White Paper: “Machine Learning for Physics and the Physics of Learning” (IPAM Long Program, Fall 2019)

Tristan Bereau, Kathy Breen, Xavier Brumwell, Bing Brunton, Steven Brunton, Jared Callaham, Onur Çaylak, Maghesree Chakraborty, Kathleen Champion, Nicholas Charron, Yu-Chia Chen, Cecilia Clementi, Gianni De Fabritiis, Alex Goeßmann, Rhys Goodall, Francois Gygi, Richard G. Hennig, Moritz Hoffmann, Brooke Husic, Lukas Kades, Tom Kennedy, Stefan Klus, Jonas Köhler, Dominik Lemm, Andreas Mardt, Marina Meila, Kim A. Nicoli, Frank Noé, Luca Pasquali, Patrick Riley, Adam Rupe, Matthias Rupp, Lars Ruthotto, Anastasiya Salova, Kirill Shmilovich, Jordan Snyder, Matthew Spellings, Marc Stieffenhofer, Luca Venturi, Jiang Wang, Andrew D. White, Stephen R. Xie

Table of Contents:

1. Executive Summary

This whitepaper summarizes the progress and insights from the IPAM long program on “Machine Learning for Physics and the Physics of Learning”, held in Fall of 2019.

During the last couple of decades advances in artificial intelligence and machine learning (ML) have revolutionized many application areas such as image recognition and language translation. The key of this success has been the design of algorithms that can extract complex patterns and highly non-trivial relationships from large amounts of data and abstract this information in the evaluation of new data. In the last few years these tools and ideas have also been applied to, and in some cases revolutionized problems in fundamental sciences, where the discovery of patterns and hidden relationships can lead to the formulation of new general principles.

This IPAM program focused on the opportunities and challenges in the application of ML tools in the physical sciences and if/how theoretical results in the physical sciences can help in the definition of new ML methods. The program hosted four workshops (WS) focusing on different aspects of the overarching topic:

WS1 “From Passive to Active: Generative and Reinforcement Learning with Physics” focused on novel machine learning models for designing new molecules or materials, synthesis pathways, and optimal controls for dynamical systems.
WS2 “Interpretable Learning in Physical Sciences” focused on the need to develop interpretable ML methods to understand the ML predictions in terms of physically meaningful quantities, in order to advance our understanding.
WS3 “Validation and Guarantees in Learning Physical Models: from Patterns to Governing Equations to Laws of Nature” focused on learning equations, i.e. interpretable and extrapolatory models from data, on modeling dynamical systems and low-dimensional manifolds, on bounding errors and statistical aspects model selection.
WS4 “Using Physical Insights for Machine Learning” focused on applying insights or models from physics to make progress in developing new ML models and algorithms, or to better understand why successful ML models such as stochastic gradient descent in deep learning frameworks work well.

In addition to the workshops, we formed multiple working groups that met regularly during the program and tackled different subtopics. Subsequently, the state of the discussion and outcomes on these different subtopics are described. In particular, we discuss the open challenges that have been identified and that we as a group plan to continue to investigate in the future.

2. Physics-Constrained Machine Learning

Introduction: Building known physical laws and constraints into ML models of physics can significantly improve the predictive accuracy and statistical efficiency of the model by removing physically implausible predictions from the search space. Two principal approaches to enforcing physics constraints are: (i) data augmentation (DA), and (ii) hard-coding the constraints in the ML model. DA is conceptually easy to do, but is statistically inefficient and does not preclude violation of the constraints. Building constraints into the ML structure ensures that the constraints hold exactly and reduces necessary training data, but requires designing specific structure into ML models or representations.

Symmetries, invariances, and group actions Physical systems and processes often have symmetries. For example, the energy of an isolated molecule will be invariant to rotation and translation, i.e. to operations in the symmetry groups of 3D translations and rotations. In general symmetries can be linked to a conserved property. Enforcing symmetries in machine-learned models of physical systems exactly is often essential, e.g., in order to ensure stability of time-integrating with a machine-learned force field. Invariances can be enforced in ML models either by inserting invariant features into the learning structure (e.g., distances are invariant to translation and rotation) or by using operations inside the ML structure that induce the desired invariance (e.g. sum pooling induces permutation invariance).

Equivariances and Convolutions Equivariance is a generalization of invariance. An equivariant function will commute with a symmetry group. Equivariant functions can retain properties in data that are discarded through invariance. For example, a function invariant to rotation will not be able to learn forces of a molecule, but one that is equivariant can retain the necessary orientation information. There are multiple ways to include translation equivariance into a model, e.g., convolutional layers are translation equivariant since they apply a set of filters uniformly across a function space. Equivariance to rotation is less straightforward. Graphs and pairwise distances are rotation invariant, but separate the information of the system into two-body interactions which have to be recombined to get more complex interactions. Also, although these functions appear natural for molecules, they do not apply for other data structures. Spherical harmonics can be used as a tool to encode rotation equivariance while retaining the spatial information of the whole data structure. One method to derive features from data which are both rotation and translation equivariant is to use a convolutional network where each filter is the product of a spherical harmonic and a radial function.

Asymptotics Extrapolation to regimes outside the training data is difficult, especially for highly nonlinear ML models such as neural networks. To avoid predicting arbitrary, incorrect results, the extrapolation behavior of the ML structure must be considered. In physics we often have known asymptotics that can be explicitly built in. Examples: (1) When learning all-atom or coarse-grained molecular potentials, we can add a prior energy that ensures that the learned energy will go to infinity as particles approach each other. (2) For ground-state solutions of the Schrödinger equation the derivatives of the wavefunction is known to have certain limits as electrons approach nuclei or other electrons. In these limits, the probability of finding electrons also vanishes, i.e. hard-coding the correct asymptotics into trainable wavefunctions is important.

Physics-derived learning problems Given that we want to solve a particular physics problem with machine learning, what is the “correct” optimality principle, or loss function? In supervised learning problems this is straightforward: we want to minize the difference between predictions and given labels, e.g., using mean square error or cross-entropy. In unsupervised machine learning, it is often less clear what the best objective is, but physics often helps defining the learning problem. Examples include: (i) Electronic structure problem: Rayleigh-Ritz variational principle, (ii) Molecular kinetics: variational approach of Markov processes (VAMP), (iii) Coarse-grained molecular potential: thermodynamic consistency, can be implemented by force matching or minimization of Kullback-Leibler divergence, (iv) Unbiased one-shot generation from a target density (e.g. Boltzmann): Variational energy loss.

Work done at IPAM Investigating the effect of permutational symmetries on low dimensional embeddings and clustering of MD data; which invariances to enforce, depending on the system and data. (Kondor, Meila, Pan, Shlyakhtenko)

Outlook Efficiently encoding physical constraints into ML structures is an active area of research. Open problems include Gauge invariance in lattice theories and efficiently encoding permutational antisymmetry in fermionic wavefunctions.

3. Error bounds and Theory of Machine Learning

Introduction: For physics models learned from data, it is often important to guarantee that model quantities lies within a certain error interval, at least to guarantee that the model obeys certain structural constraints. One way to obtain such guarantees is to analyze the model after it has been learned from data. Results in this area are still few.

For trained multilayer neural networks, methods based on Rate Distortion theory solve the “credit assignment problem” of identifying the input components that have most influence on the current network prediction. For a learned clustering, one can obtain a prediction interval in the space of clusterings, by solving a convex optimization problem.

In many situations, models fit on a given large training set will be applied to data distributions that differ from that of the original data. In this case, for optimal performance the model must be retrained with data from the new domain. This paradigm is known as transfer learning, and one fundamental question is how much new data is needed. Intuitively, this would depend on “how far” the new data distribution is from the old, and recently it has been understood rigorously how this distance must be defined. Estimating the distance in practical settings is possible in special circumstances, and it is a matter of current research. With this estimate, one would have *a priori* guarantees that the accuracy of the retrained predictor would match its accuracy on the original training distribution.

Modeling the properties of over-parameterized neural networks: A pattern often observed in the current ML literature is the use of highly overparameterized models, in particular, Neural Networks (NN), which have become de facto "universal approximators". The current trend is often to construct a large, heuristically motivated, model—potentially including billions of parameters—and optimise it using stochastic 1st order optimisation (gradient descent) algorithms. While it is clear that more complex models are able to fit a larger variety of functions, it is somewhat surprising that simple optimisation methods are able to optimize these models while also avoiding overfitting. It is vitally important to understand the convergence of such methods on three different levels: (1) How fast are we reaching a global optimum (and why can we reach it)? (2) What are the characteristics of the optima we find? (3) What are the tradeoffs between number of parameters, number of data points, and data dimensionality?

Thermodynamic limit: Recent developments in the field of NNs theory study the “thermodynamic” limit as the number of machine learning model parameters growing to infinity. This has recently proven to be a successful idea that provides many new theoretical guarantees. In this framework, depending on the regularization (normalization) applied to the model, two different regimes arise: the mean field regime and linearized regime. These regimes allow us to study optimization dynamics and—in the linearized case—generalisation properties, e.g. the double descent phenomenon. Outstanding challenges include creating a unified framework for understanding the thermodynamic behavior of data-driven models, understanding the tradeoffs between models in the regimes listed above, connecting thermodynamic behavior to the finite-sized case, and showing that the double descent phenomenon exists for a larger family of models.

Navigating energy landscapes: Techniques such as the replica trick for certain classes of problems (spin glass, spiked tensor) precisely characterize the computational and statistical complexity and individuate phase transitions. Unfortunately, at the current level they are limited to a range of ‘toy’ models.

Outlook: In the analysis of NN, the challenges are (1) to analyze theoretical models that replicate the complexity of actual NN training, and (2) the validity of the very concept of data distribution; in some settings it is not clear that the training set captures the real conditions when the model is used. There exists few but very powerful frameworks for giving guarantees for a given trained model, without making untestable statistical assumptions. Here, the challenge is to refine these frameworks in order to give informative guarantees (e.g. narrow prediction intervals) for more cases.

4. Predictive Uncertainties and Domain of Applicability

Introduction For ML predictions to be useful, whether for applications such as dynamics simulations and active learning or for human assessment and decision making, an accurate estimate of their reliability (“predictive uncertainty”) is required. Quantitative predictive uncertainties also aid in quantifying the generalizability and reproducibility of ML models, as well as for model selection and hyperparameter tuning. While in the experimental sciences it is established best practice to report measurements with error bars, these are not often included with ML predictions. However, predicting uncertainties in addition to prediction errors is crucial for scientific ML.

Sources of uncertainty Predictive uncertainty can come from many sources, including data density and model limitations (for example, wrong functional form or missing features). For experimental data, noise is an additional strong source of uncertainty. Loss functions optimizing the L2 norm, which is commonplace in ML, correspond to assuming a homoskedastic Gaussian noise model. However, from experimental data and physical considerations it is known that such a noise model is often inappropriate, such situations call for more physically motivated choices of noise-models and loss-functions. Another distinction related to noise that is not often made is the one between confidence and predictive distributions.

Evaluation metrics It is currently unclear which evaluation metrics are best suited to validate predictive uncertainties. In ML it is common to use metrics such as log predictive density or continuous ranked probability score, but these often yield little additional signal beyond summary statistics of prediction errors (the latter, for example, is a generalization of mean absolute error).

Datasets with observed distributions Almost all currently available benchmark and validation datasets do not contain information about the uncertainty of recorded observations, which precludes direct comparison of observed and predicted distributions.

Domain of applicability An important aspect of predictive uncertainty is identifying the domain of applicability of a model (the input range where the model is valid, that is, has low predictive uncertainty) to prevent uncontrolled errors, which may accumulate or lead to unphysical behavior in simulations. For example, an ML potential can perform well on a test set, but still have catastrophically large errors in unphysical regions of input space, such as configurations with very close atoms. Although the domain of applicability concept has a long history in cheminformatics, reliable identification of an ML model's domain of applicability remains an open challenge.

Locality of errors It is good practice to consider, in addition to the summary statistics, the distribution of prediction errors. However, global error metrics ignore the spatial dependence of prediction errors, that is, their dependency on the input space. In recent work, del Rosario et al. show that an acquisition function for active learning that focuses on the Pareto front can have higher global error than other acquisition functions while identifying better candidates, and Sutton et al. identify input space regions where model error is consistently and significantly lower than the average error over the whole dataset. The connection between spatially resolved prediction errors and predictive uncertainties is currently unexplored.

Conformal prediction is an assumption-free confidence interval prediction framework. In the past, methods within this framework, such as full conformal prediction, required practitioners to train in a leave-one-out manner which makes interval estimation computationally infeasible. Split conformal prediction, a computationally feasible alternative, has the drawback of reducing the power of the estimated interval. Recent progresses in the field, e.g., cross conformal prediction or Jackknife+, are more powerful with less computational power while still providing reasonable guarantees.

Active learning is intimately tied to predictive uncertainties via acquisition functions, which control the exploration-exploitation trade-off and rely on accurate predictive uncertainties. Due to the challenge of small datasets (Section 8), ML models for energy surfaces depend on answers to questions such as how many and which data points should be generated to obtain a prescribed accuracy. Reliable predictive uncertainties would enable efficient on-the-fly sampling of new data points whenever a model enters part of configuration space with high uncertainty due to sparse training data.

Work done at IPAM. A working group reviewed validation metrics for predictive uncertainties from ML and other fields such as the atmospheric sciences, set up recommendations for their validation, and benchmarked uncertainties with frequently used algorithms such as Gaussian process regression and random forests. Extensions of existing algorithms were proposed.

Outlook. Increased awareness, analysis, benchmarking and improvement of predictive uncertainties of scientific ML models would lead to better understanding of model capability and applicability, increased acceptance of ML models in science and industry, as well as improved performance of related approaches such as active learning. Community efforts towards availability of benchmarking datasets with observed distributions (as opposed to observed values) would greatly aid in this.

5. Accelerating Discovery with on-the-fly Learning

Introduction: Progress in science is often constrained by practical considerations of experimental design and data acquisition. Especially where the parameter space is large, we are generally unable to run all the experiments needed to exhaustively explore and characterize many systems of interest in physical, engineering, and biological sciences.

Typical machine learning studies start with a fixed, ideally exhaustive data set to which models are fit in a *post hoc* manner. Since this approach is often not tractable as data is produced in a stream, we are increasingly interested in developing *interactive* methods that provide principled guidance for model construction and continual learning.

Active Learning: Across many fields we have the common problem of selecting the most informative experiments to run given a strict budget in either data size or model complexity. Where experiments are constrained in total duration, we seek methods that are *fast* and may incorporate streaming data to make predictions in real time. Where experiments are constrained in sample size, we seek *steerable* models that quantify uncertainty and guide principled design of experimental parameters.

Within an active learning framework, data-driven modeling tools can lead to a more integrated, continuous, and virtuous cycle between data generation and learning that allow for improved performance from less data.

Reinforcement Learning: Other fields such as control in dynamical systems or design of new molecules can be viewed as decision processes. Here the possible state-spaces much larger than the compact design-spaces that current actively learning techniques can handle. The framework of reinforcement learning offers tools that can enable us to tackle such problems. The power of reinforcement learning is that by implicitly factorising vast conditional probability spaces agents can be trained to make decisions about how to interact with their environment over extended series of events. The agent is driven by an expected future reward which is optimized to approximate the true reward as the agent is exposed to more training examples.

Work Done at IPAM: The strengths and limitations of on-the-fly approaches were presented and discussed throughout the program. This led to the exploration of how such techniques might be applied across many disparate and interesting physical systems:

Example system - Large scale *ab-initio* molecular dynamics: Atomistic simulations based on quantum theory provide a consistent description of both structural and electronic properties. These *ab-initio* simulations are extremely costly and currently consume large amounts of computing resources on large supercomputers. ML-based adaptive algorithms as well as on-the-fly definition of electronic response models may lead to substantial acceleration in the exploration of electronic excitations and optical properties, using approaches similar to those discussed in the section on Machine Learning of Energy Landscapes.

Example system - Asymptotic safety in quantum gravity: The functional renormalization group represents one possible way to find a consistent and predictive theory for quantum gravity. Finding such a theory entails locating saddle points of a set of ordinary differential equations. Apart from incorporating prior physical knowledge no reliable methods exist to detect saddle points in higher dimensions. Reinforcement learning is a promising tool to reduce the number of differential equation evaluations, helping to find more accurate asymptotically safe theories of a gravitational field.

Example system - Optimisation of molecular species: By deploying an active learning procedure over a learned representation of pi-conjugated peptides molecules candidates for testing can be selected in a guided manner that balances the trade-off between exploration of undersampled regions and exploitation in high confidence regions of chemical space. The aim of using active learning to navigate this large chemical space is to be able to identify promising chemistries with optimal optoelectronic properties in fewer design iterations than other screening approaches.

Outlook: On-the-fly and interactive learning offers a powerful suite of tools for scientists - particularly experimentalists. Whilst many of these ideas are well established the recent innovations and progress within Machine Learning have made application of these techniques to higher dimensional problems and larger data sets more feasible. Development of software and standardisation best practises is now needed to make these tools more accessible with the potential to enable acceleration across a wide range of domains.

6. Machine Learning of Equations

Introduction: Scientific inquiry often aims to produce new descriptions and understanding of regularities in nature. Historically, these regularities have been expressed in concise mathematical laws. However, modern, complex machine-learned models are often not expressible in such a compact way. Small mathematical expressions have significant advantages over black box models: they can be more easily analyzed, extrapolation and transferability can be more intuitively understood, validity and sensitivity to nuisance factors are more apparent.

The automated discovery of equations is known mostly as “systems identification” in dynamical systems and “symbolic regression” in machine learning. Current approaches aim to use the power of modern computational techniques to produce concise and interpretable mathematical laws. There is a fundamental tension in equation discovery between descriptive capability and model complexity, and the goal is often to find parsimonious equations that balance accuracy and efficiency. These equations can then be used as surrogate models, e.g. to control dynamical systems or find novel materials.

Review of techniques: Approaches so far have generally fallen into three categories based on how the search space of possible models is parameterized. One category is evolutionary algorithms, where promising equations are mixed together to produce new ones (Eureqa). Another set of approaches rely on sparse regression to select relevant terms in the equation from a large library of candidate terms (SINDy, MANDy, SISSO). A third class represents equations as parse trees (GVAE, SD-VAE, NG-MCTS).

Major challenges: In many systems, the most natural variables for concise equations are not obvious. Measurement noise and latent variables often further confound this choice. Promising directions include the use of time-delay coordinates to uncover latent variables, regularized or robust statistics to denoise and remove outliers, and simultaneously learning coordinate transforms and equations in a joint optimization.

The optimal parameterization of the search space is another challenge. There is typically a tradeoff in terms of the generality/size of the search space and the ease of search. The parameterization can be informed by the known physics and symmetries of the system, but this is currently done ad-hoc and problem-specific. High-dimensional data often suffers from the curse of dimensionality, although tensor-based methods (e.g. MANDy) provide a promising framework. Further, the effect of various parameterization techniques on the overall success on a problem is poorly understood.

Given the search space, significant challenges still remain in search/optimization over that space, including balancing accuracy and parsimony (especially in the presence of sensor noise), fitting of constants at arbitrary points in an expression, informing the search with known physics, and appropriate normalization of data.

Lastly, convergence guarantees and bounds on the sufficient quality and quantity of data for a successful recovery are required to develop a more fundamental understanding of model identification. There are theoretical results for guaranteed recovery in sparse optimization, but these assumptions are often not met in practice.

Work at IPAM: A powerful way to spur further research is the creation of Common Task Frameworks (CTFs): benchmark datasets abstracted from real-world problems with clear evaluation metrics. CTFs assist researchers and have been enormously successful in a variety of machine learning areas. A key focus of our work is to build new CTFs for equation discovery.

The set of previously developed empirical potentials for interatomic energy yields a rich set of scientifically useful equations. From the LAMMPS library, we identified 151 many-body interaction equations. The equations are stratified by the incremental complexity of these interactions, from simple distance-based potentials like Lennard-Jones to complicated equations with angular and Coulombic terms. Additional complexity can be added by increasing the variety of chemical species and limiting the amount of information provided from the simulation.

In addition, we are developing several open benchmark problems in dynamical systems theory that will incorporate issues that are relevant for modeling physical systems. We are also developing a software package for SINDy that will be publically available and well-documented on Github. The goal is to promote the use of these methods for equation discovery in new scientific application areas.

Outlook: Despite the limitations and outstanding challenges discussed, machine learning for equation discovery has already moved beyond toy problems and been applied to real science problems, including predicting adsorption energy and superconducting critical temperature, discovering reduced-order models for complex fluid flows, identifying structural models for building safety, distinguishing metals and insulators, and identifying stable perovskites. We believe that further methods development and further applications for novel discovery will proceed hand in hand.

7. Data-driven Approaches for Complex Dynamical systems

Introduction: Recovering information from and exerting control over dynamical systems is a problem that is present in several scientific fields, including but not limited to molecular dynamics, fluid dynamics, climate science, and social systems. The analysis of these systems is challenging due to high dimensionality, strong nonlinearity, difficulty in understanding the governing equations of the system, or even acquiring these equations. Despite these limitations, there are methods for gaining qualitative and quantitative information about their behavior.

A traditional approach involves designing low-dimensional theoretical models. In this case, the models are often interpretable but do not necessarily provide an accurate description of the system of interest. Development of such a model also requires system-specific expertise, and thus limits transferability. Recent years have seen a rapid increase in the availability of data for dynamical systems, thanks to advances in experimental and simulation techniques as well as computing power. As a consequence, data-driven methods have in turn become the prevalent tool of choice for analyzing dynamical systems.

Data-driven methods: Several fields have developed different modelling methods, most of them closely related to Markovian models, due to their effectiveness in describing the behavior of a complex dynamical system through simple analysis. Most of these approaches are based on the identification of a low-dimensional linear representation of the system that preserves dynamical and stationary information. This is done by approximating the Koopman operator, a linear, infinite-dimensional operator that describes the dynamics of observables of the system. These methods are known as Markov models for the stochastic systems community. The advantage over classical dimensionality reduction techniques such as (kernel) PCA and manifold learning approaches such as diffusion maps is that Koopman-based methods explicitly take temporal information into account.

The Koopman operator is the optimal linear operator that evolves state observables through time. The operator eigenvalues and eigenfunctions then describe mode frequencies and spatial modes respectively. The simplest algorithm to approximate a Koopman operator is the Time-Independent Component Analysis (TICA) algorithm, which was independently discovered in the fluid dynamics community as Dynamic Mode Decomposition (DMD). Many different extensions have been proposed, using reproducing kernel Hilbert space theory, tensors, or neural networks. These methods try to address the curse of dimensionality, i.e., the number of basis functions required for obtaining accurate estimates of eigenvalues, eigenfunctions, and modes grows exponentially with the system size. The hyperparameter search problem of choosing the right observables or the right kernel is often neglected. The recently variational approach for Markov processes (VAMP) defines a general ML loss function for linear approximations of dynamical systems that can be employed to find suitable hyperparameters.

Molecular Dynamics Applications: Data-driven methods based on the Koopman operator, e.g. TICA or VAMPnet, are used to isolate slow modes in the dynamics of large molecules. In particular, a spectral gap in the approximated operator signifies a separation of time scales and eigenfunctions of slow timescales give the transition structure between metastable states in the stationary dynamics. In general, such spectral gaps might not exist and thus the truncated linear model might not be accurate.

Fluid Dynamics Applications: PCA and DMD are commonly used for model reduction. Galerkin methods can reduce Navier-Stokes PDEs into a system of coupled ODEs, and similar methods are used to identify localized oscillatory flow features, such as large-scale atmospheric oscillations (e.g. El Nino Southern Oscillation). The Koopman operator and other methods, like geodesic Lagrangian Coherent Structures and local causal states, show promise in capturing more complicated localized features like coherent vortices and jet-core flow barriers in turbulent flows and hurricanes from climate data. Yet, because ground truth does not exist for these complicated features, method validation is an open challenge.

Work done at IPAM: We started integrating the VAMP framework into dynamical systems analyses from fluid dynamics. Efforts at IPAM have led to the development of community software tools in TICA/DMD/VAMP/system identification methods. This will help bridge gaps between different fields utilizing similar methods and organize approaches as dynamical systems analysis techniques advances in the future.

Outlook: Incorporating symmetry constraints into machine learning methods has been shown to make problems more tractable and physical where equivariances and invariances are present. Similar techniques can be explicitly applied to studying other dynamical systems of interest using the methods discussed above. For example, imposing symmetry constraints with respect to continuous groups can be important for problems in fluid dynamics, and the analysis of complex networks can benefit from considering permutational symmetries in their structure (see section 2).

In the end, all the aforementioned methods are developed and used by several different communities and a plethora of extensions and modifications have been proposed over the last years. Unifying these methods remains a challenging task.

8. Machine Learning of Energy Landscapes

Introduction: Accurate numerical simulation of molecules and materials is at the heart of physics, chemistry, and materials science. Still, established approaches are limited by some combination of accuracy, transferability, and computational cost. ML potentials of energy landscapes can enable the large-scale exploration of structure and dynamics by increasing the accuracy and reducing the computational cost of models.

Recent progress was made with respect to (i) symmetries, such as curl-free kernels and covariance in features and models to increase the efficiency of learning, (ii) ML models that reproduce first-principles energies for small molecules and simple materials at a comparable accuracy, and (iii) applications to, among others, reactions, structure searches, phonon calculations, and spectroscopy.

Outstanding issues. In this IPAM program, we identified the specific issues of (i) the limitations in computational speed of ML potentials, (ii) the lack of long-range interactions, (iii) the small size of dataset due to the cost of first-principles calculations, and (iv) the need to incorporate physical knowledge into the models to provide robust models the community can use and develop.

Faster potentials. Current ML models, while faster than first-principles calculations, are still orders of magnitude slower than efficient but fixed-form empirical potentials, aimed at long simulations of large systems. To enable applications and community adoption of ML models as surrogate models for the exploration of energy landscapes, there is a need for ML potentials that approach the speed of empirical potentials while retaining some of the accuracies of the underlying first-principles reference calculations.

Long-range interactions. Many current ML representations describe atoms in their local environment, which limits the models to encode only short-range interactions. Extensions to long-range interactions that account for Coulomb forces, polarizability, and dispersion are of great importance for applications to ionic compounds and magnetic materials.

Physical intuition and constraints. Incorporating physical knowledge and constraints directly into ML methods can reduce the data requirement and lead to more robust and efficient models. Also a direct connection between physics and ML can enable models that are more physically understandable. However, research on physics-inspired machine learning models is in its infancy and requires close collaborations between the communities. See “Physics-constrained Machine Learning” below.

Robust models for community use. The robustness of a machine learning model is a crucial factor for its establishment and use by the community. Knowing and defining the limitations of models helps researchers understand appropriate use cases and reduces the number of failed simulations.

Distributed community efforts. Currently, once a machine learning model is published, there is rarely the possibility to improve the model due to training and testing data not being public and the non-existence of a platform to easily share data. Developing a cyberinfrastructure to build and refine models by continuously incorporating new challenging structures would not only make the models more robust for the whole community but also reduce the number of redundant computations. These efforts benefit particularly from linking the communities of potential developers to applied math and computer scientists as done in this IPAM program to utilize prior domain knowledge.

Small data challenge. ML potentials models are only as good as the data used in their development. Accurate models rely on expensive quantum mechanical datasets. The resulting small datasets can limit conventional ML methods. Kernel methods are non-parametric and are more data-efficient than, e.g., neural networks. Still, in practice, kernel methods work only in the low data regime because the size of the kernel scales quadratically with the number of data points. On the other hand, parametric models such as neural networks are more efficient but require large data sets.

The community needs more labeled datasets because the current datasets are either limited in scope or not sufficiently accurate enough for some applications. The use of GPUs is also still at its infancy in quantum software, with limited speed-up apart from some exceptions, while very established in biomolecular dynamics simulations. GPUs could help reduce the larger prefactor on more scalable methods like quantum Monte Carlo and in so doing obtain very accurate datasets.

Work at IPAM. A working group at IPAM investigated many-body expansions for ML potentials and the possibility to learn such potentials using physically derived ML models by linear regression on many-body simulation data. This approach could enable ML potentials with a computational speed comparable to Lennard-Jones and Stillinger-Weber potentials and simplifies the incorporation of physical constraints.

Outlook. The incorporation of physical domain knowledge and constraints, as well as extension to long-range interactions can provide efficient and robust ML potentials that enable large-scale structure searches and simulations of the dynamics of molecules and materials ranging from biomolecular systems to structural and functional compounds. The proposed community efforts for database creation, validation, and sharing as well as the cyberinfrastructure to continuously build, refine, and validate models would empower the broader community to utilize these ML potentials in a vast range of applications.

9. Machine Learning for Coarse-Graining

Introduction. Many interesting physical phenomena, such as cellular division or the origin of stars, have an enormous number of constituent particles, making direct simulation computationally infeasible. One way to study these physical phenomena with simulation is by *coarse-graining*. Coarse-graining reduces simulation complexity by grouping together similar particles, with the goal of retaining the important physical behavior present in the simulation. The original simulation, with the complete set of particles, is called fine-grained (FG) and the coarsened system is called coarse-grained (CG).

Problem Statement. Coarse-graining is a two step process. First, one must choose how to group the fine-grained particles into CG large particles, called beads. This mapping is called “crisp” if particles are placed entirely into one coarse bead or “fuzzy” if particles are distributed among multiple beads. In either case, the mapping choice has more degrees of freedom than particle number, which is large. The second step in coarse-graining is choosing how to define the governing potential energy equation for the CG system. This search for an effective CG potential requires knowledge of the mapping, although these two steps could be iterative. The selection of mapping and effective CG potential are optimization problems. The choice of the optimization objective and treatment of high-dimensionality are important challenges, especially the often neglected optimization of mapping.

Machine Learning opportunities. ML enables new approaches to previously intractable challenges in the definition of CG models. For example, the effective CG potential that is thermodynamically consistent with the FG system must include multi-body terms among the CG beads due to the removal of degrees of freedom. This is true even if the FG potential includes only two-body interactions between particles. As the integral equations that define a thermodynamically consistent CG potential are not analytically tractable in practice, existing methods define an approximate CG energy by variationally optimizing the parameters of an ansatz functional energy form. It is difficult to design these functional forms as multi-body and thus most past work neglects this. This can now be solved with ML. NNs can naturally include multi-body terms and nonlinearities without the need of explicitly providing a functional form due to their universal function approximation power. Several participants in this IPAM program have developed CGnet, a specific NN architecture for the coarse-graining of biomolecules like protein systems.

Work at IPAM. The choice of CG map has been rarely studied, partially due to the large space of choices (e.g., linear vs non-linear, crisp vs fuzzy) and its high dimension. A core conclusion of our group is that it must be studied because the mapping sets fundamental limits on the quality of CG model. Current state-of-the-art is human intuition, although recent work presented at IPAM is pushing for systematic choices. In addition to the challenges of the mapping object (dimension, choice), there is no consensus on the objective to optimize when choosing the mapping. Ideas identified during the program include mapping entropy, force matched noise, force matched error and information content of CG beads.

During this IPAM program, we started to address open challenges for coarse-graining by focusing on a highly simplified model system - a quadratic potential - for which at least two possible objective quantities can be computed in closed form. We also have good intuition for choice of CG maps. We studied the role of the noise and coarse-grained entropy in evaluating the quality of a CG map. The noise measures the extent to which the force on CG coordinates is under-specified by the CG coordinates themselves, and the coarse-grained entropy is simply the entropy of the CG distribution. We find that in the simplified models we studied, the noise and the CG entropy need not be correlated. We also find that minimizing noise while maximizing CG entropy is achieved by performing PCA on an ensemble of system configurations.

Outlook. There are a number of open challenges relating to ML in coarse-graining. The first challenge is optimizing the CG mapping. In our model systems, we find no correlation between CG mapping entropy and noise, two possible mapping optimality measures. Are they truly independent measures of quality? Is there a best objective for optimizing CG mapping? The second challenge is maintaining the physics of the original system. The goal of coarse-graining is to keep relevant physics when coarsening, but can this be measured or quantified? The third challenge is developing transferable CG models. Developing effective CG potentials and mappings that can be used across multiple systems would drastically reduce the data and computation time required for developing new CG models. The last challenge is the incorporation of asymptotics into effective CG potentials. For example, the effective CG potential should go to infinity when approaching non-physical configurations, such as those with overlapping particles (See also “Physics-constrained Machine Learning”).

10. Generative Models for Physical Systems

Introduction. Generative models are capable of representing high dimensional, potentially conditional, distributions. These models can generate unseen samples according to a learned distribution reflecting the underlying data distribution. Prominent examples include generating realistic human faces, inpainting images, and generating text. In the context of physical systems, generative models have experienced success in producing molecular graphs, molecular structures, and lattice configurations.

Given a set of physical objects that share a number of common properties, an instance of property values may not uniquely define an object. For example, a molecule is defined not only by its chemical composition but also by its 3D structure. To generate a valid molecular structure from chemical composition, a generative model must learn a distribution over the structure of molecules that share that composition.

Challenges and approaches. While learning distributions is appealing, evaluating the similarity between the target and model distribution is challenging. Information theoretic approaches are the Kullback-Leibler divergence and/or the Jensen-Shannon divergence to evaluate the similarity of the distributions, while optimal transport theoretic methods give rise to the family of Wasserstein-p metrics. Existing generative models can be classified into three categories: exact likelihood methods, approximate likelihood methods, and likelihood-free methods. Prominent examples are Flows, Variational

Autoencoders (VAEs), and Generative Adversarial Networks (GANs), respectively. Flow models shape a candidate distribution (e.g. a normal distribution) to match a target distribution by applying a sequence of bijective transformations. VAEs use an encoder-decoder architecture with an approximate likelihood approach to parameterize a lower-dimensional latent distribution that can be purposed for generation. In case of GANs, two competing networks are adversarially trained. A generator network transforms a simple prior distribution into a target distribution, while a discriminator network attempts to distinguish real from generated data. If the two networks reach a Nash equilibrium, the data distribution is approximately represented by the generator.

Work at IPAM. Generative models can be used in physical domains as proposal distributions (for instance to explore molecular conformations or generate lattice configurations) or to model inverse problems (such as suggesting new candidates for drug-like molecules or to reconstruct high-resolution molecular structures from low-resolution representations).

One example of a generative model developed during the long program is a backmapping scheme. Molecular dynamics simulations can be prohibitively expensive for large systems, motivating coarse graining to effectively integrate out undesired degrees of freedom enabling access to more physical relevant time- and length-scales (cf. Section 9). However, recovering a temporally coherent super-resolution of the low resolution representation is a challenge. Temporal coherence and backmapping can be important to compute observables in the original space. We approach this problem by using GANs to generate high-resolution structures by incorporating information of past coarse and fine grained representations as inputs into both the generator and the discriminator.

Another example of a generative model developed during the program are the so-called "equivariant normalizing flows". A general theory for designing equivariant flows based on dynamical systems was developed and instantiated for toy systems.

Outlook. A topic discussed during the program aims to incorporate physical symmetries directly into generative models, which can tremendously reduce the complexity of the learning problem. Many challenges nevertheless remain both in the implementation and numerical stability. Lastly, assessing the quality of these learned distributions remains difficult and necessities establishing new, robust evaluation metrics.

11. Outlook

Introduction: Machine learning now plays several roles in physics, namely discovery, providing fast emulators/surrogate models, and understanding physical systems. Techniques like active learning, reinforcement learning and Bayesian optimization are used to accelerate the search for new materials and compounds. Data driven predictors, including kernel ridge regression and neural networks, are being used to approximate the results of costly physical calculations, achieving orders of magnitude speed-ups. In the physical sciences, quantitative prediction and explaining phenomena by compact equations and high level concepts go hand in hand; machine learning is being used to interpret and understand physical data, either from observation or from simulation, and to assist in the discovery of physical laws. Finally, as field matures, ML for physics must become production-ready and show that it can consistently beat the state of the art for real-world problems while maintaining efficiency and stability.

Data production, benchmarks, infrastructure: Producing accurate benchmark datasets that are relevant to real physical challenges requires domain expertise, great care, and significant time. These datasets provide great value to a community by focusing effort and providing clear relative evaluation of methods. We believe more such datasets are needed across a variety of scientific areas, along with protocols for evaluation of existing methods. We started several dataset creation and benchmarking efforts in this program. Further, the community should better acknowledge and reward the scientists who diligently produce these valuable community artifacts. Lastly, along with the datasets themselves, the infrastructure for sharing the data, fairly comparing models, and making those models widely available needs further development.

Physics informed ML Learning problems from Physics differ fundamentally from other applications of ML in industrial or IT scopes: Physics gives us guidance, or even constraints on defining the learning or optimization problem, on symmetries, conservation laws and asymptotic behavior that our predictions should obey. Incorporating physics constraints or intuition into ML models is a very active area of research, it can significantly simplify the learning problem as the optimization needs to search over a smaller set of functions than an uninformed blackbox ML model, and it is a great field for researchers from machine learning and physicists to interact. Concrete progress has been made in various application fields, such as ML for molecular sciences. There is also progress in general understanding, such as the deep mathematical relationships between invariances, equivariances, group theory and convolutions. Challenging open problems persist though, e.g. (1) For certain symmetries we do not know an efficient representation that can scale to large problem instances, such as the antisymmetry constraint in fermionic wavefunctions. (2) While a lot of research has focused on building physical symmetries, conservation laws or intuition into inference models, doing this for generative models is much harder and relatively underexplored.

Physical analysis methods for ML The use of analysis methods from physics for understanding and improving ML models is an active research area. We saw a variety of promising results during our program. However, there remains a significant gap between the high degree of complexity of models actually used in applications and the simplicity of the models that can be effectively analyzed with provable properties. We look forward to the continued narrowing of this gap and insights from the theoretical analysis directly affecting modelling choices for real problems.

Reliability of methods: interpretability, convergence guarantees, uncertainty State-of-the-art ML tools already have important applications in Physics. However, they are still mostly seen as black boxes. Using these tools to advance the physical sciences in a more fundamental way still appears to be a big and promising challenge. There is an urgent need for more interpretable methods that can provide physical intuition and help in the formulation of new general principles (see, e.g., section “Machine Learning Equations”). Additionally, it would be important to be able to associate uncertainties to ML predictions, as it is customary when using classical approaches in Physics. A crucial challenge is to address the reliability of ML models (see, e.g., section “Predictive Uncertainties and Domain of Applicability”). One would like to have ML architectures able to make “good” prediction in the physical domain, however it is crucial to define what “good” means in this context. In order for ML surrogate models to be applied more broadly in the physical sciences we need to be able to trust their predictions. The quantification of uncertainties and the issue of reliability are active areas of research in ML, and the results need to be ported to the physical sciences.

Surpassing the state of the art: While there are classical examples where physics ideas have entered ML (e.g. Boltzmann Machines), and also physicists have employed classical machine learning algorithms such as kernel methods and dense neural networks since decades, the field of machine learning for physics and the physics of learning has received a surge of interest with the recent breakthroughs in deep learning and is fast-growing. In almost every aspect of physics, chemistry and other domain sciences, new ML models are incorporated, extended or even redesigned. We are in a state where this field is very “excited” and moving rapidly. In many cases, even the very fact that a modern ML model is “translated” into a domain science is considered progress and attracts attention.

As the field matures, we must answer an important question for all of these application fields: can we actually surpass the previous state of the art in a fair comparison? For example, a key component in ML is universal function approximators, and there is no doubt that inserting ML components into traditional simulation or modeling workflows adds expressiveness. But can we beat classical state-of-the-art methods in terms of efficiency, as e.g. calling neural network packages to evaluate functions is often still much slower than simpler function representations that have been extensively optimized in code. Also, more flexible ML implementations of physics problems need to demonstrate stability for production purposes, e.g., can we run large-scale instances of simulators that contain novel ML components reliably?

The goal of bringing ML into physics is to better understand the function of natural systems, discover new theories, and design new useful physical objects. ML in physics is just moving from proof of concept to accomplishing these grander goals and we look forward to greater impact and discoveries.