participant) and Steven R. White we showed that ML can learn the exact functional from exact data, provided by DMRG. The resulting model allowed for self-consistent calculation of the functional. The remarkable aspect is that not too many training points were required to generate a model to chemical accuracy. Extensions to three-dimensions were investigated by one of us, Felix Blockherde

(workshop III), Klaus-Robert Müller (workshop III), and Kieron Burke. The goal is to transform an approximate, fast calculation like LDA or Hartree-Fock into a suitable basis set. This basis set would ideally be used to represent a few-site problem for an exact method, such as DMRG or quantum Monte Carlo, and obtain the exact ground state. Hence, the approximate answer was turned into the correct one without solving the expensive many-body problem.

2.3.3 Future Directions

A major goal in this area is to replace the expensive *ab initio* MD simulator, which is accurate on all time scales and operates at atomic resolution, with a cheap, coarse grained simulator, which is only accurate on long time scales (M. Maggioni, Workshop II). With existing tools, this goal can be achieved at least partially offline, i.e. after simulation data has been collected. But to be truly useful, methods are needed that can construct and use the cheap simulator online, to speed up exploration of the simulation. A collective variable should be tractable to compute, should lead to understanding not only calculation, and should ultimately have a clear physical meaning and thus be measured experimentally. The above methods give evidence that such goals are attainable in the future.

2.4 Potential Energy Landscapes

C. Anderson, S. Chiama, O. Isayev, Y. Li, A. Shapeev, S.Shankar

2.4.1 Background

Interatomic potential energy surfaces (PES) and force fields are used for atomic simulations in chemical, materials, and biological systems. These potentials and force fields are incorporated into molecular dynamics and Monte Carlo simulations with the objective of accurate predictions (better than 1 kcal/mol) without the need to compute detailed quantum mechanical interactions using *ab initio* calculations. The challenge is to use machine learning techniques to identify, construct and validate interatomic potential energy surfaces that can then be used to efficiently explore and model systems in the many practical applications that are not accessible using *ab initio* methods.

2.4.2 Recent Developments

Given the exponential scaling of accurate quantum methods, interatomic potentials are practically viable for molecular dynamics of complex macro molecules like proteins and DNA for applications in protein folding and ligand-binding (Isaev), material simulations for complex material structures including multi-grains and interfaces (Sinnott, Hart, Csanyi), chemical reactions by estimating transition states (Henkelman). The interatomic potentials and/or force fields could be physically-based or ML-based. The

differences between the two could be based on a variety of factors: the form of the functions (based on physics or empirical fitting of data), the basis of descriptors (based on physics or empirical fitting) , the number of parameters (small vs. large), etc..

2.4.3 Methods

Machine learned potentials are only as good as the quality of the underlying physics captured. Improvement in the quality thus requires the inclusion of better quantum mechanical simulations that are used to train the machine learing models. Several talks were given that focused on the improvement of the underlying quantum mechanical simulations, either improvements in the computational efficiency leading to more rapid generation of data (Burke et. al) and/or improvements in the modeling of the underlying physics such as the accurate modeling of long range quantum mechanical effects (Tkatchenko).

The machine learning potentials are not expected to give reliable answers beyond their training region, but if they can detect that they are extrapolating and possibly be re-trained automatically (i.e., they can do active learning), then they can be used to make predictions for the configurations they were not trained for. Two talks (A. Shapeev, A. De Vita) highlighted importance of active learning and proposed methods of doing it.

2.4.3.1 Methods: Descriptors

Work by researchers prior to the workshop has resulted in the consensus opinion that a key ingredient to the creation of a machine learning potential is the identification of the descriptors or machine coordinates that are used. Talks were given concerning the machine-learned models where descriptors based upon the physical characterization of the system (Ramprasad, Phillipott) and talks that combined mathematical and physical descriptions (Shapeev, Csanyi).

2.4.3.2 Methods: Kernel and Network-based

Several talks addressed the development of machine learning methods (ML) for speeding up molecular dynamics (MD) simulations. Novel neural networks architectures were introduced (Schuett, Isayev), as well as new descriptors and kernels for linear predictors such as Kernel Regression Methods, Gaussian Processes (Csanyi, Shapeev, DeVita, Chmiela). The importance of sampling was a frequent topic of discussion and new active learning schemes were presented (Noe, Shapeev, DeVita, Kevrekedis). Furthermore, new approaches for ML model analysis were proposed including Tensor Networks (Mller).

Potential energy surfaces (PES) trained with generic machine learning methods are not guaranteed to have correct gradients due to topological inconsistencies. The topology of such models is not sufficiently constrained by a low number of energy training samples. Being true quantum-mechanical observables within the BO approximation, atomic forces are cheaply available to train dedicated ML force fields. Force fields that satisfy the energy conservation constraint can even be integrated to obtain the PES (Csanyi, De Vita, Chmiela)

2.4.3.3 Methods: Chemical Reactions Potentials

Two widely used reactive potentials were discussed during the workshop. Both these methods use with variable-charge schemes that are desirable for multicomponent or multifunctional systems: the charge-optimized many-body (COMB) and the reactive force field (ReaxFF). Both reactive potentials can describe chemical bond formations and breakages when performing atomic simulations, which COMB (Susan Sinnott and Simon Phillpot) has more emphasise on the macroscopic behaviors, such as the bulk modulus, etc, whereas ReaxFF (Amy Ying Li) has focuses on detailed quantum chemically verified bond formations and breakages. Both potentials can simulate heterogeneous systems in a large scale.

2.4.4 Future Directions

There are multiple challenges in ensuring the computational efficiency enabled by PES and their applications: Accuracy, where the potentials need to be valid for a wide variety of systems (Philippott), chemical and reaction systems with 1 kcal/mole (Henkelman), roughness of landscapes (Csanyi), and sampling methods (Kevrekedis, Clementi, Noe). In addition, validation is a key challenge for ensuring that the experimental systems and the modeling systems are consistent.

2.5 Benchmarks and Data Repositories

Farnaz Heidar-Zadeh, Olexandr Isayev, Katerine Saleme Ruiz, Mohammad Atif Faiz Afzal, Matthias Rupp, and Claudia Draxl

2.5.1 Background

Reliable data are the basis for exploring ML models. This concerns both, the choice of the ab initio method to represent the target property best as well as the consistency of data. Focusing on computed data, accuracy is an important aspect; it requires validation, i.e. confronting computed data with experiment or with those obtained by higher-level theory. The other aspect is precision, i.e., making sure that the respective equations were solved accurately, and applied approximations dont affect the results beyond a tolerable threshold (verification).

While in quantum chemistry benchmarking and dedicated datasets have been common for more than two decades, in materials science efforts along these lines have just started, the most prominent example being the so-called Delta test.

While the latter investigation was based on and, actually, required calculations just carried out for this purpose, the design of new machine-learning methods asks for data collections that are ready to be used. It would be desirable to test different approaches on the basis of such dedicated datasets to evaluate their performance. Lack of such data for some particular cases was expressed during a panel discussion in WS III.

Conducting large numbers of high-quality calculations is a substantial investment of both human and computer time. Sharing and reuse of such data is therefore highly

beneficial, and several large-scale data repositories for QM calculations for materials and molecular data are being developed. Each of them serves a different purpose, and a few of them are briefely described below, selected in view of ML aspects and their relevance to the IPAM workshop.

2.5.2 Recent Developments

IPAM hosted a considerable number of key players of data repositories and benchmarking efforts. Among the invited speakers were Chris Wolverton (WS I) who runs the OQMD database), a collection of nearly half a million inorganic compounds.

Matthias Scheffler (WS I, core participant) and Claudia Draxl (WS I, core participant) presented the NoMaD Repository and data analytics tools developed on the basis of these data within the NOMAD Center of Excellence. The NoMaD Repository accepts input and output files of all major electronic-structure and force-field codes. It enables sharing of data, reuse and repurposing, giving access to the raw data. It currently hosts more than 3.3 Mio calculations, and is the only repository for materials recommended by Nature Scientific Data. This large data collection is the basis for the activities of the NOMAD CoE. A major effort is placed on the normalization of data (NOMAD Archive), i.e. dealing with the necessity to make data produced by different codes and applied approximations comparable. These data are then used for data mining and the development of a Materials Encyclopedia. At all steps, i.e. the Repository, the Archive, the Encyclopedia, and the Data Analytics platform, APIs are or will be made available to provide the respective data.

AFLOW is a repository of computational materials databases constructed from high-throughput ab initio calculations using the AFLOW framework. It is developed and maintained by an international collaboration between 15 academic groups. The AFLOW (Automatic FLOW) code works with the VASP and Quantum ESPRESSO DFT packages. It includes preprocessing functions for generating input files for the DFT package; obtaining the initial geometric structures by extracting the relevant data from crystallographic information files or by generating them using built-in prototype databases, and then transforming them into standard forms which are easiest to calculate. It then runs and monitors the DFT calculations automatically, detecting and responding to calculation failures, whether they are due to insufficient hardware resources or to runtime errors of the DFT calculation itself. Finally, AFLOW contains postprocessing routines to extract specific properties from the results of one or more of the DFT calculations, such as the band structure or thermal properties. The repository is continuously updated and currently contains over 1.47 Mio. entries for inorganic materials and 227 Mio. calculated properties.

During this IPAM program, O. Isayevs (WS III, core participant) group developed a materials-informatics web-based application and RESTful API for machine learning models for electronic and thermo-mechanical properties. The module is fully integrated into AFLOW and may be used directly to screen for materials with a desired property.

The Harvard Clean Energy Project Database (CEPDB), run by Aln Aspuru-Guzik (organizer, WS III), compiles 2.3 million candidate organic electronic materials to design next generation of plastic solar cell materials.

For benchmarking purposes, the qmml.org website was set up by Matthias Rupp (WS I, core participant) to provide easy access to suitable datasets for the community. It contains established datasets such as those derived in last years from the Generated Database (GDB), containing all small organic molecules satisfying simple chemical stability and synthetic feasibility rules, as well as new datasets contributed by Gbor Csnyi (organizer, WS III; water monomers/dimers, solids) and Gus Hart (WS III; solids). These benchmarking datasets serve to develop, train, validate and test machine learning algorithms for accurate interpolation of electronic structure calculations.

The website quantum-machine.org - maintained by the group of Klaus-Robert Mller (core participant) - hosts molecular reference data for medium-sized organic molecules, covering both compositional and configurational degrees of freedom. These are used by several published and unpublished works as benchmarks for potential energy surfaces and force field ML models. Several datasets include force labels in addition to energy labels for each (non-equilibrium) geometry. These benchmarking datasets serve to develop, train, validate and test machine learning algorithms for accurate interpolation of electronic structure calculations.

We remark that several databases were created as a specific outcome of the IPAM program.

2.5.3 Future Directions

Data collections are an emerging issue in materials and chemical science. These collections are an important basis for developing machine-learning and data-mining concepts and tools. IPAM has brought together key players from both areas, disussing needs and synergies, and establishing interdisciplinary connections.

Particular emphasis should be placed on data quality. This concerns the trust level of calculations, e.g. by assigning error bars with respect to computational parameters, approximations, and methodology. Large enough curated datasets should be created. These could be extracted from existing collections and / or come from dedicated high-throughput calculations. More userfriendly tutorials are needed to train young scientists and get them involved into the exploration of how to extract knowledge from data.

2.6 Nano-Engineering

F. Stefan Tautz, with inputs from Philipp Leinen, Kristof T. Schütt, Christian Wagner

2.6.1 Background

The idea to freely control the structure of matter down to the atomic scale has intrigued scientists for many decades. Scanning probe microscopy (SPM) is the method of choice for arranging atoms and molecules with precise control. We anticipate that SPM can be developed into a molecular assembly device, capable of 3D printing on the single-molecule level, to create a wide variety of functional supramolecular nanostructures. However, the wide variability of molecules with very different sizes, shapes and functionalities that provides the exceptional design freedom also implies a challenge, namely

the impossibility to perform the manipulation experiments in the essentially stochastic way that up to now forms the basis of most SPM manipulation experiments.

2.6.2 Recent applications with some results

It has been demonstrated that the SPM tip can be used as nano-robotic arm with sufficient spatial resolution to handle complex molecules in a continuous and deterministic way, such that a much larger number of states than accessible by spontaneous assembly can be reached. However, when the tip is engaged in manipulation it is not usable for imaging, which means that knowledge about the state of the molecule needs to be reconstructed from sparse measurements at all stages of the manipulation process. Here machine learning (ML) offers unique opportunities, because it is especially suitable for uncovering non-trivial patterns and correlations in data.

We have identified two complementary approaches to apply ML in SPM-based nano-engineering:

1. *Data-analytics approach:* Large datasets are generated by repeated manipulation experiments, the resultant data is structured by a sequence of unsupervised and supervised ML steps into discrete situations, and finally these situations are identified with the help of atomistic simulations. As a result, a comprehensive map of situation space becomes available, on the basis of which experiments can be planned and executed under control of the experimenter.

2. *Autonomous control approach:* Avoiding explicit data analysis, one can use deep reinforcement learning to steer the manipulation process toward a specific target. Given a suitable reward function, this has previously been used to acquire human-level skill in video games or handle complex control tasks, e.g. in robotics. We anticipate that it can also be successfully applied in nano-engineering, turning the SPM into an autonomous robot for assembling supramolecular structures from single molecules.

Over the course of the IPAM program we have developed a detailed road map for both approaches, demonstrated the feasibility of all individual steps in the roadmap and started their implementation. For example, we trained a reinforcement learning model to peel a molecule off a surface in a simulated environment. The model has to find a tip trajectory below a predefined force threshold, while having access only to its previous actions and observations. This proof of concept can be regarded as a first step towards solving more general control problems in automated nano-structure assembly. Furthermore, we initiated a project in which state-of-the-art supervised machine learning of intramolecular potentials, trained on DFT data, is used for developing a high precision molecular mechanics model for simulating manipulations. Such a potential is required for the identification of situations in the data-analytics approach.

2.6.3 Conclusions

SPM-based nano-engineering benefits from accurate machine-learned interatomic potentials at various levels, up to real-time simulations that can be used to visualize the com-

plete manipulation process while it is going on. It also relies on state-of-the-art methods across the board of ML (i.e., supervised/unsupervised learning, kernel methods, support vector machines, deep neural networks etc.).

On a more general note, the problem of molecular manipulation illustrates a usage of experimental data in ML that has not yet been exploited widely. So far, in materials science experimental data in the context of ML has mainly been used as benchmarks, and to a limited extent for providing training data across compound space. In the present context, we use data from a particular experiment as training data for ML in order to uncover vital information that is only accessible in statistical patterns and correlations. The model thus learned can be used for guiding the experiment.

2.6.4 Future Directions

In SPM-based nano-engineering one may anticipate adaptive control, being able to modify previously learned situation spaces in response to changes of uncontrolled experimental parameters or minor modifications of the experimental setup (type of surface or molecule), by combining the data-analytic and autonomous approaches mentioned above.

Machine learning is relatively new in the natural sciences, as opposed to artificial intelligence in an everyday environment (speech recognition, image processing, translation, autonomous driving etc.), where no specialized domain knowledge beyond ML is required. Currently there is rapid progress of ML in the natural sciences. This is significantly catalyzed by the present IPAM workshop, were researchers from the fields of ML on the one hand and physics, chemistry and materials science who provide the relevant domain knowledge on the other hand, interact strongly. So far, however, this is mainly restricted to the theoretical branches of theses natural sciences. We anticipate that ML has the potential to benefit the corresponding areas of experimental science as well. As demonstrated above for the example of nano-engineering, in the experimental context ML offers, besides data analytics, the additional advantage of providing an interface to process control, allowing the execution of experiments which would be too difficult to do in a purely human control loop.

3. Applications

Although this is an emerging field, several applications of techniques were presented during the IPAM program. These included application of ML methods to metallic glasses, organic photovoltaics, catalysts, electronics materials, biological applications and bioinformatics for cancer identification.

Each of the applications used ML and an appropriate physical or chemical descriptors for predicting material behavior. One of the applications showed that there is structure hidden in the disorder of glassy materials which can be quantified by softness. A simple Arrhenius relaxation for each softness, coupled with the time evolution of softness, led to the observed relaxation dynamics of glassy liquids below an onset temperature (Kaxiras & co-workers). Another application used combination of quantum methods

such as time-dependent density-functional theory and ML methods such as neural networks and Bayesian methods to screen through several millions of molecules to identify specific molecules for organic light-emitting diode molecules across the visible spectrum (Aspru-Guzik & co-workers). Another application used linear energy relations as a dimensionality reduction tool in catalyst searches which have helped in finding leads for novel heterogeneous catalysts and electro-catalysts (Bligaard & co-workers). An industrial application showed use of ML methods for estimating adhesion between different interfaces of materials, yield stresses, and in synthesis of materials (Shankar & co-workers). Michel & co-workers presented a business model which uses ML methods to build a unified system for storing broad sets of materials and chemical data, generating machine learning models on top of that data, and the effectiveness of adding physical models to that predictive system.

For biological molecules, Beratan & co-workers used a simple few-parameter coarse-grained Hamiltonian models, inverse design via the linear combination of atomic potentials method, diversity-oriented molecular library design, and property-biased molecular library design to use conceptual framework for identifying molecules with specific chromopore design. In addition Stahlberg presented a recent collaborative effort spanning multiple DOE national laboratories and National Cancer Institute are exploring how predictive insight into complexities of cancer can take advantage of emerging machine learning capabilities to accelerate insight and understanding.

4. Conclusions

The long program was intended to look at the possibilities of applying machine learning to model and predict properties and behavior of many-particle systems. Although this is an emerging field, several substantial breakthroughs in the discovery of molecules and materials were presented during the program. The field is extremely competitive with new players coming every month and we expect many more unexpected discoveries in this field in the near future. It was also clear from studying both classical and quantum systems that there are several challenges including availability of sufficient high quality data and improved calculation methods, descriptors for collective variables that tie atomic descriptions to properties other than energy, and appropriate machine learning algorithms. At this point, these machine-learning algorithms are often developed using different programming architectures, and can require a higher level of familiarity and expertise for most of the applications.

In addition, physically-based models in scientific and engineering applications are extrapolative, i.e. the parameters are determined by observing limited data in some domain, and the models are tested in extended or even wholly different domains, and the performance of such models is evaluated according to how well they do in such a situation. In contrast, high dimensional ML models are best at interpolation. Several applications were demonstrated that demonstrated the integration of both extrapolative (physics-based) and interpolative (ML-based) techniques. We expect many further developments in this area.

In the program, there were several discussions touching the areas of machine learning

models and representations, many-body interactions: methods, algorithms, and applications; dimensionality reduction and collective variables; potential energy landscapes, nano-engineering with specific experimental manipulation of atoms, and benchmarks and data repositories. The emerging areas are likely to be exciting in terms of research and applications for materials. The cautionary message is that the methods need to be carefully assessed in applications to estimating experimentally verifiable properties of real materials.